Haozhe He[1], Brian J. Soden[1], Ryan J. Kramer[2,3]

[1] Rosenstiel School of Marine, Atmospheric and Earth Science, University of Miami, Miami, FL, USA.

[2] Climate and Radiation Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA.

[3] Goddard Earth Science Technology and Research II, University of Maryland at Baltimore County, Baltimore MD, USA.

Corresponding author: Haozhe He (haozhe.he@miami.edu)

Key Points:

- The underestimate of ECS by the Gregory method arises from an underestimate of the radiative forcing.
- The effective feedback parameter estimated from the Gregory method agrees well with the equilibrium feedback parameter.
- Approximately one-third of CMIP6 models fall within the "hot model" category with the true ECS greater than 5K.

Abstract

The most recent generation of climate models exhibits an alarming increase in high climate sensitivity models compared to previous generations. Because the calculation of equilibrium climate sensitivity (ECS) requires simulations of a thousand years or more, most studies estimate ECS using shorter model integrations. However, the most widely used method for estimating ECS from shorter simulations underestimates ECS. Previous studies attributed this underestimate to the time-dependence of climate feedbacks. Here we demonstrate that it actually arises from an underestimate of the radiative forcing. We present a modified method that corrects for this underestimate and is shown to better agree with ECS calculated from "long run", millennium-scale simulations. This method reveals that the actual number of "too hot" models is roughly double that previously diagnosed, with one out of every three CMIP6 climate models having an ECS greater than 5K - the "very likely" upper bound on ECS.

Plain Language Summary

Equilibrium climate sensitivity (ECS) defines the susceptibility of the climate to external forcing. The most recent generation of climate models exhibits an alarming increase in high climate sensitivity models (i.e., ECS > 5K) compared to previous generations. If such high sensitivities are correct, the impacts on civilization would be devastating and keeping global warming below the 2 °C threshold set by the Paris Agreement would require even more rapid reductions in CO_2 emissions than already proposed. Unfortunately, such concerns are based on a methodology that underestimates ECS. In this study, we present a modified method that corrects for this underestimate and reveals that the actual number of "too hot" models is double that previously diagnosed, with

one out of every three climate models having an ECS greater than 5K – the IPCC estimated upper bound on ECS, suggesting a potentially appalling future if without major action to reduce emissions.

1 Introduction

Equilibrium climate sensitivity (ECS) is formally defined as the change in global-mean surface air temperature required to restore radiative equilibrium in response to a doubling of CO_2 and is the most widely used metric to quantify the susceptibility of the climate to external forcing; i.e., ECS $= -F_{2\times}/\,_{eq}$, where $F_{2\times}$ is the effective radiative forcing (ERF) from a doubling of CO_2, and $_{eq}$ is the equilibrium feedback parameter, the efficiency at which radiative equilibrium is restored per unit change in surface air temperature (Rugenstein & Armour, 2021). The $_{eq}$ is calculated from radiative anomalies between two equilibrated climate states and represents the integration of all processes that brought the climate system to the new equilibrated state.

It is already known that climate models from the Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al., 2016) show, on average, a notably higher ECS than CMIP5 (Taylor et al., 2012) models (Zelinka et al., 2020). Most of the increase is attributed to the increase in the radiative feedback parameter, especially a stronger cloud feedback (Zelinka et al., 2020).

The larger ECS also contributes to a "hot model" problem. According to previous estimates, approximately one-sixth (10 out of 58) of CMIP6 models possess an ECS of 5 K or greater, which considerably exceeds the upper limit of the range projected by CMIP5 models and the 'very likely' ECS range of IPCC AR6 report, i.e., 2–5 K (Forster et al., 2021). If such high sensitivities are correct, the impacts on civilization could be devastating and keeping global warming below the 2 °C threshold set by the Paris Agreement (UNFCCC, 2015) would require even more rapid reductions in CO_2 emissions than already proposed.

However, many believe these models are unrealistic and that their projections exaggerate the impact of global warming (Forster et al., 2020; Tokarska et al., 2020; Zhu et al., 2020). To reduce the influence of these "too hot" models, an emerging literature recommends that future projections should exclude models with an ECS that lies outside the estimated likely range (Hausfather et al., 2022).

