https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Standardising the "Gregory method" for calculating equilibrium climate sensitivity

Anna Zehrung[1], Andrew D. King[1,2], Zebedee Nicholls[1,3,4], Mark D. Zelinka[5], Malte Meinshausen[1,4]

[1] School of Geography, Earth and Atmospheric Sciences, The University of Melbourne, Melbourne, 3010, Australia
[2] Australian Research Council Centre of Excellence for Weather of the 21st Century, Clayton, 3800, Australia
[3] International Institute for Applied Systems Analysis (IIASA), Schloßplatz 1, 2361 Laxenburg, Austria
[4] Climate Resource, Melbourne, 3000, Australia
[5] Lawrence Livermore National Laboratory, Livermore, CA, 94551, USA

Correspondence to: Anna Zehrung (azehrung@student.unimelb.edu.au)

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Abstract. The equilibrium climate sensitivity (ECS) – the equilibrium global mean temperature response to a doubling of atmospheric CO_2 – is a high-profile metric for quantifying the Earth system's response to human-induced climate change. A widely applied approach to estimating the ECS is the 'Gregory method' (Gregory et al., 2004), which uses an ordinary least squares (OLS) regression between the net radiative flux and surface air temperature anomalies from a 150-year experiment in which atmospheric CO_2 concentrations are quadrupled. The ECS is determined at the point where net radiative flux reaches zero i.e. the system is back in equilibrium. This method has been used to compare ECS estimates across the CMIP5 and CMIP6 ensembles and will likely be a key diagnostic for CMIP7. Despite its widespread application, there is little consistency or transparency between studies in how the climate model data is processed prior to the regression, leading to potential discrepancies in ECS estimates. We identify 20 alternative data processing pathways, varying by different choices in global mean weighting, annual mean weighting, anomaly calculation method, and linear regression fit. Using 41 CMIP6 models, we systematically assess the impact of these choices on ECS estimates. While the inter-model ECS range is insensitive to the data processing pathway, individual models exhibit notable differences. Approximating a model's native grid cell area with cosine of the latitude can decrease the ECS by 11%, and some anomaly calculation methods can introduce spurious temporal correlations in the processed data. Beyond data processing choices, we also evaluate an alternative linear regression method – total least squares (TLS) – which appears to have a more statistically robust basis than OLS. However, for consistency with previous literature, and given physical reasoning suggests that TLS may further reduce the ECS compared to OLS, i.e. make a known bias in the Gregory method worse, we do not feel there is sufficient clarity to recommend a transition to TLS in all cases. To improve reproducibility and comparability in future studies, we recommend a standardised Gregory method: weighting the global mean by cell area, weighting the annual mean by number of days per month, and calculating anomalies by first applying a rolling average to the piControl timeseries then subtracting from the CO_2 quadrupling experiment. This approach implicitly accounts for model drift while reducing noise in the data to best meet the pre-conditions of the linear regression. While CMIP6 results of the multi-model mean ECS appear robust to these processing choices, similar assumptions may not hold for CMIP7, underscoring the need for standardised data preparation in future climate sensitivity assessments.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

1. Introduction

The equilibrium climate sensitivity (ECS) – the steady state global mean temperature response to a doubling of atmospheric CO_2 relative to preindustrial levels – has long been a cornerstone metric for quantifying future climate change (Sherwood et al., 2020). The ECS is commonly estimated using climate models, with Charney et al. (National Research Council, 1979) first proposing an uncertainty range of 1.5 to 4.5 K. The most recent climate model-based estimate places this range at 1.8 to 5.6 K (Zelinka et al., 2020), which was then narrowed to 2 to 5 K based on multiple lines of evidence in the Intergovernmental Panel on Climate Change's (IPCC's) most recent assessment report (Forster et al., 2021).

The most direct method for calculating the ECS involves Earth system models (ESMs) simulating the climate until it reaches thermal equilibrium following a doubling of atmospheric CO_2. However, such an experiment is computationally expensive and it can take multiple millennia of simulation years for a model to equilibrate (Rugenstein et al., 2020). Previously, researchers often relied on the less computationally expensive atmospheric general circulation models coupled with a motionless upper ocean mixed layer, or 'slab ocean'. This approach, however, can affect the ECS estimate because it excludes the effects of thermal inertia and the dynamic and thermodynamic responses of the mixed layer (Boer and Yu, 2003).

Since 2004, fully coupled ESMs have been used instead to estimate the ECS using the "Gregory Method" (Gregory et al., 2004), hereafter GM, which allows for an estimate of the ECS from abrupt CO_2 perturbation simulations that are centuries rather than millennia in duration. We acknowledge that many researchers refer to the metric calculated using the GM as the *effective* climate sensitivity (Caldwell et al., 2016; Dunne et al., 2020; Rugenstein et al., 2020; Rugenstein and Armour, 2021; Sanderson and Rugenstein, 2022; Zelinka et al., 2020), given that the model has not run to true equilibrium. However, we use the term ECS because this study does not consider the potential non-linearities within this method (such as an inconstant feedback parameter).

The GM is based on the zero-dimensional energy balance model, which relates the global mean radiative flux anomaly at the top of the atmosphere, N, to the global mean effective radiative forcing, F, and the global mean radiative response T, where λ is the global mean feedback factor, and ΔT is the temperature change relative to preindustrial levels:

$$N = F - \lambda \Delta T$$

To calculate the ECS, Gregory et al. (2004) take the first 150 years of an abrupt CO_2 quadrupling experiment (abrupt-4xCO_2) relative to the model's preindustrial control experiment (piControl) and calculate an ordinary least squares (OLS) linear regression of annual mean values of N against ΔT. The steady state – equilibrium – is estimated at N=0, i.e. at the T-intercept. The radiative forcing is, according to this model, the N-intercept, and the feedback factor is the (negative) slope of the regression. To express the ECS and radiative forcing relative to a doubling of CO_2 rather than a quadrupling, the T- and N-

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

intercepts are divided by two, as per the original study. Note that scaling by a factor of two implicitly assumes the forcing due to a quadrupling of CO_2 is twice that of a CO_2 doubling, which does not hold if the relationship between forcing and CO_2 concentrations is non-linear (Byrne and Goldblatt, 2014; Etminan et al., 2016; Meinshausen et al., 2020).

The GM is extensively used and cited across literature. It has been applied to assess the fifth and sixth phases of the coupled model intercomparison projects (CMIP) (Andrews et al., 2012; Caldwell et al., 2016; Forster et al., 2013; Zelinka et al., 2020), to investigate ECS state dependence, e.g. (Andrews et al., 2015; Armour et al., 2013; Bloch-Johnson et al., 2021; Dai et al., 2020; Dunne et al., 2020; Mitevski et al., 2023), and as a reference method for comparing other climate sensitivity estimation approaches (Chao and Dessler, 2021; Sherwood et al., 2020).

While the GM calculation is relatively simple, several choices must be made during data preparation. Here we define 'data preparation' as the processing steps applied to the data before performing the N-ΔT regression. Many studies lack transparency regarding these preparatory steps, leading to potential inconsistencies in approach. To our knowledge, no study has to date systematically assessed how different data preparation methods may influence ECS results.

Many researchers do not describe their data preparation entirely, instead presenting the ECS estimate as a direct result of the N-ΔT regression over the 150 year timeseries (Dai et al., 2020; Dessler and Forster, 2018; Geoffroy et al., 2013; Klocke et al., 2013; Lutsko et al., 2022; Meehl et al., 2020; Mitevski et al., 2021, 2023; Ringer et al., 2014; Zhou et al., 2021). Others provide only limited details, such as specifying the model member used (Wang et al., 2025; Zelinka et al., 2013).

For studies that do address data preparation, the focus is typically on anomaly calculations and how to account for model drift. Here, the term anomaly refers to – in its simplest form – the difference between the corresponding abrupt-4xCO_2 and piControl timeseries. However, methods for calculating anomalies vary widely:

Linear trends in the piControl: Some studies apply a linear fit across the portion of the piControl experiment that corresponds with the abrupt-4xCO_2 experiment, subtracting this linear fit from the corresponding abrupt-4xCO_2 timeseries (Andrews et al., 2012; Armour, 2017; Bloch-Johnson et al., 2021; Dong et al., 2020; Flynn and Mauritsen, 2020; Forster et al., 2013).

