Equilibrium Climate Sensitivity and Transient Climate Response biased low in historical simulations of CMIP6 models

Yue Dong[1], Kyle C. Armour[1], Cristian Proistosescu[1], Timothy Andrews[2], David S. Battisti[1], Piers M. Forster[3], David Paynter[4], Christopher J Smith[3], and Hideo Shiogama[5]

[1]University of Washington
[2]Met Office
[3]University of Leeds
[4]GFDL/NOAA
[5]National Institute for Environmental Studies

November 24, 2022

Abstract

This study assesses the effective climate sensitivity (EffCS) and transient climate response (TCR) derived from global energy budget constraints within historical simulations of 8 CMIP6 global climate models (GCMs). These calculations are enabled by use of the Radiative Forcing Model Intercomparison Project (RFMIP) simulations, which permit accurate quantification of the historical effective radiative forcing. We find that long-term historical energy budget constraints generally underestimate EffCS from CO2 quadrupling and TCR from CO2 ramping, both by 12%, owing to changes in radiative feedbacks and changes in ocean heat uptake efficiency. Atmospheric GCMs forced by observed warming patterns produce lower values of EffCS that are more in line with those inferred from observed historical energy budget constraints. Understanding the discrepancies between modeled and observed historical surface warming patterns remains critical for constraining EffCS and TCR from the historical record.

Table S1. Estimates of radiative feedback parameter, EffCS and TCR. λ_{his} and $\text{EffCS}_{\text{his}}$ values from *amip-piForcing*(*amip*) simulations are calculated from linear regressions over 1870 - 2014 (1979 – 2014); from *historical* simulations are over 1870 – 2014 and 1979 – 2014 (in bracket). λ_{4xCO2} and $\text{EffCS}_{\text{4xCO2}}$ from *abrupt4xCO2* simulations are calculated from regressions of ΔN against ΔT over 150yrs of the simulations. TCR_{his} values from *historical* simulations are calculated by taking differences between 1995 – 2014 and 1869 – 1882. (Note that multi-model mean is calculated by averaging over all 8 models, except for *amip-piForcing* estimates, in which case multi-model mean is average of 6 available models)

	Feedback parameter [$\mathrm{W\,m^{-2}\,K^{-1}}$]			EffCS [K]			TCR [K]	
Models	amipPF(amip)	historical	4xCO2	amipPF(amip)	historical	4xCO2	historical	1pctCO2
CanESM5	-1.46(-1.46)	-0.72(-0.66)	-0.65	2.49(2.50)	5.11(5.51)	5.64	2.60	2.75
CNRM-CM6-1	-1.26(-1.21)	-0.75(-0.67)	-0.74	3.01(3.13)	5.04(5.68)	4.90	2.45	2.23
GFDL-CM4	-1.90(-2.44)	-1.55(-1.40)	-0.82	2.04(1.6)	2.51(2.77)	3.89	1.39	2.05
GISS-E2-1-G	N/A (-1.64)	-1.26(-1.20)	-1.45	N/A (2.01)	2.61(2.76)	2.71	1.44	1.66
HadGEM3-GC31-LL	-1.33(-1.72)	-0.80(-0.63)	-0.63	2.92(2.26)	4.82(6.16)	5.55	1.96	2.47
IPSL-CM6A-LR	-1.64(-1.96)	-1.07(-0.89)	-0.75	2.30(1.93)	3.54(4.28)	4.56	2.16	2.39
MIROC6	-1.50(-1.75)	-1.23(-1.09)	-1.40	2.30(1.96)	2.79(3.15)	2.60	1.55	1.58
NorESM2-LM	N/A (-3.17)	-1.82(-1.60)	-1.34	N/A (1.18)	2.05(2.32)	2.56	1.15	1.48
Multi-model mean	-1.52(-1.92)	-1.15(-1.02)	-0.97	2.51(2.07)	3.54(4.06)	4.05	1.84	2.08

August 18, 2021, 8:30pm

Equilibrium Climate Sensitivity and Transient Climate Response biased low in historical simulations of CMIP6 models

Yue Dong[1], Kyle C. Armour[1,2], Cristian Proistosescu[3], Timothy Andrews[4], David S. Battisti[1], Piers M. Forster[5], David Paynter[6], Christopher J. Smith[5], Hideo Shiogama[7]

[1]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA
[2]School of Oceanography, University of Washington, Seattle, WA, USA
[3]Department of Atmospheric Sciences and Department of Geology, University of Illinois at Urbana-Champaign, IL, USA
[4]Met Office Hadley Centre, Exeter, UK
[5]School of Earth and Environment, University of Leeds, UK
[6]NOAA/Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, NJ, USA
[7]National Institute for Environmental Studies, Tsukuba, Japan

Key Points:

- Within CMIP6 models, historical energy budget constraints underestimate equilibrium climate sensitivity and transient climate response
- Atmosphere-only models forced by observed surface warming patterns produce lower values, in line with historical observations.
- Discrepancies between modeled and observed historical surface warming patterns account for the differences in feedbacks and climate sensitivity estimates

Corresponding author: Yue Dong, `dongy24@uw.edu`

Abstract

This study assesses the effective climate sensitivity (EffCS) and transient climate response (TCR) derived from global energy budget constraints within historical simulations of 8 CMIP6 global climate models (GCMs). These calculations are enabled by use of the Radiative Forcing Model Intercomparison Project (RFMIP) simulations, which permit accurate quantification of the historical effective radiative forcing. We find that long-term historical energy budget constraints generally underestimate EffCS from CO_2 quadrupling and TCR from CO_2 ramping, both by 12%, owing to changes in radiative feedbacks and changes in ocean heat uptake efficiency. Atmospheric GCMs forced by observed warming patterns produce lower values of EffCS that are more in line with those inferred from observed historical energy budget constraints. Understanding the discrepancies between modeled and observed historical surface warming patterns remains critical for constraining EffCS and TCR from the historical record.

Plain Language Summary

Here we use climate models to evaluate the extent to which future warming can be inferred from observations of historical warming. To do so, we employ the RFMIP simulations of 8 models to calculate the historical radiative forcing, and assess historical energy budget constraints on two metrics of global warming: effective climate sensitivity and transient climate response. We find that model historical simulations generally underestimate the climate sensitivity and transient climate response, because historical warming patterns differ from projected warming patterns. Simulations with atmosphere-only models using observed warming patterns produce even lower values of climate sensitivity, suggesting that estimates of climate sensitivity based on recent observations may be similarly biased low.

1 Introduction

Equilibrium climate sensitivity (ECS) and transient climate response (TCR) are two fundamental metrics for evaluating climate change projections. ECS represents the *equilibrium* surface warming in response to a doubling of atmospheric CO_2 concentration relative to pre-industrial levels. Although idealized, ECS has been found to explain most of the spread in projected 21st century global temperature change under realistic emission scenarios (Grose et al., 2018; Sherwood et al., 2020). TCR represents the *tran-*

sient surface warming at the time of CO_2 doubling under an idealized 1% per year CO_2 increase. As a measure of transient response, TCR is better constrained and is also informative about the projected degree of global warming in the coming century.

In principle, ECS and TCR can be inferred from the global energy balance framework (Gregory et al., 2002):

$$\Delta N = \Delta F + \lambda \Delta T, \tag{1}$$

where ΔN is the global-mean top-of-atmosphere (TOA) radiation anomaly (approximately equal to ocean heat uptake), ΔF is the effective radiative forcing (ERF; Myhre et al., 2013), ΔT is the global-mean surface air temperature anomaly, and λ is the radiative feedback parameter (negative for a stable climate). ECS is inferred as: $\mathrm{ECS} = -F_{2\mathrm{x}}/\lambda_{eq}$, where $F_{2\mathrm{x}}$ is the ERF from CO_2 doubling, and λ_{eq} is the radiative feedback when a new equilibrium is reached ($\Delta N = 0$). Ideally, ECS can be estimated from equilibrium states within global climate models (GCM) forced by an abrupt CO_2 doubling (*abrupt2xCO2*) or CO_2 quadrupling (*abrupt4xCO2*), after sufficiently long integration (Rugenstein et al., 2020). In practice, ECS is often extrapolated from a linear regression of ΔN against ΔT for the first 150yrs of *abrupt4xCO2* simulations (Gregory et al., 2004). This extrapolation generally underestimates the true ECS due to changes in radiative feedbacks as climate equilibrates (Rugenstein et al., 2020), owing to time-evolving surface warming patterns (e.g., Armour et al., 2013; Andrews et al., 2015; Dong et al., 2020), and nonlinear mean-state dependence of radiative feedbacks (e.g., Caballero & Huber, 2013; Bloch-Johnson et al., 2015, 2021). Therefore, we refer the ECS values estimated from these non-equilibrium states to as an *effective* climate sensitivity (EffCS; Andrews et al., 2015; Sherwood et al., 2020):

$$\mathrm{EffCS} = -\frac{F_{2\mathrm{x}}}{\lambda_{\mathrm{eff}}}, \tag{2}$$

assuming the effective radiative feedback (λ_{eff}) at a transient state would remain constant to equilibrium. Specifically, in this paper we denote EffCS values from *abrupt4xCO2* simulations (through regressions of annual-mean ΔN against ΔT for the first 150yrs) as $\mathrm{EffCS}_{4\mathrm{xCO2}}$ (data from Zelinka et al., 2020). We also estimate EffCS values from the historical energy budget constraints (using Eqs. 1 and 2), in which case we refer to it as $\mathrm{EffCS}_{\mathrm{his}}$.

For TCR, it is commonly calculated as the global-mean surface air temperature change averaged over a 20-year period centered on year 70 of the *1pctCO2* simulations where

CO_2 concentration is doubled (denoted here as TCR_{1pct}). The values of TCR can also be estimated from historical energy budget constraints (see more details in Section 4), in which case we will refer to it as TCR_{his}.