Ideally, the ECS can be estimated from equilibrium states in climate models forced by an abrupt doubling or quadrupling of CO_2, after millennial-length or "long run" integrations (Rugenstein et al., 2020). Due to the computational burden of such long integrations, only a handful of models have performed such simulations. In practice, the ECS shown in most studies (Andrews et al., 2012, 2019; Flato et al., 2013; Forster et al., 2021; Gettelman et al., 2019; Hausfather et al., 2022; Meehl et al., 2020; Wyser et al., 2020; Zelinka et al., 2020) is extrapolated from a linear regression of global-mean energy imbalance (N) at the top-of-atmosphere (TOA) against surface air temperature anomalies (T) for the first 150 years of abrupt-4×CO2 simulations (Gregory et al., 2004; referred

to as Gregory method hereafter). It has been established that the Gregory method generally underestimates the true ECS by around 15% compared to more accurate "long run" integrations (Dunne et al., 2020; Rugenstein et al., 2020; Rugenstein & Armour, 2021). Taking a CMIP6 model as an example (Figure S1), the ECS of CESM2 has previously been reported as 5.3 K (Gettelman et al., 2019), obtained using the Gregory method. However, the surface air temperature anomaly at year 999 after a quadrupling of CO_2 is around 12 K, suggesting the true ECS from a doubling of CO_2 is at least 6 K and likely greater, since the model is still far from reaching equilibrium at year 999. Previous studies attribute the ECS underestimate to a time dependence of climate feedbacks (Dunne et al., 2020; Rugenstein et al., 2020), owing to time-evolving surface warming patterns (Andrews et al., 2015; Armour et al, 2013; Dong et al., 2020), and nonlinear state-dependence of radiative feedbacks (Bloch-Johnson et al., 2021; Caballero & Huber 2013).

In this study, we demonstrate that the ECS underestimate actually arises from its underestimate of the ERF, which is 15% lower on average compared to the ERF diagnosed from the recommended method (Forster et al, 2016) of using fixed-SST simulations and correcting for land-warming effects. In contrast, the effective feedback parameter estimated from the Gregory method agrees well with the equilibrium feedback parameter obtained from fixed-SST ERFs and "long run" coupled integrations. We develop a modified method to correct for the ERF bias that provides estimates of ECS in much better agreement with "long run" integrations. Because it only uses the standard 150-yr-long abrupt-4×CO2 integrations, this modified method can still be applied to the full suite of CMIP6 models. When applied to the CMIP6 models, we find that approximately one-third of models fall within the "hot model" category (i.e., ECS $>$ 5K). The preponderance of such high sensitivities has important implications both for the interpretation and application of model projections, as well as for mitigating climate change under such high sensitivity scenarios.

2 Materials and Methods

2.1 LongRunMIP and CMIP5 / 6 model simulation output

The millennial-length abrupt-4×CO2 simulations of 11 models taking part in the LongRunMIP (Rugenstein et al., 2019) and two CMIP6 models (999-yr-long simulation of CESM2 and 900-yr-long simulation of IPSL-CM6A-LR) are comprehensively used in proposing and evaluating the modified method. Meanwhile, the fixed-SST simulations [sstClim / sstClim4×CO2 of CMIP5 and piClim-control / piClim-4×CO2 of Radiative Forcing Model Intercomparison Project (Pincus et al., 2016) endorsed by CMIP6] are used to provide the most accurate ERF estimations. In addition, the piControl and abrupt-4×CO2 simulations of CMIP5 / 6 models, including 29 CMIP5 models and 58 CMIP6 models, are used to revisit the differences in both the effective climate sensitivity and equilibrium climate sensitivity between CMIP5 and CMIP6 ensembles.

2.2 Land-warming correction of the ERF

The two land-warming correction methods (Forster et al, 2016) recommended by Andrews et al. (2021) are used: 1. A modified version of the tropospheric and surface correction method (TROP) by Tang et al. (2019) that assumes various "adjustments" from the fixed-SST experiment are related to the land warming; 2. The surface temperature method (SFC) applied in Smith et al. (2020a) assumes that the radiative effect of surface temperature change (i.e., the surface Planck "adjustment") is the only radiative effect of land warming in fixed-SST experiments. The only modification of the former method to that of Tang et al. (2019) is that all relative humidity (RH) radiative responses are ascribed to rapid adjustments, since the RH feedback is close to zero (He et al., 2021) and furthermore as the limited land warming of order -1 (K) is in no way able to trigger the RH radiative response of order -1 (W m^{-2}).