Rolling or climatological means:

- Some studies apply a 21-year rolling mean to the piControl and subtract the smoothed timeseries from the corresponding abrupt-4xCO_2 timeseries (Caldwell et al., 2016; Eiselt and Graversen, 2023; Po-Chedley et al., 2018; Qu et al., 2018; Zelinka et al., 2020).
- Others calculate a climatological mean of the piControl over a fixed period, such as the full simulation or a specific subset of years, prior to subtracting from the corresponding abrupt-4xCO_2 (Chao and Dessler, 2021; Jain et al., 2021).

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Extended averages: Rugenstein and Armour (2021) subtract the 1000 year average of the piControl timeseries from the abrupt-4xCO_2 timeseries.

Given the lack of transparency and consistency across literature, we aim to investigate how different choices in data preparation may influence the ECS, radiative forcing, and feedback estimates across CMIP6 models - with a particular focus on the ECS values. We identify 10 alternative data processing choices based on the various methods discussed in literature (Fig. 1). Each choice ultimately leads to two ECS estimates, given we also compare the application of two different linear regression fits: OLS, to be consistent with the literature and the original study (Gregory et al., 2004), and total least squares (TLS), given that it is not obvious that all the pre-conditions for OLS are met within the GM.

Notwithstanding the linear fit method, we do not include modifications to the regression itself. Adjustments to the GM regression, such as excluding the initial decades of the timeseries to account for inconstant feedbacks (Andrews et al., 2015; Dunne et al., 2020), including higher order terms in the energy balance equation (Bloch-Johnson et al., 2015), or applying a non-linear ECS scaling factor between abrupt-4xCO_2 and -2xCO_2 experiments (Dai et al., 2020), are already well-documented and widely cited across the literature.

This study does not aim to constrain the ECS ensemble range or address potential non-linearities within the GM calculation. Instead, our focus is on comparing differences in data preparation methods and establishing a standardised GM for future research. This is particularly relevant with the upcoming release of CMIP7 data (Dunne et al., 2024), as ECS calculations will likely be among the first steps taken to compare CMIP7 models and assess how the ensemble aligns with previous CMIP generations.

2. Methods

For our analysis, we compare the effects of data preparation choices across 41 CMIP6 models. To calculate the ECS, the GM requires five variables, the surface air temperature (TAS), top of the atmosphere (TOA) reflected shortwave radiation (rsut), TOA outgoing longwave radiation (rlut), and TOA downward shortwave radiation (rsdt) at monthly timescales, and the atmospheric cell area spatial variable (areacella), for both the abrupt-4xCO_2 and piControl experiments. If a model lacks the required variables or is unavailable for download, it is excluded from the study. For 12 models, cell area data is not available across any experiment, precluding them from this investigation, as grid averaging is one of the processing steps we consider.

We identify four key steps, each with a range of possible choices, which collectively form the basis for 20 data preparation paths we investigate in this study (Fig. 1). While we compare all 20 paths, for simplicity we label only three of them, as the Baseline, Standard, and Alternative paths. These respectively aim to replicate – to the best of our knowledge – the data

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

processing paths described in the original GM study, recent literature (Caldwell et al., 2016; Eiselt and Graversen, 2023; Zelinka et al., 2020), and an alternative anomaly calculation method.

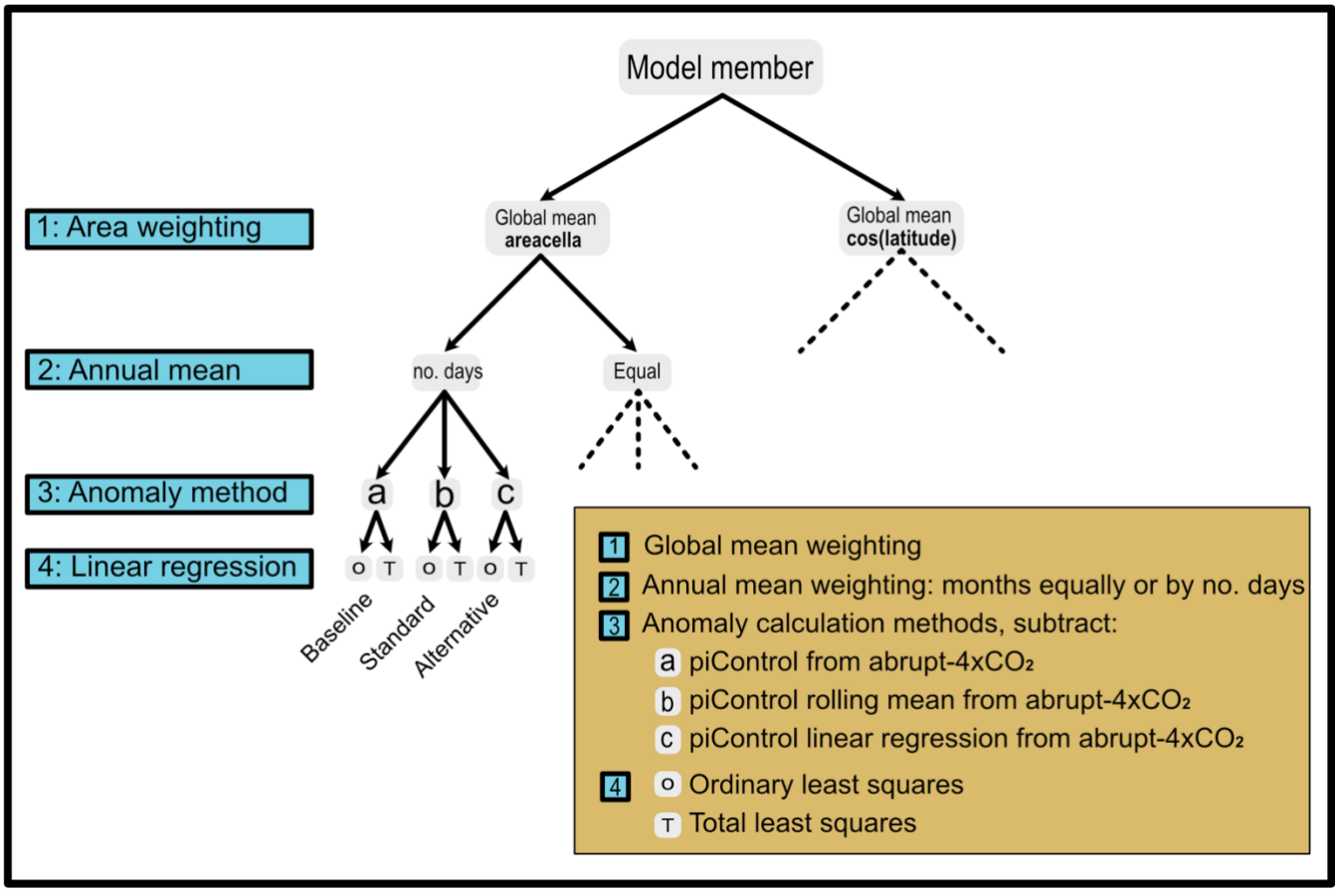


Figure 1. Decision tree illustrating the four steps and possible choices that we compare in this study. For simplicity, we have not shown all 20 paths, although these are indicated by the dashed lines. The Baseline, Standard, and Alternative paths form the basis for much of our comparison, although we investigate the differences between all paths.

We acknowledge that the choices and order of steps we identify in this study may not align with the steps taken by other researchers. Given the lack of methodological details in some studies, and given the number of data processing choices and different orders in the lead up to the regression analysis, we aim to investigate it is important to be clear about the exact path taken in any study.

In the following, we describe the choices at each data processing step. We include only one member for each model, prioritising the first member, e.g. "r1i1…" (Wang et al., 2025; Zelinka et al., 2013) where possible. To calculate the global mean, we

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

compare two common approaches, weighting by grid-cell area or by cosine of the latitude, cos(lat). For the annual mean calculation, we choose to use annual (rather than a longer) time period mean, consistent with much of the literature including the original Gregory et al. (2004) study. The choices we compare are to weight each month equally, or to weight each month by its number of days. To calculate the anomalies, we compare three approaches which broadly reflect the methods used across the literature:

A. Subtract each year of the piControl from the contemporaneous abrupt-4xCO_2 timeseries.

B. Calculate a 21-year rolling average over the piControl and subtract the resulting timeseries from the contemporaneous abrupt-4xCO_2 simulation. Note that the first use of this method by Caldwell et al. (2016) compared a range of window sizes and found that it made no difference to the ECS estimate for CMIP5 models. This anomaly calculation method has been replicated for CMIP6 models (Eiselt and Graversen, 2023; Zelinka et al., 2020) using a 21-year rolling average. However, window size has not been compared for CMIP6 models. We calculate the ECS using an OLS fit across a range of window sizes – 3, 5, 11, 21, 31, 41, 71 years – and find it makes no difference compared to the 21-year rolling average (Fig. S1). Thus, for consistency with recent studies, we retain the 21-year window size.