$EffCS_{4xCO2}$ and TCR_{1pct} estimates from fully-coupled atmosphere-ocean GCMs (AOGCMs) have been found to be substantially higher than values of $EffCS_{his}$ and TCR_{his} from observed energy budget constraints (e.g., Otto et al., 2013; Lewis & Curry, 2015, 2018; Forster, 2016), raising a key question of whether the AOGCMs are overly sensitive. However, lower values of $EffCS_{his}$ and TCR_{his} have also been found in AOGCM *historical* simulations, but only in a few models (Winton et al. 2020 for GFDL-CM4; Andrews et al. 2019 for HadGEM3-GC3.1-LL; Dessler et al. 2018 for MPI-ESM1.1). The limited number of model studies reflects the fact that the time-varying historical ERF (ΔF in Eq. 1) is not often diagnosed, precluding accurate calculation of radiative feedback and thus $EffCS_{his}$ and TCR_{his}. Some other studies have instead used *abrupt4xCO2* or *1pctCO2* simulations as a surrogate for historical warming (Armour, 2017; Proistosescu & Huybers, 2017; Lewis & Curry, 2018; Dong et al., 2020), or used a rough estimate of historical ERF taken from IPCC AR5 (Myhre et al., 2013) for CMIP5 AOGCMs (Marvel et al., 2018; Gregory et al., 2020). These approaches provide inter-model comparisons, and generally find that $EffCS_{4xCO2}$ is larger than $EffCS_{his}$, but it is unclear how accurate their estimates are given that they do not use model-specific estimates of ERF (or in some cases lack historical non-CO_2 forcings altogether).

This work is thus motivated by two key questions: (1) how robust is the finding that values of $EffCS_{4xCO2}$ and TCR_{1pct} are higher than values of $EffCS_{his}$ and TCR_{his} estimated using historical energy budget constraints? (2) How do the estimates of $EffCS_{his}$ and TCR_{his} from models compare to those from observed energy budget constraints? The answers to these questions have major implications for how the historical record informs future climate projections. Here we employ simulations of the Radiative Forcing Model Intercomparison Project (RFMIP; Pincus et al., 2016), which provide the time series of historical ERF for 8 CMIP6 AOGCMs (section 2). With ERF in hand, we assess $EffCS_{his}$ and TCR_{his} values within *historical* simulations and compare to the values of $EffCS_{4xCO2}$ and TCR_{1pct} for each of these models (section 3 and 4). We then compare EffCS and TCR estimates between models and observations, and discuss implications for observation-based historical energy budget constraints on EffCS and TCR (section 5).

2 Data

2.1 Historical Effective Radiative Forcing from RFMIP Simulations

The ERF includes rapid adjustments from the atmosphere in response to changes in CO_2 or other forcing agents (Myhre et al., 2013). It can be quantified from the TOA radiation changes within atmosphere-only GCM (AGCM) simulations wherein forcing agents are changed while SST and sea-ice concentration (SIC) fields are fixed at pre-industrial values (Forster et al., 2016). Here we make use of the fixed-SST simulations of RFMIP that are currently available for 8 CMIP6 models (CanESM5, CNRM-CM6-1, GFDL-CM4, GISS-E2-1-G, HadGEM3-GC31-LL, IPSL-CM6A-LR, MIROC6, NorESM2-LM). The time-series of historical ERF is calculated as the difference of net TOA radiative flux between a 30-year control run (*piClim-control*), where all forcing agents are fixed to pre-industrial levels, and a forcing run (*piClim-histall*), where time-varying atmospheric concentrations of all historical forcing agents are imposed. Historical ERF from a single group of forcing agents (e.g., greenhouse gases, anthropogenic aerosols, natural forcings including volcanoes and solar variability) can also be estimated from single-forcing runs of RFMIP (*piClim-histghg*, *piClim-histaer*, *piClim-histnat*, respectively). We also estimate ERF of CO_2 doubling, F_{2x}, from RFMIP *piClim-4xCO2* siulations, where CO_2 is abruptly quadrupled and held constant for 30yrs while SST and SIC fields are fixed. F_{2x} is computed from the TOA radiation changes of the 30yr-average (scaled by 1/2 to estimate the forcing for CO_2 doubling from CO_2 quadrupling simulations). For all RFMIP simulations, the ensemble mean is used when more than one member of the simulation exist.

Note that the TOA radiation flux changes derived from the fixed-SST simulations includes the effect of temperature changes over land and sea ice, which should be considered as part of the radiative response rather than ERF. We remove this portion of radiative effects by subtracting off the global-mean surface air temperature change scaled by each model's radiative feedback parameter from its *abrupt4xCO2* simulation – the method proposed in Hansen et al. (2005). Recent studies find advantages in several new correction methods, such as fixing both SST and land-surface temperatures in AGCM simulations (Andrews et al., 2021), or using surface temperature radiative kernels (Smith et al., 2020). We choose to apply the Hansen et al. (2005) method here given that it is a widely used method and readily improves ERF estimates using the available output of the RFMIP simulations. All the historical ERFs are calculated as global and annual means, spanning the period 1850 – 2014 (the same interval of fully-coupled *historical* simulations).

2.2 Historical simulations of AOGCMs and AGCMs

Within *historical* simulations of AOGCMs, we compute global mean N and T from the mean of all available ensemble members, in attempt to reduce noises from internal variability. Except for GFDL-CM4 (1 member), NorESM2-LM (3 members), and HadGEM3-GC31-LL (4 members), all models have more than 10 *historical* ensemble members available. The annual mean changes of N and T relative to pre-industrial levels are calculated by subtracting a linear fit of the global annual mean *piControl* values to remove unforced model drift. Note that ΔN, ΔF, and ΔT in the energy budget framework (Eq. 1) can also be defined as differences between two specific historical states. We will elaborate the periods over which we compute the historical energy balance in the following two sections. In order to examine the contributions of individual forcing agents to historical climate change, we also employ single-forcing historical simulations (*hist-GHG*, *hist-aer*, *hist-nat*) described by the Detection and Attribution Model Intercomparison Project (DAMIP; Gillett et al., 2016), where only one type of forcing agent is changed while all other forcing agents are fixed at preindustrial levels.

Results from the coupled AOGCMs are compared to two sets of AGCM simulations. One is the *amip* simulation, a CMIP6 DECK experiment (Eyring et al., 2016) where AGCMs are forced by time-evolving observed SST and SIC fields and by time-varying historical forcing agents. While *amip* simulations are available for all of the 8 CMIP6 models assessed here, they are performed only over 1979 – 2015. To investigate early historical energy budget in AGCMs, we also make use of *amip-piForcing* simulations described by the Cloud Feedback Model Intercomparison Project (CFMIP; Webb et al., 2017), which are available from 1870 to 2014. Similar to *amip*, the *amip-piForcing* simulations are forced by observed SST and SIC fields while all forcing agents are fixed at pre-industrial levels (i.e., ERF is zero). Radiative feedbacks are in theory identical between a model's *amip-piForcing* and *amip* runs because SST and SIC fields are the same, assuming the linearity of the global energy balance (Eq. 1) and that the ERF added to *amip* simulations are accurately quantified by RFMIP simulations. A caveat is that only 6 out of 8 models used here (CanESM5, CNRM-CM6-1, GFDL-CM4, HadGEM3-GC31-LL, IPSL-CM6A-LR, MIROC6) have *amip-piForcing* experiments available. Given that most of the variability in TOA radiative fluxes comes about through variations in SSTs which are fixed in AGCM simulations, for both sets of AGCM simulations we only use the first realization of each model.

3 Historical Energy Budget Constraints on Radiative Feedbacks and EffCS

In the energy budget framework, $\mathrm{EffCS_{his}}$ can be written as:

$$\mathrm{EffCS_{his}} = -\frac{F_{2x}}{\lambda_{\mathrm{his}}}, \tag{3}$$

where the historical effective radiative feedback parameter (λ_{his}) is given by:

$$\lambda_{\mathrm{his}} = \frac{\Delta N - \Delta F}{\Delta T}. \tag{4}$$

We first show historical variations in λ_{his}, calculated by a linear regression form of Eq. 4 in a sliding 30-year window. We find remarkable differences in decadal-scale radiative feedbacks between *historical* simulations (black line in Fig. 1a) and *amip-piForcing* simulations (blue line in Fig. 1a). While natural variability may have played a dominant role in the first half of the 20th century, where net ERF was relatively small (Fig. S1), the discrepancy between AOGCMs and AGCMs persists throughout the full historical period towards early 21st century. Notably, λ_{his} in the *amip-piForcing* simulations of AGCMs trends toward more-negative values since 1970s to present, consistent with earlier studies using CMIP5 models (Andrews et al., 2018; Silvers et al., 2018; Gregory et al., 2020; Dong et al., 2019); whereas in the *historical* simulations of AOGCMs, the trend of λ_{his} is rather weak or even slightly towards more-positive values. During the second half of the century, λ_{his} values from AOGCMs track those in *hist-GHG* simulations (red line), suggesting the simulated feedbacks are primarily driven by GHG forcing, which has dominated global net ERF over this period (Fig. S1).