Here, all the radiative responses which are assumed to be related to land warming are calculated with the radiative kernel methods, following Soden et al. (2008) and Held and Shell (2012). To minimize the uniqueness of the single radiative transfer model used to derive radiative kernels, multi-kernels ensemble mean radiative responses, obtained from results using CESM1-CAM5 (Pendergrass et al., 2018), GFDL (Soden et al., 2008), HadGEM3 (Smith et al., 2020b), and RRTM-ERAI (Huang et al., 2016) radiative kernels, are used. The ERF from method TROP is the difference between the time-mean global-mean energy imbalance at the TOA and the sum of surface Planck responses, RH-fixed Planck responses and surface albedo responses, while the ERF from method SFC is the difference between the time-mean global-mean energy imbalance and surface Planck responses only. Most of $ERF_{\text{fixed-SST}}$ diagnoses shown in the manuscripts are the results obtained from method TROP, except for those shown in Figures S3b and S4.

2.3 The equilibrium temperature change

Following Rugenstein et al. (2020), the equilibrium temperature change to a quadrupling of CO_2 is determined by regressing global-mean annual-mean energy imbalance at the TOA against surface air temperature anomaly over the final 15% of warming period of more than 1000-year simulation, although this method is also adopted for 999-yr-long simulation of CESM2 and 900-yr-long simulation of IPSL-CM6A-LR.

3 A modified method

Here we use the recommended estimates of ERF obtained from fixed-SST experiments that correct for the effects of land warming ($ERF_{\text{fixed-SST}}$) and equilibrium temperature change (T_{eq}) obtained from a large collection of millennial-length abrupt-4×CO2 simulations from LongRunMIP to investigate the cause for the underestimated ECS by the Gregory method. We then propose a simple modification to correct for this bias that is shown to provide better agreement with the best estimates of ECS obtained from "long run" integrations.

Consider Figure 1 which shows a Gregory plot of N vs. T for the 999-yr-long abrupt-4×CO2 simulation of CESM2. The traditional Gregory method

regresses N on T using the first 150 years (solid black line) with the slope providing an estimate of the climate feedback parameter ($_{1–150}$) and the y-intercept providing an estimate of the ERF ($ERF_{1–150}$). Additional regressions are shown for select time periods over the 999-yr integration (colored lines). As expected, the radiative feedback parameter increases for each sub-period as the climate approaches the equilibrium state (Andrews et al., 2015; Ceppi & Gregory, 2017; Dong et al., 2020; Geoffroy et al., 2013; Proistosescu & Huybers, 2017; Rugenstein et al., 2020).

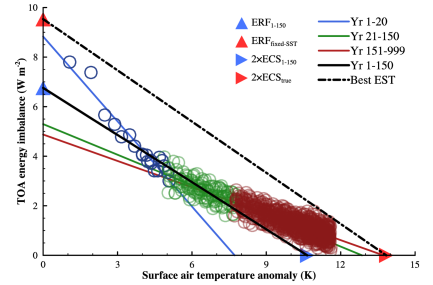

Figure 1. Gregory plot for CESM2: Global-mean annual-mean energy imbalance at the TOA against surface air temperature anomaly from its 999-yr-long abrupt-4×CO2 simulation. Open circles in blue, green, and brown represent each year during years 1–20, 21–150 and 151–999, respectively. The corresponding lines show regression fits for the three periods. The black solid line indicates the regression fit for years 1–150 (or the Gregory method), while the black dashed line is made by using the recommended ERF diagnosis (Forster et al, 2016) and inferred equilibrium temperature change (Rugenstein et al., 2020). The slope of each solid line represents the radiative feedback parameter during each period, while the slope of black dashed line represents the equilibrium feedback parameter, derived by $_{eq} = -ERF_{fixed\text{-}SST}/\ T_{eq}$.

However, the slope from years 1–150 ($_{1–150}$) agrees very well with the best estimate of the equilibrium feedback parameter, $_{eq} = -ERF_{fixed\text{-}SST}/\ T_{eq}$ (black dashed line). This is true in all 7 models for which both fixed-SST ERF and "long run" coupled abrupt-4×CO2 simulations are available (Figures S2a–g) and suggests that the underestimate of ECS from the Gregory method is not attributable to errors in estimating the feedback parameter .

In contrast, the ERF estimated from the Gregory method (upward blue triangle in Figure 1) is ~2.5 W m^{-2} smaller than the more accurate estimate of $ERF_{fixed\text{-}SST}$ (upward red triangle). Similar low biases in $ERF_{1–150}$ relative to $ERF_{fixed\text{-}SST}$ are noted for the other 6 models (Figures S2b–g), indicating a systematic low bias in ERF estimated using the Gregory method. This suggests that it is the bias in $ERF_{1–150}$, not $_{1–150}$, that causes the low bias in ECS from the Gregory method.