C. Calculate a linear regression over 150 years of the piControl timeseries for each variable and subtract this linear fit from the corresponding years of the abrupt-4xCO_2 timeseries.

In addition to the steps described above, it is necessary to align the abrupt-4xCO_2 experiment with the piControl at the prescribed branch time. We perform branch alignment after calculating the global mean. While this is a necessary step in data processing, we do not identify alternative choices and thus do not analyse its impact on the ECS. Furthermore, we note that the branch times are not always reliable and for some models the correction may not be accurate. Introducing validation of branching information at the point of simulation submission for CMIP7 would greatly reduce the total time spent on these corrections after initial submission.

A final data processing step is calculating the TOA net radiative flux (RNDT), which is equal to $rsdt - rsut - rlut$. We identify this as a potential step, given the RNDT can be calculated before or after the anomalies. However, upon investigation, we find the order of RNDT calculation relative to the anomalies makes zero difference to the ECS estimate, thus we do not include it in the remainder of the analysis.

Following the data processing, we fit a linear regression line over the first 150 years of the RNDT and TAS anomalies using two methods. First, for consistency with previous literature, we perform an OLS regression with TAS as the independent variable. Additionally, we fit a TLS – alternatively called ‘orthogonal regression’ – line to the data. The key differences

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

between these two methods are that OLS minimises the sum of squared residuals in the y-variable, whereas TLS minimises the sum of squared perpendicular distances between the data points and the regression line (Isobe et al., 1990), thereby removing the need to choose an independent variable. For both regression methods, we take the T-intercept (divided by two) as the ECS, the N-intercept (divided by two) as the radiative forcing due to doubling CO_2, and the slope as the feedback parameter.

To assess the uncertainty of each individual ECS calculation, we use two bootstrapping approaches. The first approach uses a standard bootstrap by sampling over the RNDT and TAS anomaly timeseries 150 times with replacement, calculating the ECS and repeating 10,000 times. The second approach uses a moving block bootstrap (Gilda, 2024) to account for interannual dependence in the timeseries. This approach randomly samples blocks of consecutive data points with replacement, calculating the ECS and repeating 10,000 times to obtain a 95% confidence interval.

3. Comparing the Gregory method data processing choices

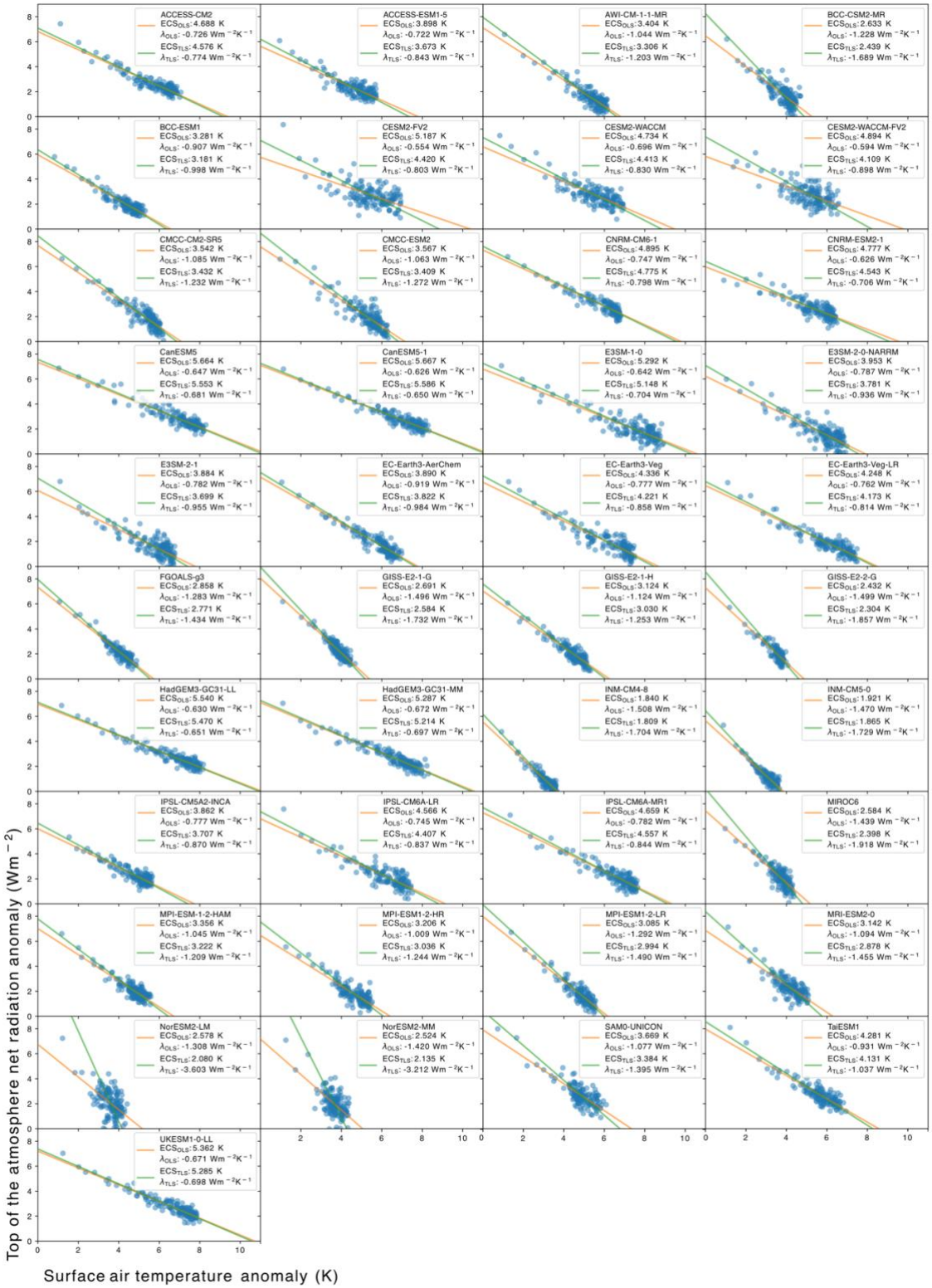
We calculate 20 ECS estimates for each model using the data processing choices described in the methods. An example of the Gregory plot for each model (the scatterplot of the 150-year N-ΔT anomalies with an OLS and TLS regression fit), calculated using the Baseline pathway, is shown below (Fig. 2). Using the Baseline pathway as our point of comparison, we apply a Kolmolgorov-Smirnov test to compare the inter-model ECS distributions between the remaining 20 paths. The test reveals no significant difference in inter-model ECS range between paths, even when comparing paths calculated using an OLS and TLS fit.

Despite the lack of significance between paths for the ensemble ECS range, we find that the preparation choices matter for a subset of individual models. In the following subsections we discuss the implications of the different choices for each data processing step. This analysis leads to a recommended path for a standardised GM. Note that in the following we use an OLS fit for the ECS estimates unless otherwise specified.

3.1 Global mean weighting

We compare two global mean weighting methods: by grid cell area and cosine of the latitude. For most models, the choice of global mean weighting method has little to no impact (likely because these models have regular grids, Fig. 3a), as the median ECS difference across the ensemble when comparing weighting methods is effectively zero. However, we observe four outlier models for which the global mean weighting makes a difference. For AWI-1-1-MR, MPI-ESM-1-2-HAM, and MPI-ESM1-2-HR, weighting the global mean by cos(lat) reduces the ECS estimate by 0.29 K (9%), 0.36 K (11%), and 0.21 K (7%), respectively. For HadGEM3-GC31-MM, weighting by cosine of the latitude increases the ECS estimate by 0.16 K (4%).

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025



https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Figure 2. The Gregory plots calculated from the Baseline pathway for each model. The blue scatter plot represents the anomalies over time in the surface air temperature and radiative flux timeseries. The orange and green lines show linear fits calculated using ordinary and total least squares regression, respectively.