We next assess $\mathrm{EffCS_{his}}$ from energy budget constraints within the *historical* simulations of the AOGCMs and the AGCMs. To compute the energy budget in Eqs. 3 and 4, the time interval over which anomalies (Δ) are calculated needs to be carefully chosen to avoid short-term variability and effects of volcanic eruptions (Lewis and Curry 2014; Forster 2016). Previous studies have often used two methods: (1) taking finite differences between a base period and a final period (Lewis & Curry, 2015, 2018; Winton et al., 2020; Sherwood et al., 2020); or (2) using regression over the full period of interest (Gregory et al., 2020; Andrews et al., 2019). Since we are comparing EffCS between

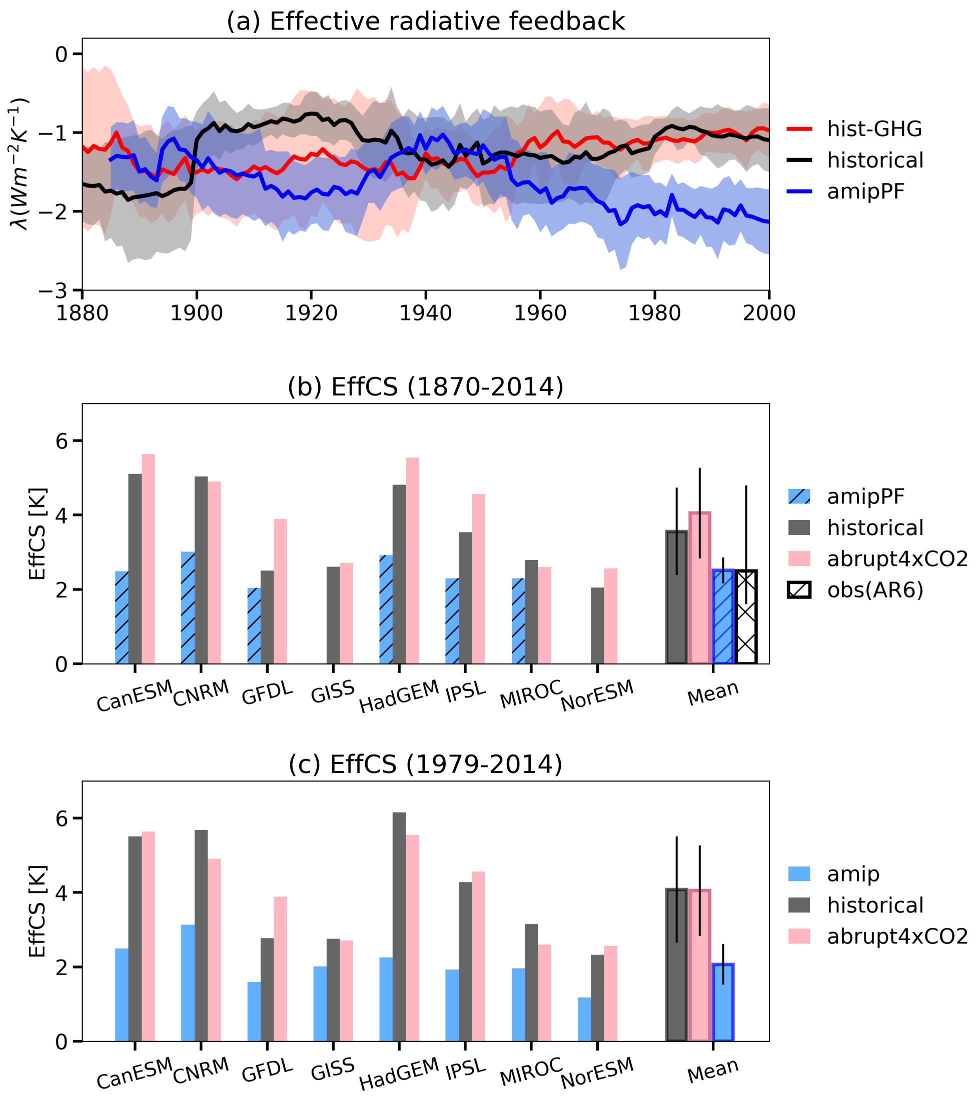

Figure 1. Historical energy budget constraints on net radiative feedback and EffCS. (a) Time series of the estimated historical radiative feedbacks (λ_{his}). Thick lines denote multi-model means, shadings denote one standard deviation across models. (b, c) EffCS estimated from the energy budget of (b) full historical record (1870 – 2014) and (c) recent decades (1979 – 2014). The outlined colored bars on the right in (b, c) denote the multi-model mean values of EffCS from corresponding simulations, with error bars indicating one standard deviation across models. The white hatched bar in (b) denotes the median $\mathrm{EffCS}_{\mathrm{his}}$ value of 2.5K based on observed energy budget changes reported in IPCC AR6 (Forster et al., in press), with the error bars denoting 5-95% range of 1.6 – 4.8 K. Models listed (from the left to right) are: CanESM5, CNRM-CM6-1, GFDL-CM4, GISS-E2-1-G, HadGEM3-GC31-LL, IPSL-CM6A-LR, MIROC6, NorESM2-LM.

AOGCMs and AGCMs (including *amip* simulations which are only available from 1979 onwards), we choose to use the regression method here. That is, the λ_{his} used to compute EffCS$_{\mathrm{his}}$ (Eq. 3) is calculated via ordinary least squares (OLS) regression of Eq. 4, over two periods of our interest: the full historical period 1870 – 2014 (Fig. 1b), and the recent decades of the Satellite Era 1979 – 2014 (Fig. 1c).

3.1 EffCS$_{\mathrm{his}}$ from Long-term Historical Energy Budget (1870 - 2014)

The values of EffCS$_{\mathrm{his}}$ inferred from long-term energy budget in *historical* simulations are generally lower than EffCS$_{\mathrm{4xCO2}}$ from *abrupt4xCO2* simulations (Fig. 1b). As noted above, the difference between EffCS$_{\mathrm{his}}$ and EffCS$_{\mathrm{4xCO2}}$ has been documented in a few models. For GFDL-CM4, Winton et al. (2020) found an EffCS$_{\mathrm{his}}$ of 1.8K and an EffCS$_{\mathrm{4xCO2}}$ of 4K (EffCS$_{\mathrm{4xCO2}}$= 5K if using yrs 51-300 of the model's extended *abrupt4xCO2* simulation). For HadGEM3-GC3.1-LL, Andrews et al. (2019) found an effective $F_{2\mathrm{x}}$ of 3.49 $\mathrm{W\,m^{-2}}$ and a historical feedback of 0.86 $\mathrm{W\,m^{-2}\,K^{-1}}$(average of 4 ensembles), implying an EffCS$_{\mathrm{his}}$ of 4.1K, in contrast to the model's EffCS$_{\mathrm{4xCO2}}$ of 5.5K. For MPI-ESM1.1, Dessler et al. (2018) found an ensemble-median EffCS$_{\mathrm{his}}$ of 2.72K, slightly lower than the value of EffCS of 2.93K estimated from a CO_2 doubling simulation. Here we show that, within 6 out of 8 fully-coupled CMIP6 AOGCMs assessed, historical energy budget constraints underestimate EffCS$_{\mathrm{4xCO2}}$ from CO_2 quadrupling, with an average of 12% lower across all models (Table S1). Averaging over all 8 AOGCMs, EffCS$_{\mathrm{his}}$ is 3.56 K ($\pm$1.17K; one standard deviation across models, unless noted elsewhere) and EffCS$_{\mathrm{4xCO2}}$ is 4.05K ($\pm$1.46 K), corresponding to an averaged λ_{his} of -1.15 $\mathrm{W\,m^{-2}\,K^{-1}}$ ($\pm$0.37 $\mathrm{W\,m^{-2}\,K^{-1}}$) and λ_{4xCO2} (from the regression of 150yrs *abrupt4xCO2* simulations, data from Zelinka et al. 2020) of -0.97 $\mathrm{W\,m^{-2}\,K^{-1}}$ ($\pm$0.34 $\mathrm{W\,m^{-2}\,K^{-1}}$), respectively. Lower values of EffCS$_{\mathrm{his}}$ are found in AGCM *amip-piForcing* experiments over the same historical period, with a mean EffCS$_{\mathrm{his}}$ value of 2.51K ($\pm$0.35K) across 6 available models, which is lower than the mean EffCS$_{\mathrm{4xCO2}}$ value of 4.52 K ($\pm$1.04 K) across the same 6 models by 44%. Using the Winton et al. (2020) method, i.e., taking ΔN, ΔT and ΔF as differences between 1869 – 1882 and 1995 – 2014, yields nearly the same result (the difference between the two methods is statistically insignificant in a T-test): the mean value of EffCS$_{\mathrm{his}}$ is 3.50 K ($\pm$1.31 K) from *historical* simulations across all 8 AOGCMs and 2.54 K ($\pm$0.4 K) from *amip-piForcing* simulations across 6 available AGCMs.

The EffCS and radiative feedback differences between historical energy budget constraints and CO_2 quadrupling in models arise primarily from differences between historical and near-equilibrium warming patterns (Fig. 2). Under CO_2 quadrupling, AOGCMs generally project a warming pattern that is representative of the equilibrium response showing polar amplification and weakened tropical Pacific west-east SST gradient (Fig. 2c; Andrews et al., 2015; Ceppi & Gregory, 2017; Andrews & Webb, 2018; Dong et al., 2020); whereas the SST trend pattern in *historical* simulations appears to be more spatially uniform (Fig. 2a). It has been argued that the projected enhancement of warming in the tropical eastern Pacific relative to the tropical western Pacific in models tends to weaken the lower tropospheric stability, thereby weakening the negative low cloud feedback and negative lapse-rate feedback, producing a higher EffCS (Zhou et al., 2016; Ceppi & Gregory, 2017; Andrews & Webb, 2018; Dong et al., 2019). In contrast, the relatively uniform tropical warming patterns simulated in *historical* simulations would maintain negative cloud feedback and therefore lower EffCS. The fact that $\mathrm{EffCS_{his}}$ estimates from *amip-piForcing* simulations are even lower reflects that the observed historical warming pattern shows slightly enhanced warming in the Indo-Pacific Ocean and delayed warming in both the eastern Pacific Ocean and part of the Southern Ocean (e.g., Zhou et al., 2016; Dong et al., 2019, 2020; Silvers et al., 2018).