While all CMIP5 / 6 models provide the standard 150-yr-long abrupt-4×CO2 simulations necessary to derive $_{1–150}$ from the Gregory method, only a small subset of them also provide the necessary simulations to diagnose $ERF_{fixed\text{-}SST}$. Below we outline a modification to the Gregory method that enables a more accurate estimation of ERF, and thus ECS, from standard 150-yr-long abrupt-4×CO2 simulations. Since the climate feedback parameter becomes less negative as the climate approaches equilibrium (Rugenstein et al., 2020), we can expect more accurate ERF estimations from regressions on shorter periods. Following Qin et al. (2022), we calculate regressed ERFs using all possible windows of 10–30 yr duration with starting at year 1 and then diagnose the correlation and root-mean-square deviation (RMSD) between every regressed ERF value and corresponding recommended $ERF_{fixed\text{-}SST}$ diagnosis (Figure S3a).

Although ERFs from first 10-yr and 14-yr regressions have the lowest RMSD and the highest correlation with the $ERF_{fixed\text{-}SST}$, they are not much different than the ERF diagnosed by regressing years 1–20 ($ERF_{1–20}$) (Figure S3a), a more conventional record length used previously by Forster et al. (2016). We therefore use the $ERF_{1–20}$ hereafter. Compared to the uniform ~15% underestimation of ERF by $ERF_{1–150}$, the first 20-yr regressions provide the ERF estimations much closer to the $ERF_{fixed\text{-}SST}$ in almost all models (Figures 2a, 2b and S4). We emphasize that the results discussed below are insensitive to either the above-mentioned first 10–30 yr time window used in the regression or the method chosen to correct for land-warming effect when deriving $ERF_{fixed\text{-}SST}$ (Figure S3).

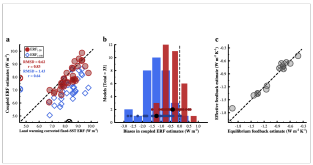

Figure 2. Evaluation of coupled ERF and feedback parameter estimates. (a) Comparisons between coupled ERF estimates using years 1–150 and 1–20 and the $ERF_{fixed\text{-}SST}$ diagnosis, obtained from sstClim4×CO2 and piClim-4×CO2 simulations. (b) Histogram displaying number of models that fall within 0.5 W m^{-2} bins of biases in coupled ERF estimates, relative to the $ERF_{fixed\text{-}SST}$ diagnosis. The markers which lie on axes in (a) are multi-model ensemble means of corresponding estimates. The thin and thick dots in (b) indicate the biases in coupled ERF estimates of each model and corresponding multi-model ensemble means. Note the $ERF_{fixed\text{-}SST}$ diagnoses shown here are obtained using the tropospheric and surface correction method (TROP). (c) A comparison between the estimates of the effective feedback parameter ($_{1–150}$) and the equilibrium feedback parameter ($_{eq}$). The $_{eq}$ is derived by $_{eq} = -ERF/\ T_{eq}$, where the ERF is either $ERF_{fixed\text{-}SST}$ or $ERF_{1–20}$, depending on the availability of fixed-SST ERF experiments. The solid circles with cross inside in (c) indicate the models with fixed-SST ERF experiments available.

As expected, the effective feedback estimates from standard 150-yr regressions ($_{1–150}$) agree well with the equilibrium feedback estimates (Figures 2c and S2), derived by $_{eq} = -ERF/\ T_{eq}$, where the ERF is either $ERF_{fixed\text{-}SST}$ or $ERF_{1–20}$, depending on the availability of fixed-SST ERF experiments. In this case, we propose a modified method to estimate the true ECS in short-integration simulations using the $ERF_{1–20}$ and $_{1–150}$, $ECS_{Modified} = -ERF_{1–20}/2\ _{1–150}$.

Compared to the underestimated ECS by the original Gregory method, the modified method replicates the true ECS diagnosed from the "long run" coupled simulations (Figure 3), providing an easy and effective way to diagnose the true ECS. As expected, the impact of a forcing bias on ECS increases as the climate feedback parameter increases, and is largest for the two models with the largest feedback parameter (i.e., CESM2 and FAMOUS). Even if we exclude the two high ECS models, the ECS estimates from this modified method still have a better performance than the original Gregory estimates.