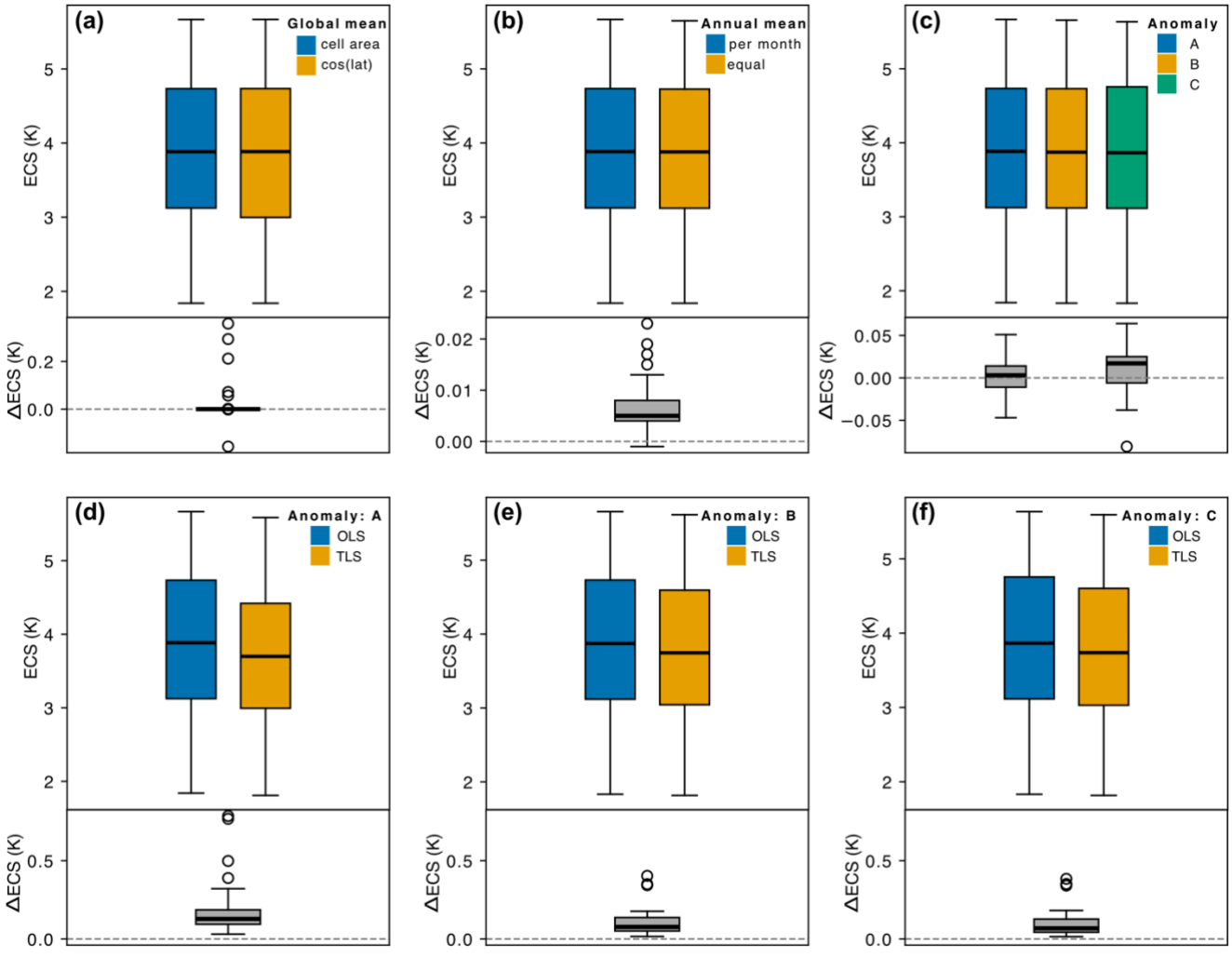


Figure 3. Each subplot shows the inter-model ECS range (upper) and differences between these ranges (lower) comparing the choices at each of the data preparation steps. **a)** Global mean weighting comparing cell area and cosine of the latitude. **b)** Annual mean weighting compares weighting by number of days per month, or by each month equally. **c)** Anomaly calculation method, with uppercase letters denoting the raw piControl, A, rolling mean, B, and linear trend, C. **d), e), f)** OLS compared to TLS regression for the three anomaly methods. Note that the differences in range are always calculated as orange subtracted

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

from blue (or green subtracted from blue, in the case of plot **c)**). Additionally, note that the difference in ECS range for plots **d), e), f)** share a y-axis.

The differences in ECS for global mean weighting methods primarily arise from the model's treatment of grid cell areas at high latitudes, especially for AWI-1-1-MR, MPI-ESM-1-2-HAM, and MPI-ESM1-2-HR (Fig. S2). Given the strong influence of polar regions on the global mean, differences in weighting at the poles can lead to variations in the ECS estimate. This will be prevalent if a model's native grid cells are irregular in shape or size, meaning that weighting by cos(lat) may introduce errors in comparison to the true cell area.

Many researchers may use regridding to calculate the global mean. For this study, we do not consider regridding techniques. Instead, we highlight the potential differences in using a cos(lat) approximation for a model's native grid cell area. Where possible, we recommend weighting the global mean by cell area and working with the model's native grid, as this reduces the number of choices to be made. Where cell area is not available, cos(lat) may be used as an approximation, however this may introduce small errors. This is a clear demonstration of the importance of the "areacella" variable in CMIP submissions.

3.2 Annual mean weighting

The two different annual mean weighting methods we compare – weighting each month equally or by the number of days – results in a median difference of 0.005 K (Fig. 3b). The maximum difference is 0.023 K (0.04%) for CESM-FV2, indicating the amount the ECS reduces when weighting each month equally. Given these results, we conclude that the ECS is largely insensitive to annual mean weighting choices.

In the original study, Gregory et al. (2004) identify the potential of using annual or longer-period means. However, we find that most studies use annual means, so for consistency with previous literature we recommend that annual means remain standard. We recommend calculating the annual mean weighting each month by the number of days, given this is a true reflection of the annual value and all the information is provided in the model data.

3.3 Anomaly calculation method

Of the data processing steps analysed in this study, the anomaly calculation method is the most commonly described in the literature. We compare three methods that broadly reflect the different approaches between studies. These methods form the basis for our Baseline, Standard, and Alternative paths, which respectively calculate the anomalies relative to a raw piControl, a 21-year rolling average, and a linear trend. To evaluate the impact of these different approaches, we calculate the differences in the inter-model ECS range between the Baseline and Standard paths, as well as between the Baseline and Alternative paths

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

(Fig. 3c). The median difference between the Baseline and Standard paths is 0.013 K, with a maximum difference of 0.05 K (1.3%) for the IPSL-CM5A2-INCA model. The median difference between the Baseline and Alternative paths is a decrease of 0.02 K, and the maximum difference is an increase of 0.08 K (1.6%) for the CESM-FV2 model.

Previous studies which compute anomalies relative to a climatological mean or linear trend cite their methods as aiming to reduce the effects of model drift (Andrews et al., 2012; Armour, 2017; Caldwell et al., 2016; Flynn and Mauritsen, 2020). Since these methods are replicated and cited by more recent research, we assume that these researchers also aim to reduce model drift (Dong et al., 2020; Eiselt and Graversen, 2022; Po-Chedley et al., 2018; Zelinka et al., 2020).

Model drift refers to the long-term unforced trend due to incomplete spin-up or non-closure of global energy mass budgets (Irving et al., 2021). Studies typically diagnose model drift in unforced experiments (Gupta et al., 2012, 2013; Irving et al., 2021), although Hobbs et al. (2016) find that energy biases in CMIP5 models are largely insensitive to the forcing experiment, suggesting that the drift present in the piControl is likely also observed in the abrupt-4xCO_2 experiment. While drift in forced experiments has not been explicitly examined for the CMIP6 ensemble, Irving et al. (2021) assume it to be equivalent to that in the piControl, based on the findings of Hobbs et al. (2016) for CMIP5. Thus, assuming an equivalent drift is present in both the abrupt-4xCO_2 and piControl experiment, each of the anomaly calculation methods we investigate will implicitly remove the model drift following the subtraction. It is only if, for example, a trend is removed from only one of the experiments prior to the anomaly calculation, that biases may be introduced.

While the ECS estimates are relatively insensitive to the anomaly calculation method when using an OLS fit, we observe larger differences when comparing the inter-model range of each method between an OLS and TLS fit (Fig. 3d,e,f). The median difference between OLS and TLS for the baseline is 0.13 K, whereas the median differences for the Standard and Alternative paths are 0.08 K and 0.07K respectively. In addition, the difference in inter-model range for the latter two anomaly methods is narrower than for the Baseline. The Baseline exposes an outlier of 0.8 K (16%) difference for CESM-WACCM-FV2, and the Standard and Alternative paths share an outlier of 0.4 K (16%) for NorESM-LM.