The historical pattern effect that leads to lower values of $\mathrm{EffCS_{his}}$ may partially result from various non-CO_2 forcing agents that have operated in the historical period (e.g., Marvel et al., 2016; Forster, 2016). Gregory et al. (2020) suggest that volcanic forcing may bias estimate of EffCS from CO_2 quadrupling by causing different surface warming patterns in CMIP5 models. Winton et al. (2020) find that a large portion of the $\mathrm{EffCS_{his}}$ underestimate in GFDL-CM4 is attributable to its large efficacy of aerosol forcing. To test this possibility within other CMIP6 models, we make use of the DAMIP non-GHG forcing simulations, namely, *hist-aer* and *hist-nat* (Fig. S2). Within all but one model, natural forcing (volcanoes and solar variability) alone produces even lower values of $\mathrm{EffCS_{his}}$ than those from *historical* simulations (i.e., a larger historical pattern effect). In comparison, when forced by anthropogenic aerosol forcing alone, four models show a larger historical pattern effect while three models show a reduced pattern effect. These results suggest that natural forcing may be key to the historical pattern effect, while the effect of aerosol forcing is less robust across models.

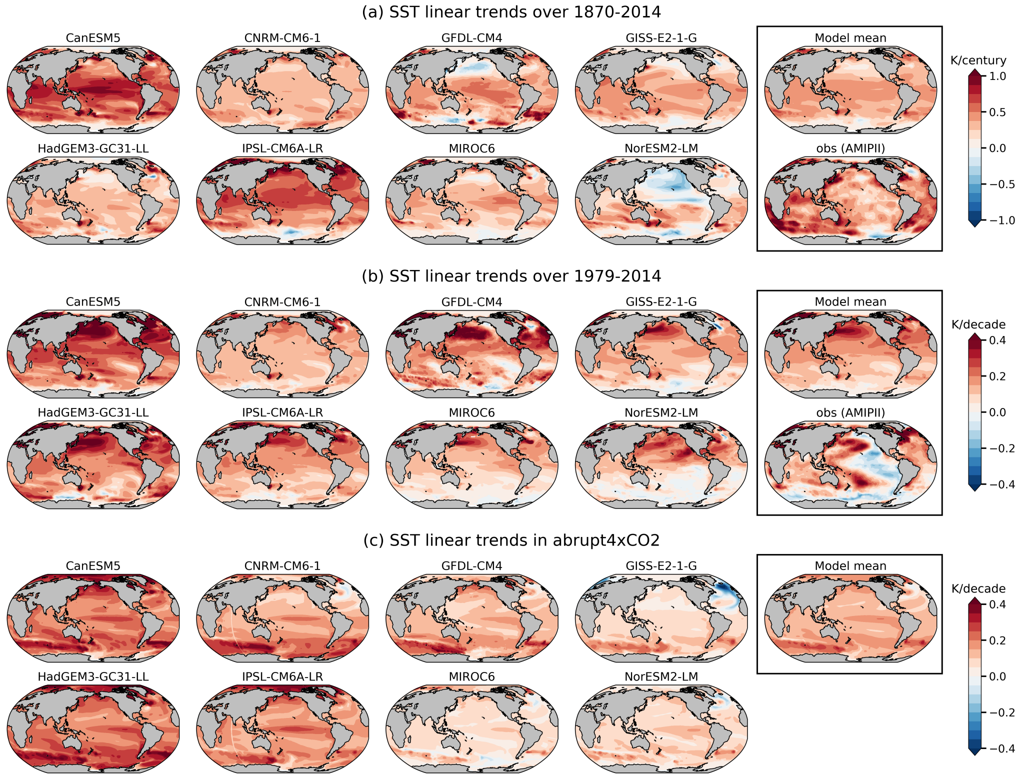

Figure 2. Historical and equilibrium SST trend patterns. SST linear trends over (a) 1870 – 2014, (b)1979 – 2014, and (c) 150 years of *abrupt4xCO2* simulations, calculated via OLS regressions of annual-mean SST against time. The observed SST trend patterns in (a, b) are calculated using AMIPII dataset (Hurrell et al., 2008). The model-mean SST trend patterns is all panels are calculated by averaging over all 8 AOGCMs.

3.2 $\mathrm{EffCS_{his}}$ from Recent Energy Budget (1979-2014)

Having quantified the long-term historical energy budget constraint on EffCS, we next focus on the most recent decades 1979 – 2014 (Fig. 1c), where observations of global SSTs have been improved by satellite products. This is also the period where GHG forcing has increased dramatically while aerosol forcing trends are relatively small (Fig. S1). With stronger ERF having operated over this period, nearly all coupled AOGCMs produce higher values of $\mathrm{EffCS_{his}}$, with a multi-model mean $\mathrm{EffCS_{his}}$ of 4.08K (corresponding to a mean radiative feedback of -1.02 $\mathrm{W\,m^{-2}\,K^{-1}}$), comparable to the mean $\mathrm{EffCS_{4xCO2}}$ of 4.05K.

Does this imply that the historical pattern effect is weak in recent decades? In fact, the $\mathrm{EffCS_{his}}$ values over this period from all 8 AOGCMs are substantially higher than values of $\mathrm{EffCS_{his}}$ from their AGCM counterparts driven by observed warming patterns within *amip* simulations (blue bar in Fig. 1c) and *amip-piForcing* simulations (not shown). Averaging over all AGCMs, the mean $\mathrm{EffCS_{his}}$ and the corresponding λ_{his} from *amip* experiments is 2.07 K (±0.57 K) and -1.92 $\mathrm{W\,m^{-2}\,K^{-1}}$ (±0.58 $\mathrm{W\,m^{-2}\,K^{-1}}$), respectively. The $\mathrm{EffCS_{his}}$ difference between coupled AOGCMs and their counterpart AGCMs can be traced to the difference between modeled and observed SST patterns over recent decades. The ensemble-mean SST trend pattern in *historical* simulations of AOGCMs fails to capture many key features in observations (Fig. 2b), including the pronounced cooling trends over the eastern Pacific and Southern Ocean. In particular, the observed enhancement of tropical Pacific zonal SST gradient has been linked to the observed increase in low clouds over the stratocumulus deck, which contributes to a more-negative radiative feedback and lower EffCS (Zhou et al., 2016; Ceppi & Gregory, 2017; Dong et al., 2019; Fueglistaler, 2019). A few studies have argued that the observed tropical Pacific SST pattern may be driven by anthropogenic sulfate aerosol forcing (Takahashi & Watanabe, 2016) or volcanic forcings (Gregory et al., 2020). Using the DAMIP simulations, we found that the SST trend patterns driven by anthropogenic aerosol forcing and natural forcing are indeed more spatially heterogeneous, with some models showing weak cooling in the tropical eastern Pacific (Fig. S3). However, the cooling trends produced in these non-GHG simulations are much weaker than that observed, and are therefore generally overwhelmed by the warming trends produced by GHG forcing (Fig. S3a).

It is also possible that the observed warming pattern is in part a result of natural variability and therefore expected to differ from the ensemble mean of model simulations. For example, Watanabe et al. (2021) found that the observed equatorial Pacific west-east SST gradient over a longer historical period (1951-2010) lies within the range of large ensembles of model simulations. Olonscheck et al. (2020) also suggests a general consistency between observed SSTs and those simulated by CMIP5 and CMIP6 models, when focusing on specific regions. However, such regional analyses may be insufficient to explain the broad pattern of observed SST trends beyond the equatorial Pacific (e.g., the cooling off the coast in the subtropics and in the Southern Ocean), and their results may be sensitive to the region and time interval selected. We have examined $\mathrm{EffCS_{his}}$ and the equatorial Pacific zonal SST gradient for all individual members of *historical* simulations. We define the zonal SST gradient following Watanabe et al. (2021): the difference between the eastern Pacific (180°- 80°W, 5°S - 5°N) and the western Pacific (110°E-180°, 5°S - 5°N). But we calculate SST linear trends over 1979 – 2014 instead of 1951 – 2010 as in Watanabe et al. (2021). Over these recent decades, nearly all of the 180 CMIP6 ensemble members fail to capture the low $\mathrm{EffCS_{his}}$ values from the corresponding *amip* simulations and the observed zonal SST gradient (Fig. S4), suggesting a significant bias in the pattern effect between AOGCMs and observations.

Identifying the causes of the recent observed SST trend pattern is beyond the scope of this study. Our results on the historical energy budget constraints suggest that $\mathrm{EffCS_{his}}$ estimates from *historical* simulations generally underestimate $\mathrm{EffCS_{4xCO2}}$ from CO_2 quadrupling due to the pattern effect. However, the historical pattern effect is relatively small over recent decades in AOGCMs, owing to the fact that their historical warming patterns over recent decades are not substantially different from their equilibrium warming patterns. In section 5, we will further examine model-observation comparisons.

4 Historical Energy Budget Constraints on TCR

In the energy budget framework, TCR can be inferred from sufficiently long-term historical record where ΔT increases approximately proportional to ΔF:

$$\mathrm{TCR_{his}} = \Delta T \frac{F_{2\mathrm{x}}}{\Delta F}. \tag{5}$$

Under the global energy framework (Eq. 1), $\mathrm{TCR_{his}}$ is governed by both historical radiative feedback (λ_{his}) and ocean heat uptake (OHU) efficiency (κ_{his}), with the relationship between these approximated as (Gregory & Mitchell, 1997; Raper et al., 2002; Gregory & Forster, 2008; Gregory et al., 2015):

$$\mathrm{TCR_{his}} = \frac{F_{2\mathrm{x}}}{\kappa_{\mathrm{his}} - \lambda_{\mathrm{his}}}, \tag{6}$$

where κ_{his} is defined as:

$$\kappa_{\mathrm{his}} = \frac{\Delta N}{\Delta T}. \tag{7}$$

Here we calculate $\mathrm{TCR_{his}}$ from *historical* simulations using Eq. 5, where anomalies (Δ) are averaged over 1995 – 2014 relative to 1869 – 1882. This period is chosen to cover a sufficiently long time of historical record, and also to be largely consistent with several recent studies (Lewis & Curry, 2018; Winton et al., 2020). As noted above, Winton et al. (2020) found a $\mathrm{TCR_{his}}$ of 1.27K for GFDL-CM4, lower than the model's $\mathrm{TCR_{1pct}}$ of 2.05K. Here we find that most of the AOGCMs are consistent with GFDL-CM4 – the historical energy budget constraint underestimates TCR values from *1pctCO2* simulations, by about 12% on average (Fig. 3a). The mean $\mathrm{TCR_{his}}$ value across 8 AOGCMs is 1.84K ($\pm$0.51K), lower than the mean $\mathrm{TCR_{1pct}}$ value of 2.08K ($\pm$0.43K).