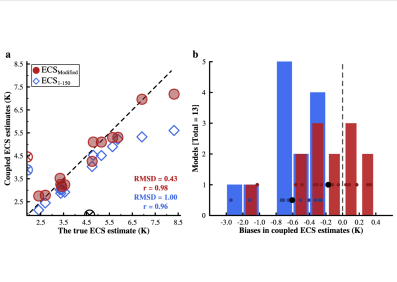

Figure 3. Evaluation of coupled ECS estimates. (a) Comparisons between coupled ECS estimates from the Gregory method ($ECS_{1–150}$) and the modified method ($ECS_{Modified}$) and the true ECS estimate using Rugenstein et al. (2020) with millennial-length abrupt-4×CO2 simulations. (b) Histogram displaying number of models that fall within 0.2 K bins (except for first two bins with 1.0 K) of biases in coupled ECS estimates, relative to the true ECS estimate. The markers which lie on axes in a are multi-model ensemble means of corresponding estimates. The thin and thick dots in (b) indicate the biases in coupled ECS estimates of each model and corresponding multi-model ensemble means.

There have been several other previously proposed regression-based methods that attempt to provide more accurate estimates of ECS by capturing the slow-pattern responses of climate models; e.g., using years 21–150 (Andrews et al., 2015) and years 51–150 (Dunne et al., 2020; Winton et al., 2013). We provide a comprehensive evaluation of our proposed method along with each of these approaches in Table S1. The modified method is superior in all respects to the previously proposed methods. The bias in ECS is reduced to $-2.5\% \pm 7.9\%$ with the modified method, close to the highest accuracy we can acquire from 400-yr-long coupled simulations by regressing years 100–400 ($-2.2\% \pm 2.8\%$; Rugenstein & Armour, 2021).

Since many modeling centers are unable to provide millennial climate perturbation simulations for the ECS evaluations due to their large computational expense, our proposed method provides an easy and effective way to diagnose the true ECS using the standard, widely-available 150-yr-long abrupt-4×CO2 simulation. Note that the method does not improve the estimate of ECS for fewer than a tenth of the climate models (BNU-ESM and CNRM-CM5 of CMIP5; CNRM-ESM2-1, CNRM-CM6-1-HR, GISS-E2-2-G, GISS-E2-2-H and MIROC-ES2L of CMIP6). In contrast to typical behavior, all of these models have a

decrease in the radiative feedback between years 1–20 and 21–150, so barely any mitigation can be expected in the ERF underestimations, even with the years 1–20 regression (Tables S2 and S3). A preliminary analysis suggests that this decrease in the radiative feedback parameter is due to overly strong variation in radiative responses at interannual timescales. In the case of CNRM-ESM2-1, the CO_2 concentrations are relaxed towards the quadrupling level below 560 hPa and then allowed to propagate throughout the atmosphere (Michou et al, 2020), therefore taking around 15 years to reach an approximate uniform atmospheric concentration (Smith et al., 2020a), leading to a time-evolving stratospheric temperature adjustment and contaminating the initial feedback estimations. Therefore, these models are dismissed in the following discussions (Figures 4, S5 and S6), although our conclusions are insensitive to whether these models are retained in the analysis or not (results including these models shown in Figures S7–S9).

4 Analysis of ECS in CMIP5 and CMIP6

Equipped with both the Gregory method and the modified method, we revisit the effective climate sensitivity (referred to as EffCS hereafter for distinction) from the traditional Gregory method and the best estimate of ECS for both CMIP5 and CMIP6 ensembles computed using the modified method proposed here (referred to hereafter as ECS) in Figure 4. CMIP6 models not only possess higher EffCS and ECS values than CMIP5 models, but have a much larger inter-model spread in both the EffCS and ECS than CMIP5 models. The larger uncertainties arise from the larger uncertainty in the $_{1–150}$, although more consistent ERF estimates are obtained in CMIP6 models (Figure S5).

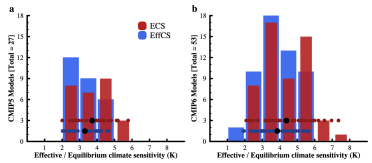

Figure 4. Climate sensitivity distributions. Histograms displaying number of (a) CMIP5 and (b) CMIP6 models that fall within 1 K EffCS and ECS bins. The thin and thick dots indicate the EffCS and ECS values of each model and corresponding multi-model ensemble means.

In contrast to previous statements based on the EffCS (Forster et al., 2021; Meehl et al., 2020; Zelinka et al., 2020), our results show there are also "hot models" with ECS greater than 5K in CMIP5 and that more than one-third of CMIP6 models fall within the "hot model" category. This is roughly double the previous estimate of high ECS models and further amplifies concerns surrounding the tendency for the CMIP6 generation of climate models to be "too hot". Hausfather et al. (2022) argue that such high ECS models are inconsistent with the observational record and bias both the multi-model ensemble mean and inter-model uncertainty range in CMIP6. They caution that such "hot models" should be restricted from conventional warming projections or limited to assessments that focus on the most extreme projections. These results also bring into question arguments that the "hot tail" models are inconsistent with the historical record (Hausfather et al., 2022), since the number of such models is now twice the previous estimate.