The differences between anomaly methods when comparing OLS and TLS results from a reduction in scatter for anomalies calculated following the application of a trend or climatology. The median correlation between RNDT and TAS for the Baseline, Standard, and Alternative paths are -0.89, -0.93, and -0.94 respectively. The largest differences in correlations, however, we observe for our outlier models, such as a difference in correlations for CESM-WACCM-FV2 of -0.15 comparing both the Standard and Alternative paths to the Baseline.

The differences in correlation likely results from a reduction in variance for the Standard and Alternative paths in comparison to the Baseline which retains the raw piControl for the anomaly calculation method. For TAS, the variance is less sensitive to

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

the anomaly calculation method, with median variances across all models being 0.80, 0.78, and 0.73 for the Baseline, Standard, and Alternative paths, respectively. However, for RNDT, the median variances show a more substantial difference: 0.83 for the Baseline, 0.71 for the Standard, and 0.72 for the Alternative path. Notably, the model with the largest correlation difference – CESM-WACCM-FV2 – exhibits the largest reduction in variance for RNDT, from 0.73 for the Baseline to 0.46 and 0.48 for the Standard and Alternative paths, respectively (although there is little difference in TAS variance for this model across anomaly calculation methods).

Given the increase in correlation between RNDT and TAS for the Standard and Alternative anomaly methods, indicating the reduction of some scatter, we recommend calculating the anomalies relative to a climatological mean or linear fit. To ensure consistency between future studies, we suggest using a 21-year running mean over the piControl, as this follows the method of the widely cited Zelinka et al. (2020) paper which calculates the ECS across the CMIP6 ensemble.

3.4 Linear regression method

In this study, we consider two linear regression fits: ordinary and total least squares regression. To the best of our knowledge, most researchers use the OLS fit of N against T to calculate the slope (λ) and ECS when using the Gregory method, e.g. (Andrews et al., 2012, 2015; Armour, 2017; Bloch-Johnson et al., 2021; Caldwell et al., 2016; Chao and Dessler, 2021; Dai et al., 2020; Dong et al., 2020; Rugenstein and Armour, 2021; Zelinka et al., 2020; Zhou et al., 2021). This is consistent with the original approach of Gregory et al. (2004), who treated temperature as the "arbitrary" choice of independent variable. However, across CMIP6 models, this choice is not arbitrary. The median slope (λ) across models is affected by the choice of independent variable; 0.89 W/m²/K when using TAS and 0.74 W/m²/K when using RNDT (Fig. 4a). For individual models, the dependent variable of choice may result in even more substantial variation (Fig. 4b), notably impacting the derived climate sensitivity.

For OLS to provide a reasonable fit, the data must meet two key conditions: there should be a clear dependent variable, and the independent variable must be measured without error (Isobe et al., 1990). In contrast, TLS accounts for errors in both variables, treats them symmetrically, and is more appropriate when seeking to determine a relationship between variables rather than establishing a causal link. Here, errors are not measurement errors, but instead are the random variations on top of the signal we are trying to fit. So, while it is not strictly an error, natural variability plays basically the same role in this study.

Gregory et al. (2004) justify using OLS over alternate regression methods on the basis of the minimal "scatter about a straight line resulting from internally generated variability". They find that the minimal scatter in the data leads to a negligible difference in slope regardless of the choice of dependent variable. However, this rationale was based on a single abrupt-4xCO_2 experiment from the HadSM3 slab ocean model. This assumption of minimal scatter does not hold for many of the fully coupled models developed since 2004. We observe substantial scatter across a range of CMIP6 models (Fig. 2), suggesting that the original justification of OLS is worth reconsidering.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Previous research has justified using temperature as the independent variable. Murphy et al. (2009) found that, on short timescales, temperature variations drive changes in outgoing radiation. Similarly, Forster and Gregory (2006) observed that temperature generally leads radiative flux, and Gregory et al. (2020) followed the physical intuition that temperature determines the magnitude of radiative flux. However, these justifications are primarily grounded in observations. For idealised model simulations, the leading relationship between radiative flux and temperature is not always evident from the timeseries alone, especially for the strongly perturbed abrupt-4xCO_2 experiments.

Given the absence of a clear causal direction from which to define an independent variable, we turn to the second key assumption of OLS: the identification of error. If one variable exhibits errors that are uncorrelated with the other variable, we typically assign the former as the dependent variable, assuming the independent variable is perfectly known (see Appendix B in Gregory et al., 2020). However, if both variables contain uncorrelated errors, TLS provides a more appropriate regression approach, as it accounts for errors in both variables rather than treating one as exact.

Unlike in observational timeseries, where errors are often well-characterised – such as instrumental uncertainty or random measurement errors – errors in climate models primarily arise from unforced variability (Gregory et al., 2020). This variability functions similarly to noise in a statistical sense, obscuring the signal we aim to extract. While it does not introduce randomness in the same way as observational errors, it complicates regression analysis by adding fluctuations that are unrelated to the primary forcing-response relationship of interest.

We can remove some of the variability in the TAS and RNDT timeseries through the anomaly calculation method. The methods which apply a climatology or linear fit to the piControl experiment removes some of the variability from the timeseries and increases the correlation between the two variables. However, to our knowledge no method exists which removes all natural variation from the model while leaving the pure forced signal. Gregory et al. (2020) used the historical ensemble mean of multiple members of MPI-ESM1.1 to argue that temperature exhibits minimal noise, supporting its use as the independent variable. However, they also acknowledge that this assumption may not hold for other ESMs. Given we cannot confidently justify treating either RNDT or TAS as the perfect independent variable, OLS may not be the most robust regression method in this context.

While we find that statistical arguments favour TLS, a number of arguments exist for retaining OLS as the preferred regression method. Firstly, retaining OLS is consistent with the last two decades of ECS research, allowing for comparisons between and within CMIP generations (although recalculating using new methods is an option given the long-term archival and access provided by the ESGF). Secondly, physical reasoning regarding ECS bias supports OLS. ECS estimates from the GM have a known low bias compared to *true* ECS values obtained from fully coupled simulations run for multiple millennia of simulation years (Rugenstein et al., 2020). We find that TLS systematically yields lower ECS values compared to OLS (Fig. 4c). This is consistent with findings of Forster and Gregory (2006), who deliberately chose the regression method which gave the largest

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

sensitivity estimate. The low bias of TLS likely arises from the TLS favouring earlier years of the regression compared to OLS, which may result in an overestimated effective radiative forcing.

Clearly, the choice of regression matters. While we analyse and compare OLS and TLS fits, exploring additional regression methods, such as the York method, or Deming regression, may provide further insights (Him and Pendergrass, 2024; Wu and Yu, 2018). We recommend that future ECS studies clearly report the regression method used and we encourage future research into more robust regression methods. Despite this, in the absence of clearer evidence, we believe that OLS should remain the basis of comparison to remain consistent with the majority of the literature.