As shown in Eq. 6, the difference between $\mathrm{TCR_{his}}$ and $\mathrm{TCR_{1pct}}$ could arise from changes in radiative feedbacks and/or changes in OHU efficiency over time (Gregory et al., 2015). To separate these two factors, we estimate λ and κ from *historical* and *1pctCO2* simulations, following Eq. 4 and Eq. 7, respectively. For *historical* estimates, ΔN, ΔT and ΔF are taken as finite differences between 1995 - 2014 and 1869 - 1882. For *1pctCO2* estimates, ΔN and ΔT are from the 20-year period centered on year 70 of the simulation when CO_2 is doubled (ΔT is equivalent to $\mathrm{TCR_{1pct}}$); ΔF at the time of CO_2 doubling is approximated by $F_{2\mathrm{x}}$, with a caveat that the true $F_{2\mathrm{x}}$ in *1pctCO2* simulations was found slightly non-logarithmic (Gregory et al., 2015, 2020). In all models κ_{his} is larger than κ_{1pct} which could contribute to the lower values of $\mathrm{TCR_{his}}$ relative to $\mathrm{TCR_{1pct}}$ (Fig. 3b). A weakening of κ over time has also been discovered within *1pctCO2* simulations, from the first doubling of CO_2 to the second doubling, which contributes to an increase in TCR during the second 70-yrs of the simulations (e.g., Gregory & Forster, 2008; Gregory et al., 2015). On the other hand, the difference between λ_{his} and λ_{1pct} varies by models (Fig. 3c). Two models show λ_{his} more negative than λ_{1pct}, along with their large κ_{his}, suggesting that the lower values of $\mathrm{TCR_{his}}$ in these models are owing to changes in both radiative feedbacks and OHU efficiency. The rest of the models show λ_{his} either very close to or slightly less negative than λ_{1pct}, suggesting a dominant role of changes in κ. One

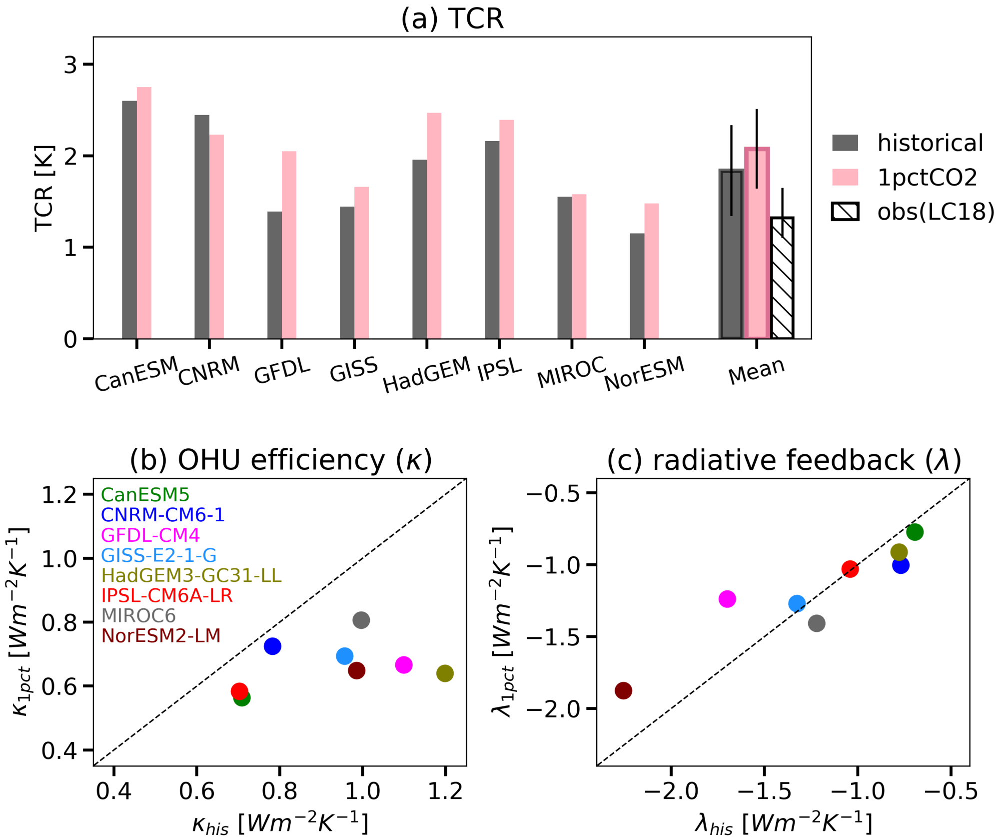

Figure 3. (a) TCR estimates from historical energy budget constraints and *1pctCO2* simulations. Black bars denote TCR_{his} values from fully-coupled *historical* simulations. Red bars denote TCR_{1pct} values from *1pctCO2* simulations. The white hatched bar denotes the best estimate of TCR_{his} of 1.32K based on observed energy budget changes reported by Lewis and Curry (2018), with a 17-83% range of 1.1-1.65K. (b) Ocean heat uptake efficiency and (c) radiative feedback from *historical* and *1pctCO2* simulations of all 8 AOGCMs.

exception is CNRM-CM6-1, whose TCR_{his} is higher than TCR_{1pct}. The overestimate of its TCR_{his} can be traced to a less-negative λ_{his} and a small difference in κ_{his}.

In summary, we find an overall underestimate of TCR of about 0.2K using historical energy budget constraints within AOGCMs. Values of TCR_{his} are generally biased low owing to the combination of too-negative radiative feedback and/or too-large OHU efficiency during the historical period. The differences in λ and κ between *historical* and *1pctCO2* are largely ameliorated when using *hist-GHG* simulations (Fig. S5), suggesting that the underestimate of TCR_{his} is mostly driven by historical non-GHG forcings. Overall, these results suggest that as time evolves and CO_2 forcing increases, the weakening of both radiative feedback and OHU efficiency could lead to higher values of TCR than those inferred from historical energy budget constraints.

5 Discussions and Conclusions

In the previous two sections, we have compared estimates of EffCS and TCR between different simulations of coupled and atmosphere-only GCMs. How do the model results compare to values of $EffCS_{his}$ and TCR_{his} from the observed energy budget constraints, and what implications do the results have for our interpretation of these observation-based estimates?

In Figs. 1 and 3 we show that the reported values of $EffCS_{his}$ and TCR_{his} from observations are much lower than the values of $EffCS_{4xCO2}$ and TCR_{1pct} from CMIP6 models. For an observation-based estimate of $EffCS_{his}$, we use values reported in IPCC AR6 (Forster et al., in press): a median value of 2.5K and 5 - 95% range of 1.6 – 4.8K based on observed energy budget changes from 1850 - 1900 to 2006 – 2019 (Fig. 1b). For TCR, we use values reported by Lewis and Curry (2018): a median value of 1.32K and a 17–83% range of 1.1 – 1.65 K based on observed energy budget changes over 1869 – 1882 to 1995 – 2016 (Fig. 3). Values of $EffCS_{his}$ from AGCM simulations forced by observed SST patterns are well in line with observation-based values of $EffCS_{his}$ (c.f. blue bars and white bar in Fig. 1b), despite the fact that AOGCM values of $EffCS_{his}$ and $EffCS_{4xCO2}$ are both higher, confirming the fidelity of the radiative response of atmospheric models given observed SST and SIC trends (Andrews et al., 2018; Loeb et al., 2020). The difference between projected $EffCS_{4xCO2}$ and observationally-constrained $EffCS_{his}$ is thus owing to changes in SST patterns with time. It implies that if nature evolves towards equilibrium in the way that AOGCMs project, we should expect higher values of EffCS and TCR (i.e., evolving toward $\text{EffCS}_{4\text{xCO2}}$ and $\text{TCR}_{1\text{pct}}$) in the future than those inferred from observed historical energy budget constraints.

Our findings are broadly consistent with earlier studies focusing on two individual CMIP6 models (Andrews et al., 2019; Winton et al., 2020): historical energy budget constraints generally (within 6 out of 8 AOGCMs) underestimate the values of EffCS from CO_2 quadrupling and TCR from CO_2 ramping, both by 12% on average. The underestimate of $\text{EffCS}_{\text{his}}$ is owing to differences in radiative feedbacks induced by the pattern effect; the underestimate of TCR_{his} is owing to a combination of differences in both radiative feedbacks and ocean heat uptake efficiency. Using observations, historical energy budget constraints provide even lower values of $\text{EffCS}_{\text{his}}$ and TCR_{his}, which are in line with the values from AGCMs forced by observed SSTs and SICs. Accounting the pattern effect and assuming the observed SST pattern will evolve towards the projected equilibrium warming pattern, the observed historical energy budget may provide a baised-low constraint on EffCS and TCR.