Since the underestimate of ERF is largely responsible for the bias in ECS, this suggests that the higher ERF in CMIP6 (Smith et al., 2020a; Figure S5) contributes to the greater number of "hot models" in CMIP6. Therefore, we diagnose the relative importance of ERF in causing both the multi-model ensemble mean EffCS and ECS to increase. Following Zelinka et al. (2020), we shift the ERF of each CMIP5 model by the multi-model mean difference in ERF between CMIP6 and CMIP5 and do the same transformation but in reverse for CMIP6 models, and then average the two results (Figure S6). Shifting the $\mathrm{ERF}_{1\text{–}150}$ ($\mathrm{ERF}_{1\text{–}20}$) changes the EffCS (ECS) by 16.8% (19.5%) of the actual CMIP5-to-CMIP6 mean difference in the EffCS (ECS), respectively. In contrast, the change in EffCS (ECS) by shifting the $_{1\text{–}150}$ for the full suite of CMIP6 models is 73.5% (71.5%) of the actual CMIP5-to-CMIP6 mean EffCS (ECS) difference. This is consistent with the results of Zelinka et al. (2020), confirming the dominant role of the $_{1\text{–}150}$ change in causing the higher mean EffCS and ECS in CMIP6 ensemble. However, if one uses the underestimated $\mathrm{ERF}_{1\text{–}150}$ from the Gregory method instead of the more accurate $\mathrm{ERF}_{1\text{–}20}$, then only 10 out of the actual 18 "hot models" in CMIP6 with ECS $>$ 5K could be identified. This again underscores the importance of accurately calculating radiative forcing in ECS estimates, especially at the high end of the ECS spectrum, further emphasizing the importance of the available $\mathrm{ERF}_{\text{fixed-SST}}$ diagnoses (Pincus et al., 2016).

5 Conclusions

In this study, we demonstrate that the ECS underestimate actually arises from its underestimate of the ERF, which is 15% lower on average compared to the ERF diagnosed from the recommended method (Forster et al, 2016) of using fixed-SST simulations and correcting for land-warming effects. In contrast, the effective feedback parameter estimated from the Gregory method agrees well with the equilibrium feedback parameter obtained from fixed-SST ERFs and "long run" coupled integrations, suggesting the underestimate of ECS from the Gregory method is not attributable to errors in estimating the feedback

parameter . Therefore, we propose a modified method to correct for the ERF bias that provides estimates of ECS in much better agreement with "long run" integrations. The modified method is superior in all respects to the previously proposed methods that attempt to provide more accurate estimates of ECS by capturing the slow-pattern responses of climate models. The ECS estimates by the modified method are close to the highest accuracy we can acquire from 400-yr-long coupled simulations. Because it only uses the standard 150-yr-long abrupt-4×CO2 integrations, this modified method can still be applied to the full suite of CMIP6 models, providing an easy and effective way to diagnose the true ECS. When applied to the CMIP6 models, we find that the actual number of "too hot" models is roughly double that previously diagnosed, with one out of every three CMIP6 climate models having an ECS greater than 5K - the "very likely" upper bound on ECS. The preponderance of such high sensitivities has important implications both for the interpretation and application of model projections, as well as for mitigating climate change under such high sensitivity scenarios.

Acknowledgments

We thank Dr. Maria Rugenstein for comments and providing LongRunMIP data. H.H., B.J.S. and R.J.K. were supported by NOAA Award NA18OAR4310269. H.H. and B.J.S. were also supported by NOAA Award NA21OAR4310351. R.J.K. was supported by NASA grant number 80NSSC21K1968 as well.

Competing Interests

Authors have no competing interests.

Data Availability Statement

The millennial-length abrupt-4×CO2 simulations from LongRunMIP are available at https://data.iac.ethz.ch/longrunmip/. The CMIP6 data are available at https://esgf-node.llnl.gov/search/cmip6/ while CMIP5 data are available at https://esgf-node.llnl.gov/projects/cmip5/. The CMIP6 / 5 models used in this work are listed in Tables S4 and S5.