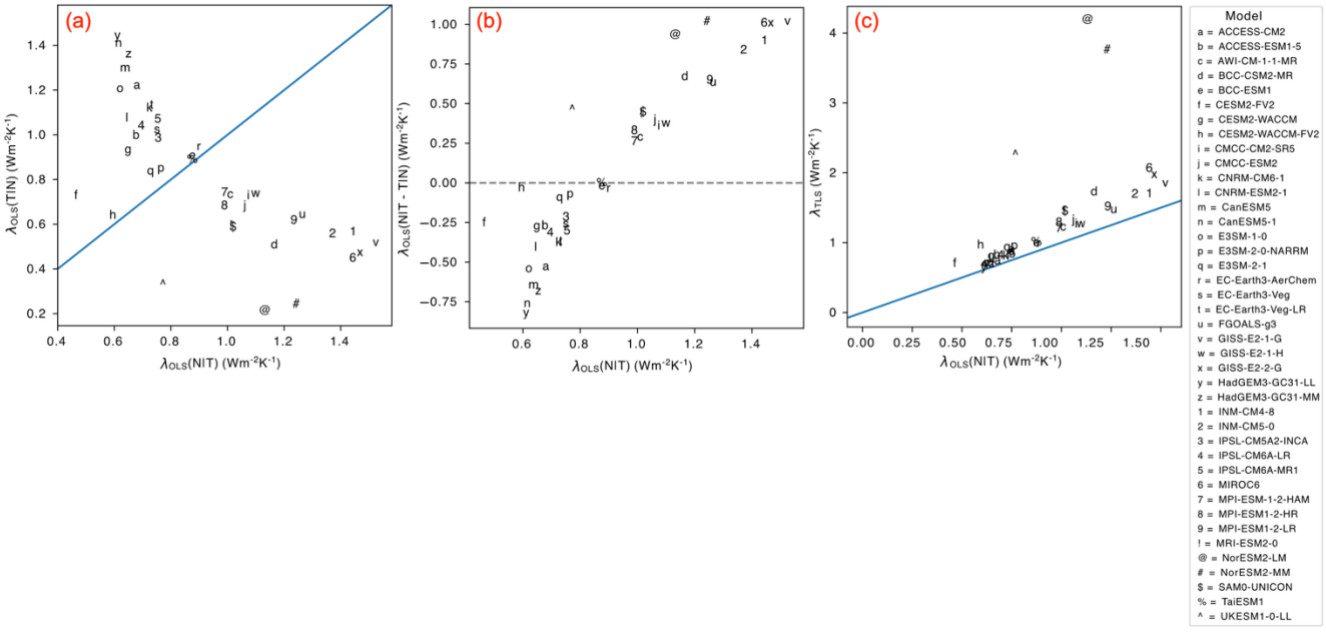


Figure 4. **a)** The slope (λ) of each CMIP6 model calculated using ordinary least squares (OLS) regression with TAS as the independent variable (x-axis) and RNDT as the independent variable (y-axis). Blue line shows the linear relationship required for the choice of independent variable to make no difference. **b)** y-axis showing the difference in slope for each CMIP6 model between the OLS regression based on TAS or RNDT as the independent variable. x-axis is the same as (a). Dashed line at y=0. **c)** The slope of each CMIP6 model calculated using total least squares (TLS) on the y-axis and OLS on the x-axis. Note that **a)** and **b)** follow the same form as Appendix C of Gregory et al. (2020), but use abrupt-4xCO_2 experiment here instead of the historical simulation.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

3.5 Uncertainty range for individual ECS estimates

While calculating uncertainty over ECS estimates has not been included in the steps we analyse, we feel this is an important step that is lacking from some climate sensitivity studies. The studies that do calculate an uncertainty range typically use a standard bootstrap approach, randomly sampling data points from the time series (with replacement) to generate 10,000 subsets for performing the Gregory regression (Andrews et al., 2012; Bloch-Johnson et al., 2021; Rugenstein et al., 2020). This is a common approach for constructing an uncertainty range; however, it assumes annual independence of data, which does not hold for some models.

To assess the level of inter-annual dependence across models, we calculate the autocorrelation of the TAS timeseries following the removal of a quadratic fit for the three different anomaly method pathways (Fig. S3). While most models exhibit the exponential decaying decorrelation of an autoregressive 1 (AR1) process, some models exhibit oscillating behaviour consistent with an AR2 process. In particular, CMCC-CM2-SR5, CMCC-ESM2, EC-Earth3-AerChem, EC-Earth3-Veg, EC-Earth3-Veg-LR, GISS-E2-1-G, GISS-E2-1-H, MIROC6, NorESM2-MM, UKESM1-0-LL show oscillations, with periods of between 3-6 years. For some of these models the AR process displayed depends on the anomaly calculation method, for example CMCC-CM2-SR5 shows an AR2 process for anomaly methods (B) and (C), whereas when using the raw piControl for anomalies it shows a decaying AR1 process.

The AR2 characteristics within these models is an unlikely feature of independent samples, suggesting the presence of an inter-annual or -decadal mode of variability. For example, a four-year period could be indicative of the El Niño Southern Oscillation (ENSO), however in the real world ENSO has a period of between 2 to 7 years (Tang et al., 2018). Thus a model with such a consistent four year ENSO – or other mode of variability – signal would be an unrealistic representation of the real world and should be considered when using the model for climate sensitivity analysis and calculating the uncertainty range. We note that this is not necessarily a feature of the anomaly calculation, however, and instead is an underlying feature of the model given the residuals of the raw abrupt-4xCO_2 time series also exhibit similar AR2 processes for the same models (Fig. S4).

It is important to consider how interannual dependence affects the confidence of ECS estimates. Gregory et al. (2004) acknowledge that interannual variability can have an impact on calculating the uncertainty range, but argue that ignoring the time dependence of the time series primarily results in a narrower uncertainty range rather than introducing bias. Jain et al. (2021) also highlight that TAS and RNDT timeseries exhibit temporal dependence, leading to an underestimation of errors. They address this by either adjusting the number of model years using an effective sample size based on time-lag correlations or by applying a standard bootstrap resampling approach, as done by Andrews et al. (2012). However, these approaches may result in different uncertainty ranges, given the standard bootstrap approach assumes independent data points, which is not true for all models.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

We find that the interannual time dependence of the data varies by model and anomaly calculation method. To account for this, we compare two bootstrap approaches: a standard bootstrap, replicating previous studies, and a block bootstrap with a block size of four years, which accounts for interannual correlations. We calculate a 95% confidence interval using the two bootstrap approaches around the ECS estimate for individual models (Fig. 5a). For simplicity, we use the Baseline pathway and the OLS fit (although we also show the same figure in supplementary, calculated using a TLS fit, Fig. S5).

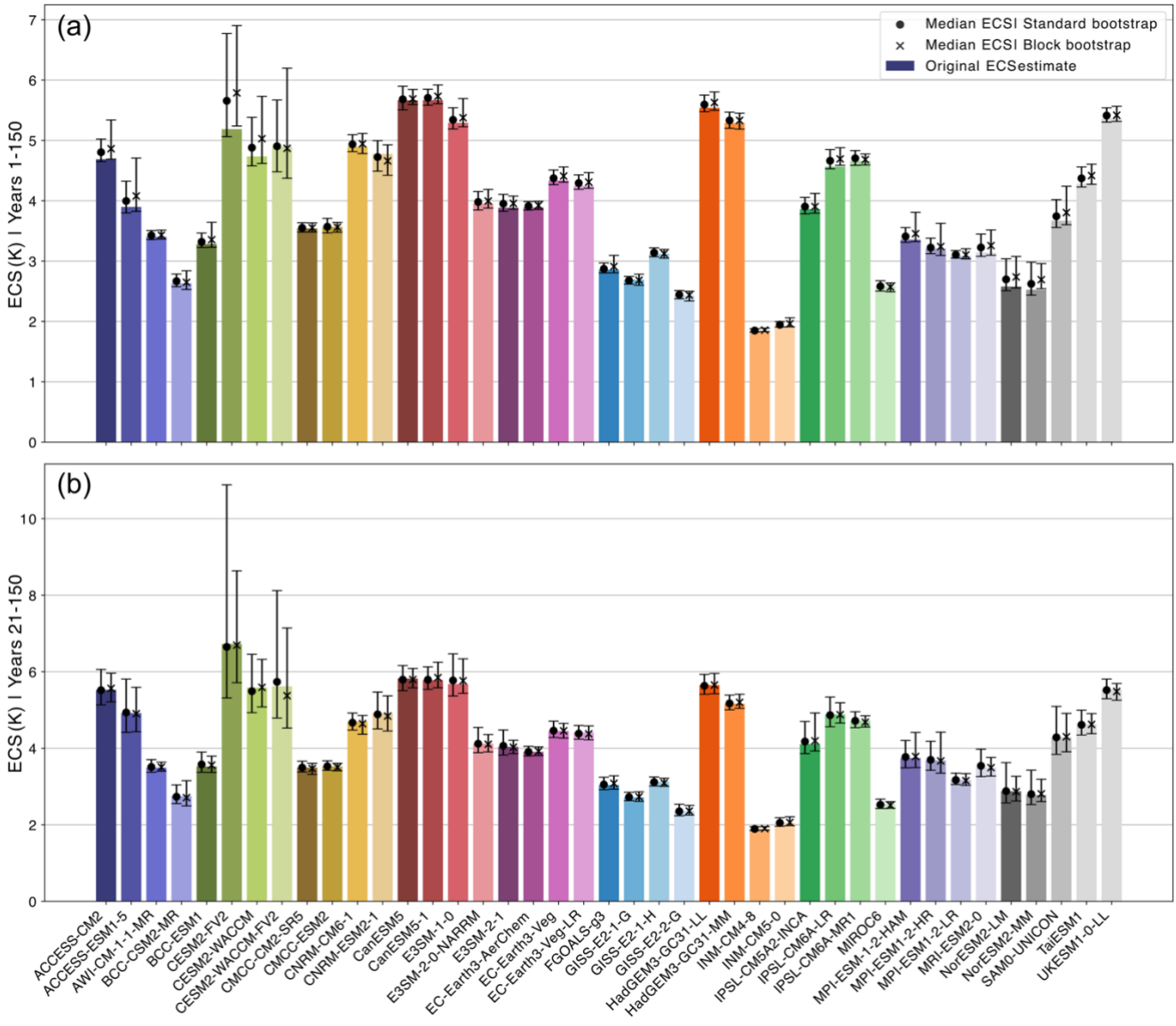
For most models the median ECS calculated using both the bootstrap approaches are larger than the original ECS estimate – for 38 models using the standard bootstrap, and 35 models using the block bootstrap. Additionally, for 31 models the median ECS calculated using the block bootstrap is larger than the median ECS calculated from the standard bootstrap. Most notably, the uncertainty range for some models sits well above the original ECS estimate (e.g. ACCESS-CM2, ACCESS-ESM1-5, CESM2-FV2, and CESM2-WACCM, NorESM2-LM, NorESM2-MM, TaiESM1).