That said, the projections by GCMs are confronted by not only uncertainties associated with atmospheric physics, e.g., cloud feedbacks (Webb et al., 2013; Zelinka et al., 2020; Sherwood et al., 2020), but also a critical open question: how reliable are model projections of future SST and SIC patterns? We find that estimates of $\text{EffCS}_{\text{his}}$ and TCR_{his} from *historical* simulations of coupled AOGCMs fall outside of the range of observation-based values, due to differences between observed and modeled SST trend patterns, which are particularly acute over recent decades. If the observed SST trend pattern (e.g., the strengthening of the tropical zonal SST gradient) is caused by internal natural variability, which will reverse sign in the coming decades according to AOGCM projections (Watanabe et al., 2021), then the higher values of EffCS and TCR found within AOGCMs may be more informative about near-future climate change under continued CO_2 forcing. If the recent observed SST trend pattern is a result of model biases in the response to anthropogenic forcing (e.g., Seager et al., 2019; Coats & Karnauskas, 2017), the lower values of $\text{EffCS}_{\text{his}}$ and TCR_{his} from observations may persist over the coming decades, in which case 21st century warming may be lower than GCMs project. This work suggests that understanding the causes of the recent observed surface warming pattern and model-observation discrepancies is critical for constraining transient and near-equilibrium climate change.

Acknowledgments

We thank Michael Winton and Nadir Jeevanjee for constructive comments. YD, KCA were supported by National Science Foundation Grant AGS-1752796. CP was supported by National Science Foundation grant OCE-2002385. DSB was supported by the Tamaki Foundation. TA was supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra, and the European Union's Horizon 2020 research and innovation program under grant agreement No. 820829 (CONSTRAIN project). Funding for PMF was provided by the European Union's Horizon 2020 Research and Innovation Programme under grant no. 820829 (CONSTRAIN). CJS was supported by a NERC/IIASA Collaborative Research Fellowship (NE/T009381/1). HS was supported by the Integrated Research Program for Advancing Climate Models (JPMXD0717935457) and Grants-in-Aid for Scientific Research (21H01161) of the Ministry of Education, Culture, Sports, Science and Technology of Japan. The CMIP6 data used in this study are available at: https://esgf-node.llnl.gov/search/cmip6/.

References

Andrews, T., Andrews, M. B., Bodas-Salcedo, A., Jones, G. S., Kuhlbrodt, T., Manners, J., et al. (2019). Forcings, feedbacks, and climate sensitivity in HadGEM3-GC3.1 and UKESM1. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4377–4394.

Andrews, T., Gregory, J. M., Paynter, D., Silvers, L. G., Zhou, C., Mauritsen, T., et al. (2018). Accounting for changing temperature patterns increases historical estimates of climate sensitivity. *Geophysical Research Letters*, *45*(16), 8490–8499.

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.

Andrews, T., Smith, C. J., Myhre, G., Forster, P. M., Chadwick, R., & Ackerley, D. (2021). Effective radiative forcing in a GCM with fixed surface temperatures. *Journal of Geophysical Research: Atmospheres*, *126*(4), e2020JD033880.

Andrews, T., & Webb, M. J. (2018). The dependence of global cloud and lapse rate feedbacks on the spatial structure of tropical Pacific warming. *Journal of Climate*, *31*(2), 641–654.

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.

Bloch-Johnson, J., Pierrehumbert, R. T., & Abbot, D. S. (2015). Feedback temperature dependence determines the risk of high warming. *Geophysical Research Letters*, *42*(12), 4973–4980.

Bloch-Johnson, J., Rugenstein, M., Stolpe, M. B., Rohrschneider, T., Zheng, Y., & Gregory, J. M. (2021). Climate sensitivity increases under higher CO_2 levels due to feedback temperature dependence. *Geophysical Research Letters*, *48*(4), e2020GL089074.

Caballero, R., & Huber, M. (2013). State-dependent climate sensitivity in past warm climates and its implications for future climate projections. *Proceedings of the National Academy of Sciences*, *110*(35), 14162–14167.

Ceppi, P., & Gregory, J. M. (2017). Relationship of tropospheric stability to climate sensitivity and Earth's observed radiation budget. *Proceedings of the National Academy of Sciences*, *114*(50), 13126–13131.

Coats, S., & Karnauskas, K. B. (2017). Are simulated and observed twentieth century tropical Pacific sea surface temperature trends significant relative to internal variability? *Geophysical Research Letters*, *44*(19), 9928–9937.

Dessler, A. E., Mauritsen, T., & Stevens, B. (2018). The influence of internal variability on Earth's energy balance framework and implications for estimating climate sensitivity. *Atmospheric Chemistry and Physics*, *18*(7), 5147–5155.

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.

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. (2016). Inference of climate sensitivity from analysis of Earth's energy budget. *Annual Review of Earth and Planetary Sciences*, *44*, 85–106.

Forster, P., Richardson, T., Maycock, A. C., Smith, C. J., Samset, B. H., Myhre, G., et al. (2016). Recommendations for diagnosing effective radiative forcing from climate models for CMIP6. *Journal of Geophysical Research: Atmospheres*, *121*(20), 12–460.

Forster, P., Storelvmo, T., Amour, K., Collins, W., Dufresne, J., Frame, D., et al. (in press). The Earth's energy budget, climate feedbacks, and climate sensitivity. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.

Fueglistaler, S. (2019). Observational evidence for two modes of coupling between sea surface temperatures, tropospheric temperature profile, and shortwave cloud radiative effect in the tropics. *Geophysical Research Letters*, *46*(16), 9890–9898.

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., et al. (2016). The Detection and Attribution Model Intercomparison Project (DAMIP v1. 0) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3685–3697.

Gregory, J. M., Andrews, T., Ceppi, P., Mauritsen, T., & Webb, M. (2020). How accurately can the climate sensitivity to CO_2 be estimated from historical climate change? *Climate Dynamics*, *54*(1), 129–157.

Gregory, J. M., Andrews, T., & Good, P. (2015). The inconstancy of the transient climate response parameter under increasing CO_2. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*(2054), 20140417.

Gregory, J. M., & Forster, P. (2008). Transient climate response estimated from radiative forcing and observed temperature change. *Journal of Geophysical Research: Atmospheres*, *113*(D23).

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., et al. (2004). A new method for diagnosing radiative forcing and climate sensitivity. *Geophysical research letters*, *31*(3).

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.

Grose, M. R., Gregory, J., Colman, R., & Andrews, T. (2018). What climate sensitivity index is most useful for projections? *Geophysical Research Letters*, *45*(3), 1559–1566.

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

Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M., & Rosinski, J. (2008). A new sea surface temperature and sea ice boundary dataset for the Community Atmosphere Model. *Journal of Climate*, *21*(19), 5145–5153.

Lewis, N., & Curry, J. A. (2015). The implications for climate sensitivity of AR5 forcing and heat uptake estimates. *Climate dynamics*, *45*(3), 1009–1023.

Lewis, N., & Curry, J. A. (2018). The impact of recent forcing and ocean heat uptake data on estimates of climate sensitivity. *Journal of Climate*, *31*(15), 6051–6071.

Loeb, N. G., Wang, H., Allan, R. P., Andrews, T., Armour, K., Cole, J. N., et al. (2020). New generation of climate models track recent unprecedented changes in Earth's radiation budget observed by CERES. *Geophysical Research Letters*, *47*(5), e2019GL086705.

Marvel, K., Pincus, R., Schmidt, G. A., & Miller, R. L. (2018). Internal variability and disequilibrium confound estimates of climate sensitivity from observations. *Geophysical Research Letters*, *45*(3), 1595–1601.

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.

Myhre, G., Shindell, D., Bréon, F., Collins, W., Fuglestvedt, J., Huang, J., et al. (2013). Anthropogenic and natural radiative forcing. *Climate Change 2013:*

The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 659–740.

Olonscheck, D., Rugenstein, M., & Marotzke, J. (2020). Broad consistency between observed and simulated trends in sea surface temperature patterns. *Geophysical Research Letters*, *47*(10), e2019GL086773.

Otto, A., Otto, F. E., Boucher, O., Church, J., Hegerl, G., Forster, P. M., et al. (2013). Energy budget constraints on climate response. *Nature Geoscience*, *6*(6), 415–416.

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.

Proistosescu, C., & Huybers, P. J. (2017). Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Science Advances*, *3*(7), e1602821.

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.

Rugenstein, M., Bloch-Johnson, J., Gregory, J. M., Andrews, T., Mauritsen, T., Li, C., et al. (2020). Equilibrium climate sensitivity estimated by equilibrating climate models. *Geophysical Research Letters*, *47*(4), e2019GL083898.

Seager, R., Cane, M., Henderson, N., Lee, D.-E., Abernathey, R., & Zhang, H. (2019). Strengthening tropical Pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. *Nature Climate Change*, *9*(7), 517–522.

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K., Forster, P., Hargreaves, J., et al. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678.

Silvers, L. G., Paynter, D., & Zhao, M. (2018). The diversity of cloud responses to twentieth century sea surface temperatures. *Geophysical Research Letters*, *45*(1), 391–400.

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

Takahashi, C., & Watanabe, M. (2016). Pacific trade winds accelerated by aerosol forcing over the past two decades. *Nature Climate Change*, *6*(8), 768–772.

Watanabe, M., Dufresne, J.-L., Kosaka, Y., Mauritsen, T., & Tatebe, H. (2021). Enhanced warming constrained by past trends in equatorial Pacific sea surface temperature gradient. *Nature Climate Change*, *11*(1), 33–37.

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., et al. (2017). The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6. *Geoscientific Model Development*, *10*(1), 359–384.

Webb, M. J., Lambert, F. H., & Gregory, J. M. (2013). Origins of differences in climate sensitivity, forcing and feedback in climate models. *Climate Dynamics*, *40*(3).

Winton, M., Adcroft, A., Dunne, J., Held, I., Shevliakova, E., Zhao, M., et al. (2020). Climate sensitivity of GFDL's CM4.0. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001838.

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

Zhou, C., Zelinka, M. D., & Klein, S. A. (2016). Impact of decadal cloud variations on the Earth's energy budget. *Nature Geoscience*, *9*(12), 871–874.