References

Andrews, T., Andrews, M. B., Bodas-Salcedo, A., Jones, G. S., Kulhbrodt, T., Manners, J., & Tang, Y. (2019). Forcings, feedbacks and climate sensitivity in HadGEM3-GC3.1 and UKESM1. *Journal of Advances in Modeling Earth Systems*, 11. https://doi.org/10.1029/2019MS001866

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

Andrews, T., Gregory, J. M., Webb, M. J., & Taylor, K. E. (2012). Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate

models. *Geophysical Research Letters*, 39, L09712. https://doi.org/10.1029/2012GL051607

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. https://doi.org/10.1029/2020jd033880

Armour, K. C., Bitz, C. M., & Roe, G. H. (2013). Time-varying climate sensitivity from regional feedbacks. *Journal of Climate*, 26(13), 4518–4534. https://doi.org/10.1175/jcli-d-12-00544.1

Bloch-Johnson, J., Rugenstein, M., Stolpe, M. B., Rohrschneider, T., Zheng, Y., Gregory, J. M. (2021). Climate sensitivity increases under higher CO2 levels due to feedback temperature dependence. *Geophysical Research Letters*, 48(4), e2020GL089074. https://doi.org/10.1029/2020GL089074

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 of the United States of America*, 110(35), 14162–14167. https://doi.org/10.1073/pnas.1303365110

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 of the United States of America*, 114(50), 13126–13131. https://doi.org/10.1073/pnas.1714308114

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

Dunne, J. P., Winton, M., Bacmeister, J., Danabasoglu, G., Gettelman, A., Golaz, J.-C., et al. (2020). Comparison of equilibrium climate sensitivity estimates from Slab ocean, 150-year, and longer simulations. *Geophysical Research Letters*, 47(16), e2020GL088852. https://doi.org/10.1029/2020gl088852

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5), 1937–1958, https://doi:10.5194/gmd-9-1937-2016

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., & Zhang, J. L. (2014). Evaluation of climate models. In Evaluation of climate models. Climate change 2013 - The physical science basis: Working group I contribution to the fifth assessment report of the intergovernmental panel on climate change (pp. 741–866). Cambridge: Cambridge University Press.

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,

7–10. https://doi.org/10.1038/s41558-019-0660-0

Forster, P. M., Richardson, T., Maycock, A. C., Smith, C. J., Samset, B. H., Myhre, G., & Schulz, M. (2016). Recommendations for diagnosing effective radiative forcing from climate models for CMIP6. *Journal of Geophysical Research: Atmospheres*, 121, 12,460–12,475. https://doi.org/10.1002/2016JD025320

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J., Frame, D., et al. (2021). The Earth's energy budget, climate feedbacks, and climate sensitivity. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, 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. Cambridge University Press.

Geoffroy, O., Saint-Martin, D., Bellon, G., Voldoire, A., Olivié, D., & Tytéca, S. (2013). 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. https://doi.org/10.1175/JCLI-D-12-00196.1

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., & Mills, M. J. (2019). High climate sensitivity in the community Earth system model version 2 (CESM2). *Geophysical Research Letters*, 46, 8329–8337. https://doi.org/10.1029/2019GL083978

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, L03205. http://doi.wiley.com/10.1029/2003GL018747

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., & Zelinka, M. D. (2022). Climate simulations: recognize the 'hot model' problem. *Nature*, 605, 26-29. doi: https://doi.org/10.1038/d41586-022-01192-2

He, H., Kramer, R. J., & Soden, B. J. (2021). Evaluating observational constraints on intermodel spread in cloud, temperature, and humidity feedbacks. *Geophysical Research Letters*, 48(17). https://doi.org/10.1029/2020GL092309

Held, I. M., & Shell, K. M. (2012). Using relative humidity as a state variable in climate feedback analysis. *Journal of Climate*, 25(8), 2578–2582. https://doi.org/10.1175/JCLI-D-11-00721.1

Huang, Y., Tan, X., & Xia, Y. (2016). Inhomogeneous radiative forcing of homogeneous greenhouse gases. *Journal of Geophysical Research: Atmospheres*, 121, 2780–2789. https://doi.org/10.1002/2015JD024569

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., et al. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, 6(26), eaba1981. https://doi.org/10.1126/sciadv.aba1981

Michou, M., Nabat, P., Saint-Martin, D., Bock, J., Decharme, B., Mallet, M., et al. (2020). Present-day and historical aerosol and ozone characteristics in CNRM CMIP6 simulations. *Journal of Advances in Modeling Earth Systems*, 12, e2019MS001816. https://doi.org/10.1029/2019MS001816

Pendergrass, A. G., Conley, A., & Vitt, F. M. (2018). Surface and top-of-atmosphere radiative feedback kernels for CESM-CAM5. *Earth System Science Data*, 10(1), 317–324. https://doi.org/10.5194/essd-10-317-2018