Clearly, the uncertainty ranges for individual models have a high bias, regardless of the bootstrap approach. This bias arises from a sensitivity to the early years of the experiment. The Gregory plots (Fig. 2) for these models show data points with low temperature anomalies and high radiative flux anomalies in the initial years. When bootstrapping across all 150 years, these early data points are often underrepresented in resampled datasets, leading to a systematic overestimation of the ECS compared to the original calculation. However, this reasoning could support the previous research which excludes early years from the data to calculate the ECS (Andrews et al., 2015; Dunne et al., 2020). Rather than overestimating the ECS, the uncertainty ranges may better represent the 'true' value.

To eliminate the differences between the bootstrap uncertainty and the original ECS estimate, we repeat the analysis while restricting both the original ECS calculation and bootstrap uncertainty estimation to years 21–150. This removes the early-year influence, yielding more consistent confidence intervals (Fig. 5b). We note that excluding the first 20 years has implications for radiative forcing estimates, as it raises the question of how long a model must run before the climate response stabilises. While this warrants further investigation, we leave this for future research, as our study focuses specifically on ECS estimation.

For future research, it is important for studies to include an ECS uncertainty range around the estimate. Ideally, modelling groups would provide multiple simulations of the abrupt-4xCO_2 timeseries to provide a more robust basis for the uncertainty assessment, given this would allow for resampling from independent experiments. However, given this is unlikely across all modeling groups, we recommend plotting the autocorrelations of the TAS and RNDT anomaly time series to assess interannual dependence in the data to inform the bootstrap resampling method. Additionally, alternative uncertainty calculation methods could be investigated which downweight the early years of the experiments, although this may be less necessary if CMIP7 abrupt-4xCO_2 experiments are run to 300 simulation years instead of the previously required 150 years.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025



445 **Figure 5.** ECS uncertainty using an ordinary least squares fit. **a)** ECS estimates for each model using the baseline Gregory Method, using years 1-150. Bars represent 95% confidence intervals, with medians calculated using a simple bootstrap (solid circle) and a moving block bootstrap with a block size of 4 (cross). **b)** The same as (a), but the ECS and bootstrap uncertainties are calculated using years 21-150 of the RNDT and TAS anomaly timeseries. See Methods for details on confidence interval calculations.

450

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

4. Discussion and conclusions

For each of the 41 CMIP6 models in this study, we compare 20 ECS estimates derived from alternative choices in data preparation steps and linear regression methods. We find no statistically significant difference between the inter-model ECS ranges across the data preparation paths, or when comparing ordinary and total least squares regression fits. Literature which compares the ECS inter-model spread across CMIP6 models, e.g. (Chao and Dessler, 2021; Dong et al., 2020; Eiselt and Graversen, 2023; Flynn and Mauritsen, 2020; Meehl et al., 2020; Rugenstein et al., 2020; Zelinka et al., 2020), are unlikely to see a meaningful difference in results by recalculating based on an alternate data preparation pathway.

Differences in ECS estimates arise, however, when comparing a subset of CMIP6 models. We find that the steps that result in the largest difference to individual ECS estimates are the choice of global mean weighting, anomaly calculation method and linear regression method. Weighting by cos(lat) compared to the model's native cell area can result in differences of around 10%, although for most models the cos(lat) approximation has almost no error. Whilst individual anomaly methods do not alter the ECS much for just the OLS fit, the range is narrower for anomaly methods which use a climatology or linear trend applied to the piControl, resolving some of the differences between OLS and TLS.

OLS has traditionally been the default linear regression method for the Gregory Method. However, we recommend further exploration of alternative approaches – such as TLS – to better balance physical understanding with statistical robustness in ECS estimation. We find that, for most models, the choice of dependent variable influences the slope of the regression, contradicting previous assumptions that the choice is arbitrary (Andrews et al., 2015; Gregory et al., 2004). Additionally, given errors – or interannual variations on top of the forced signal – are present in both variables, we do not confidently identify one variable over the other as being simulated without error. For consistency with previous research and given the physical reasoning of GM-calculated ECS low bias, OLS should remain the standard, but with room for further investigation.

One step that we do not include in this study is the choice of CO_2 perturbation experiment. Despite the ECS metric being defined as the response to CO_2 doubling, research typically uses CO_2 quadrupling. Using CO_2 quadrupling intends to maximise the signal-to-noise ratio (Bryan et al., 1988; Dai et al., 2020; Washington and Meehl, 1983). However, a large body of literature identifies a non-linear scaling for each consecutive CO_2 doubling (Bloch-Johnson et al., 2021; Chalmers et al., 2022; Hansen et al., 2005; Li et al., 2013; Meraner et al., 2013; Mitevski et al., 2021, 2022, 2023; Russell et al., 2013). This could overestimate the ECS relative to an abrupt-2xCO_2 experiment. However, research also shows that the Gregory method can underestimate the true ECS by 17% (Rugenstein et al., 2020), 14% (Dunne et al., 2020), or 10% (Li et al., 2013). Sherwood et al. (2020) propose that this underestimation, combined with the overestimation due to the nonlinear climate response to consecutive CO_2 doublings, could potentially "cancel out," resulting in an accurate sensitivity estimate using the Gregory method. However, this hypothesis has not been systematically assessed in the literature and warrants further investigation.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Based on our findings, we provide a set of recommendations for future climate sensitivity research (Table 1). These detail the steps, choices at each step, our recommendations, and the caveats in those recommendations. We acknowledge that not all studies applying the Gregory method have the ECS as their primary focus, so there may be alternative choices researchers make for their analysis that we do not explore. At a minimum, we recommend that future studies clearly report their methods, choices, and order of operations to support transparency and reproducibility (with, in our opinion, the simplest option being to simply publish code alongside studies, as this is the least ambiguous description of what was actually done). With the upcoming release of CMIP7 models, data preparation choices may play a more critical role than for CMIP6, underscoring the need for a standardised Gregory method calculation.

Table 1. The steps, choices, recommendations, and caveats we investigate in this study. These recommendations should form the basis of a standardised Gregory method for future research.

STEP	CHOICES	RECOMMENDATION	NOTES
Model member (variant)	Depends on the modelling group	r1i1p1f1	Use the first by default, although ideally calculate the ECS for all available ensemble members to quantify the sensitivity to different realisations, initialisations, and model physics.
Global mean weighting	Cell area Cosine of latitude	Cell area (areacella)	Where cell area is not available, cos(lat) is a useful approximation, although it can affect the ECS by around 10%.
Annual mean weighting	Weight each month equally Weight each month by number of days	Weight by number of days	For precision, although it effectively makes no difference.
Anomaly calculation	Subtracting from the abrupt-4xCO_2: a. Raw piControl b. 21-year rolling average c. Linear trend	21-year rolling average	We recommend this choice, although the anomaly method is not as clear cut as other steps. Other anomaly methods are likely worth investigating if sensitivity is of interest.
Linear regression method	Ordinary least squares Total least squares	OLS, with RNDT as the dependent variable, for consistency	This recommendation we make the least strongly, given the arguments for OLS may not hold against statistical scrutiny. We therefore recommend also calculating the TLS for comparison.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Code and data availability

Code required to conduct the analysis is available at https://doi.org/10.5281/zenodo.15485520 (Zehrung and Nicholls, 2025). All data used in this study are publicly available. The raw CMIP6 ESM data (Eyring et al., 2016) can be downloaded from the USA portal of the Earth System Grid Federation (https://aims2.llnl.gov/search/cmip6, ESGF LLNL Metagrid, 2025).

Author contributions

AZ, ADK, and ZN designed the experiments; AZ and ZN performed the analysis; AZ wrote the manuscript draft; ADK, ZN, MDZ, MM reviewed and edited the manuscript.