Supporting Information for "Equilibrium climate sensitivity and Transient Climate Response biased low in historical simulations of CMIP6 models"

Yue Dong[1], Kyle C. Armour[1,2], Cristian Proistosescu[3], Timothy Andrews[4],

David S. Battisti[1], Piers M. Forster[5], David Paynter[6], Christopher J. Smith[5],

Hideo Shiogama[7]

[1]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA

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

[3]Department of Atmospheric Sciences and Department of Geology, University of Illinois at Urbana-Champaign, IL, USA

[4]Met Office Hadley Centre, Exeter, UK

[5]School of Earth and Environment, University of Leeds, UK

[6]NOAA/Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, NJ, USA

[7]National Institute for Environmental Studies, Tsukuba, Japan

Contents of this file

1. Figures S1 to S5
2. Table S1

Corresponding author: Yue Dong, dongy24@uw.edu

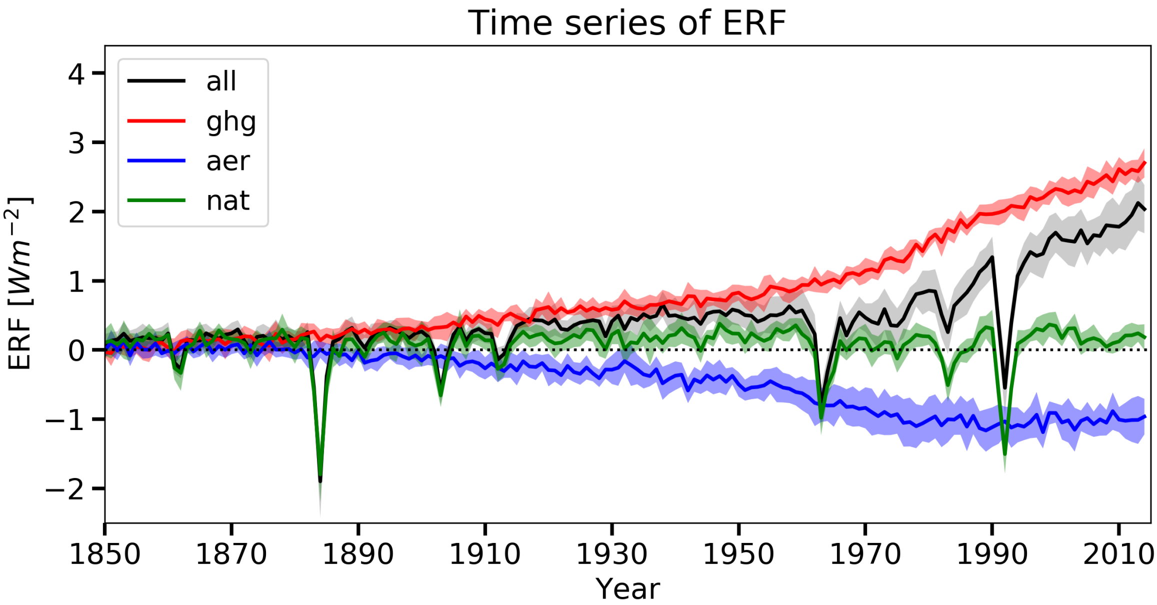

Figure S1. Time series of historical effective radiative forcing estimated from RFMIP simulations. Thick lines denote multi-model means, shadings denote one standard deviation across models.

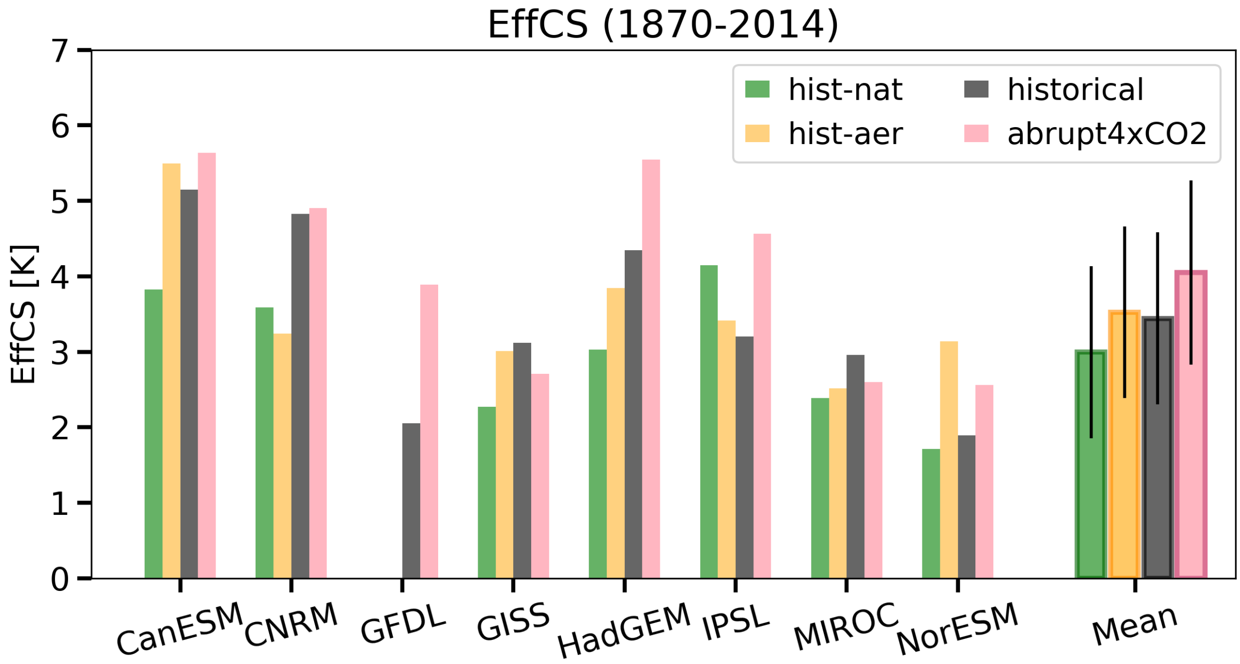

Figure S2. Similar to Figure (1b), except $\mathrm{EffCS_{his}}$ estimates are from historical energy budget constraints within historical non-GHG simulations. Yellow bars denote the values of $\mathrm{EffCS_{his}}$ from *hist-aer* simulations, and green bars denote the values of $\mathrm{EffCS_{his}}$ from *hist-nat* simulations. Note that GFDL-CM4 currently does not have single-forcing historical simulations available.

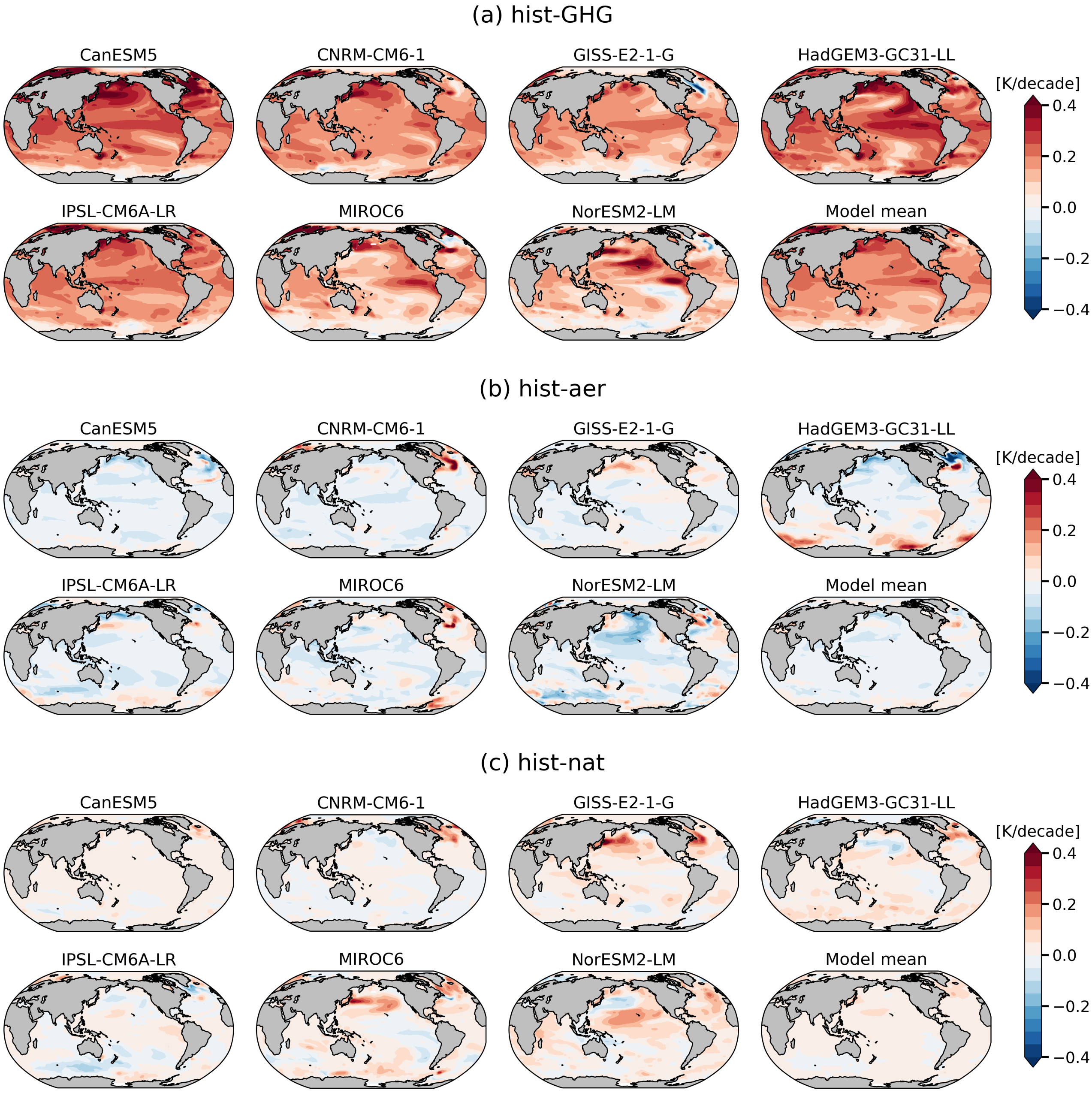

Figure S3. Similar to Figure 2(b), patterns of SST linear trends over 1979 - 2014 from (a)*hist-GHG*, (b)*hist-aer*, (c)*hist-nat* simulations.