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. https://doi.org/10.5194/gmd-9-3447-2016

Proistosescu, C., & Huybers, P. J. (2017). Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Science Advances*, 3(7), e1602821. https://doi.org/10.1126/sciadv.1602821

Qin, Y., Zelinka, M. D., & Klein, S. A. (2022). On the correspondence between atmosphere-only and coupled simulations for radiative feedbacks and forcing from CO2. *Journal of Geophysical Research: Atmospheres*, 127, e2021JD035460. https://doi.org/10.1029/2021JD035460

Rugenstein, M. A. A., & Armour, K. C. (2021). Three flavors of radiative feedbacks and their implications for estimating equilibrium climate sensitivity. *Geophysical Research Letters*, 48, e2021GL092983. https://doi.org/10.1029/2021GL092983

Rugenstein, M., Bloch-Johnson, J., Abe-Ouchi, A., Andrews, T., Beyerle, U., Cao, L., Chadha, T., Danabasoglu, G., Dufresne, J-L., Duan, L., Foujols, M-A., Frölicher, T., Geoffroy, O., Gregory, J., Knutti, R., Li, C., Marzocchi, A., Mauritsen, T., Menary, M., Moyer, E., Nazarenko, L., Paynter, D., Saint-Martin, D., Schmidt, G. A., Yamamoto, A., & Yang, S. (2019). LongRunMIP–motivation and design for a large collection of millennial-length AO-GCM simulations. *Bulletin of the American Meteorological Society*, 100(12), 2551–2570. https://doi.org/10.1175/BAMS-D-19-0068.1

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. https://doi.org/10.1029/2019gl083898

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., Boucher, O., Dufresne, J.-L., Nabat, P., Michou, M., Yukimoto, S., Cole, J., Paynter, D., Shiogama, H., O'Connor, F. M., Robertson, E., Wiltshire, A., Andrews, T., Hannay, C., Miller, R., Nazarenko, L., Kirkevåg, A., Olivié, D., Fiedler, S., Pincus, R., & Forster, P. M. (2020a). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, 20, 9591–9618. https://doi.org/10.5194/acp-20-9591-2020

Smith, C. J., Kramer, R. J., & Sima, A. (2020b). The HadGEM3-GA7.1 radiative kernel: The importance of a well-resolved stratosphere. *Earth System Science Data*, 12(3), 2157–2168. https://doi.org/10.5194/essd-12-2157-2020

Soden, B. J., Held, I. M., Colman, R., Shell, K. M., Kiehl, J. T., & Shields, C. A. (2008). Quantifying climate feedbacks using radiative kernels. *Journal of Climate*, 21, 3504–3520. https://doi.org/10.1175/2007JCLI2110.1

Tang, T., Shindell, D., Faluvegi, G., Myhre, G., Olivié, D., Voulgarakis, A., Kasoar, M., Andrews, T., Boucher, O., Forster, P. M., Hodnebrog, Ø., Iversen, T., Kirkevåg, A., Lamarque, J. F., Richardson, T., Samset, B. H., Stjern, C. W., Takemura, T., & Smith, C. (2019). Comparison of effective radiative forcing calculations using multiple methods, drivers, and models. *Journal of Geophysical Research: Atmospheres*, 124, 4382–4394. https://doi.org/10.1029/2018JD030188

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93, 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances*, 6(12). https://doi.org/10.1126/sciadv.aaz9549

UNFCCC (2015). Adoption of the Paris Agreement Tech. Rep. FCCC/CP/2015/L.9/Rev.1. https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf

Winton, M., Adcroft, A., Griffies, S. M., Hallberg, R., Horowitz, L. W., & Stouffer, R. J. (2013). Influence of ocean and atmosphere components on simulated climate sensitivities. *Journal of Climate*, 26(1), 231–245. https://doi.org/10.1175/JCLI-D-12-00121.1

Wyser, K., Noije, T. V., Yang, S., Hardenberg, J. V., O'Donnell, D., Döscher, R., et al. (2019). On the increased climate sensitivity in the EC-Earth model from CMIP5 to CMIP6. *Geoscientific Model Development*, 13, 3465–3474. https://doi.org/10.5194/gmd-13-3465-2020

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., & Taylor, K. E. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, 47, e2019GL085782. https://doi.org/10.1029/2019GL085782

Zhu, J., Poulsen, C. J., & Otto-Bliesner, B. L. (2020). High climate sensitivity in CMIP6 model not supported by paleoclimate. *Nature Climate Change*, 10, 378–379. https://doi.org/10.1038/s41558-020-0764-6