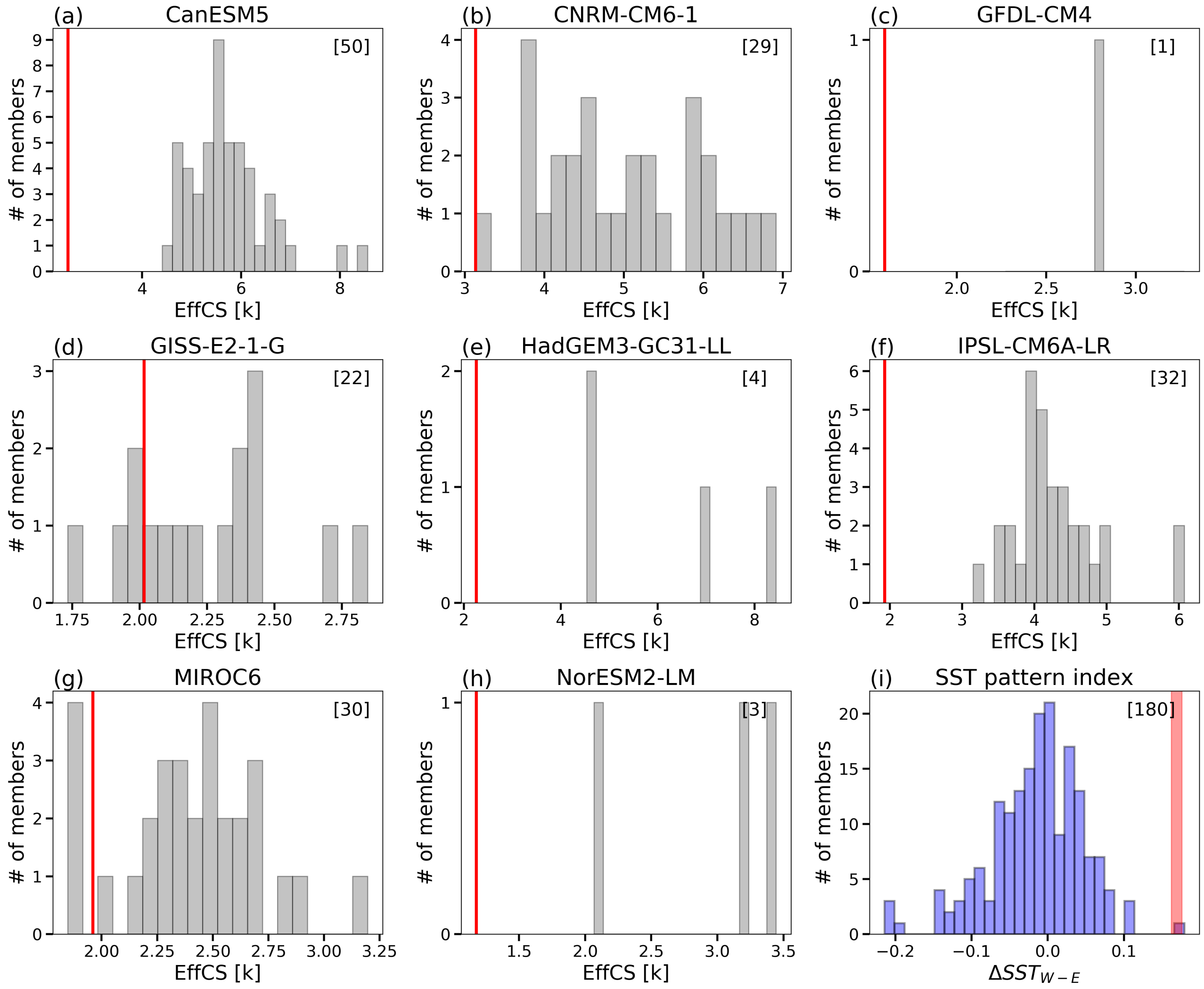

Figure S4. (a - h) $\mathrm{EffCS_{his}}$ from 1979 – 2014 of all available members of *historical* simulations (black bar) and *amip* simulations (red line) for each of the 8 models. (i) The tropical Pacific zonal SST gradient ($\Delta SST_{\text{W-E}}$) over 1979 – 2014 (defined in Watanabe et al., 2021) from all models historical ensemble members (blue bars) and observations (red shading). The observations include 4 datasets: HadISST1 (Rayner et al., 2003), AMIPII (Hurrell et al., 2008), COBE-SST2 (Hirahara et al., 2014), ERSSTv5 (Huang et al., 2017). The red shading denotes the mean $\Delta SST_{\text{W-E}}$ ±one standard deviation across 4 observational datasets. The number in the top right corner in each panel denotes the number of total model ensembles plotted.

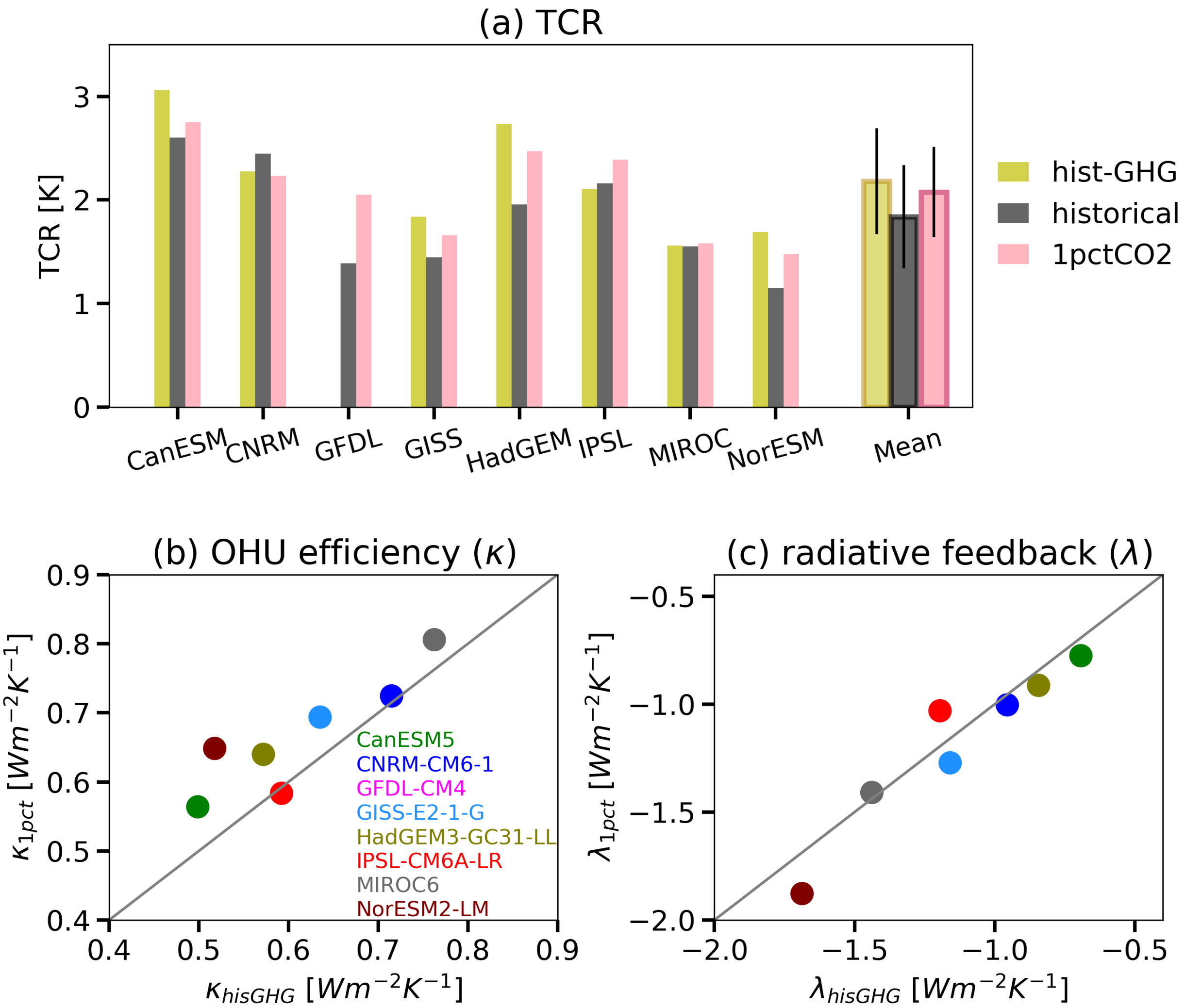

Figure S5. Same as Figure 3, except in (a) yellow bars denote TCR_{his} values from *hist-GHG* simulations, and in (b, c) κ_{his} and λ_{his} values from *hist-GHG* simulations

References

Hirahara, S., Ishii, M., & Fukuda, Y. (2014). Centennial-scale sea surface temperature analysis and its uncertainty. *Journal of Climate*, *27*(1), 57–75.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. (2017). Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *Journal of Climate*, *30*(20), 8179–8205.

Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M., & Rosinski, J. (2008). A new sea surface temperature and sea ice boundary dataset for the Community Atmosphere Model. *Journal of Climate*, *21*(19), 5145–5153.

Rayner, N., Parker, D. E., Horton, E., Folland, C. K., Alexander, L. V., Rowell, D., et al. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, *108*(D14).

Watanabe, M., Dufresne, J.-L., Kosaka, Y., Mauritsen, T., & Tatebe, H. (2021). Enhanced warming constrained by past trends in equatorial Pacific sea surface temperature gradient. *Nature Climate Change*, *11*(1), 33–37.