Competing interests

The authors declare that they have no competing interests

Acknowledgements

We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF.

Financial support

The work of MDZ was supported by the U.S. Department of Energy (DOE) Regional and Global Model Analysis program area and was performed under the auspices of the DOE by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. The work of ADK was supported by the Australian Research Council (CE230100012) and the Australian Government through the National Environmental Science Program. ZN acknowledges funding from the European Union's Horizon 2020 research and innovation programmes (grant agreement no. 101003536) (ESM2025).

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

References

Andrews, T., Gregory, J. M., Webb, M. J., and Taylor, K. E.: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, Geophys. Res. Lett., 39, https://doi.org/10.1029/2012GL051607, 2012.

Andrews, T., Gregory, J. M., and Webb, M. J.: The Dependence of Radiative Forcing and Feedback on Evolving Patterns of Surface Temperature Change in Climate Models, https://doi.org/10.1175/JCLI-D-14-00545.1, 2015.

Armour, K. C.: Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks, Nat. Clim. Change, 7, 331–335, https://doi.org/10.1038/nclimate3278, 2017.

Armour, K. C., Bitz, C. M., and Roe, G. H.: Time-Varying Climate Sensitivity from Regional Feedbacks, https://doi.org/10.1175/JCLI-D-12-00544.1, 2013.

Bloch-Johnson, J., Pierrehumbert, R. T., and Abbot, D. S.: Feedback temperature dependence determines the risk of high warming, Geophys. Res. Lett., 42, 4973–4980, https://doi.org/10.1002/2015GL064240, 2015.

Bloch-Johnson, J., Rugenstein, M., Stolpe, M. B., Rohrschneider, T., Zheng, Y., and Gregory, J. M.: Climate Sensitivity Increases Under Higher CO2 Levels Due to Feedback Temperature Dependence, Geophys. Res. Lett., 48, e2020GL089074, https://doi.org/10.1029/2020GL089074, 2021.

Boer, G. J. and Yu, B.: Dynamical aspects of climate sensitivity, Geophys. Res. Lett., 30, 2002GL016549, https://doi.org/10.1029/2002GL016549, 2003.

Bryan, K., Manabe, S., and Spelman, M. J.: Interhemispheric Asymmetry in the Transient Response of a Coupled Ocean–Atmosphere Model to a CO_2 Forcing, J. Phys. Oceanogr., 18, 851–867, https://doi.org/10.1175/1520-0485(1988)018<0851:IAITTR>2.0.CO;2, 1988.

Byrne, B. and Goldblatt, C.: Radiative forcing at high concentrations of well-mixed greenhouse gases, Geophys. Res. Lett., 41, 152–160, https://doi.org/10.1002/2013GL058456, 2014.

Caldwell, P. M., Zelinka, M. D., Taylor, K. E., and Marvel, K.: Quantifying the Sources of Intermodel Spread in Equilibrium Climate Sensitivity, https://doi.org/10.1175/JCLI-D-15-0352.1, 2016.

Chalmers, J., Kay, J. E., Middlemas, E. A., Maroon, E. A., and DiNezio, P.: Does Disabling Cloud Radiative Feedbacks Change Spatial Patterns of Surface Greenhouse Warming and Cooling?, https://doi.org/10.1175/JCLI-D-21-0391.1, 2022.

Chao, L.-W. and Dessler, A. E.: An Assessment of Climate Feedbacks in Observations and Climate Models Using Different Energy Balance Frameworks, https://doi.org/10.1175/JCLI-D-21-0226.1, 2021.

Dai, A., Huang, D., Rose, B. E. J., Zhu, J., and Tian, X.: Improved methods for estimating equilibrium climate sensitivity from transient warming simulations, Clim. Dyn., 54, 4515–4543, https://doi.org/10.1007/s00382-020-05242-1, 2020.

Dessler, A. E. and Forster, P. M.: An Estimate of Equilibrium Climate Sensitivity From Interannual Variability, J. Geophys. Res. Atmospheres, 123, 8634–8645, https://doi.org/10.1029/2018JD028481, 2018.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Dong, Y., Armour, K. C., Zelinka, M. D., Proistosescu, C., Battisti, D. S., Zhou, C., and Andrews, T.: Intermodel Spread in the Pattern Effect and Its Contribution to Climate Sensitivity in CMIP5 and CMIP6 Models, https://doi.org/10.1175/JCLI-D-19-1011.1, 2020.

Dunne, J. P., Winton, M., Bacmeister, J., Danabasoglu, G., Gettelman, A., Golaz, J.-C., Hannay, C., Schmidt, G. A., Krasting, J. P., Leung, L. R., Nazarenko, L., Sentman, L. T., Stouffer, R. J., and Wolfe, J. D.: Comparison of Equilibrium Climate Sensitivity Estimates From Slab Ocean, 150-Year, and Longer Simulations, Geophys. Res. Lett., 47, e2020GL088852, https://doi.org/10.1029/2020GL088852, 2020.

Dunne, J. P., Hewitt, H. T., Arblaster, J., Bonou, F., Boucher, O., Cavazos, T., Durack, P. J., Hassler, B., Juckes, M., Miyakawa, T., Mizielinski, M., Naik, V., Nicholls, Z., O'Rourke, E., Pincus, R., Sanderson, B. M., Simpson, I. R., and Taylor, K. E.: An evolving Coupled Model Intercomparison Project phase 7 (CMIP7) and Fast Track in support of future climate assessment, https://doi.org/10.5194/egusphere-2024-3874, 20 December 2024.

Eiselt, K.-U. and Graversen, R. G.: Change in Climate Sensitivity and Its Dependence on the Lapse-Rate Feedback in 4 × CO2 Climate Model Experiments, https://doi.org/10.1175/JCLI-D-21-0623.1, 2022.

Eiselt, K.-U. and Graversen, R. G.: On the Control of Northern Hemispheric Feedbacks by AMOC: Evidence from CMIP and Slab Ocean Modeling, J. Clim., 36, 6777–6795, https://doi.org/10.1175/JCLI-D-22-0884.1, 2023.

ESGF LLNL Metagrid: CMIP6, ESGF [data set], https://aims2.llnl.gov/search/cmip6, last access: 26 May 2025

Etminan, M., Myhre, G., Highwood, E. J., and Shine, K. P.: Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing, Geophys. Res. Lett., 43, 12,614-12,623, https://doi.org/10.1002/2016GL071930, 2016.

Flynn, C. M. and Mauritsen, T.: On the climate sensitivity and historical warming evolution in recent coupled model ensembles, Atmospheric Chem. Phys., 20, 7829–7842, https://doi.org/10.5194/acp-20-7829-2020, 2020.

Forster, P. M. F. and Gregory, J. M.: The Climate Sensitivity and Its Components Diagnosed from Earth Radiation Budget Data, J. Clim., 19, 39–52, https://doi.org/10.1175/JCLI3611.1, 2006.

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

Forster, P. M. F., Storelvmo, T., Armour, K. C., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D. J., Mauritsen, T., Palmer, M. D., Watanabe, M., Wild, M., and Zhang, H.: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, Clim. Change 2021 Phys. Sci. Basis Contrib. Work. Group Sixth Assess. Rep. Intergov. Panel Clim. Change, 923–1054, https://doi.org/10.1017/9781009157896.009, 2021.

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

Gilda, S.: tsbootstrap, https://doi.org/10.5281/zenodo.8226495, 2024.

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., Lowe, J. A., Johns, T. C., and Williams, K. D.: A new method for diagnosing radiative forcing and climate sensitivity, Geophys. Res. Lett., 31, https://doi.org/10.1029/2003GL018747, 2004.

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Gregory, J. M., Andrews, T., Ceppi, P., Mauritsen, T., and Webb, M. J.: How accurately can the climate sensitivity to CO2 be estimated from historical climate change?, Clim. Dyn., 54, 129–157, https://doi.org/10.1007/s00382-019-04991-y, 2020.

Gupta, A. S., Muir, L. C., Brown, J. N., Phipps, S. J., Durack, P. J., Monselesan, D., and Wijffels, S. E.: Climate Drift in the CMIP3 Models, https://doi.org/10.1175/JCLI-D-11-00312.1, 2012.

Gupta, A. S., Jourdain, N. C., Brown, J. N., and Monselesan, D.: Climate Drift in the CMIP5 Models, https://doi.org/10.1175/JCLI-D-12-00521.1, 2013.

Hansen, J., Sato, M., Ruedy, R., Nazarenko, L., Lacis, A., Schmidt, G. A., Russell, G., Aleinov, I., Bauer, M., Bauer, S., Bell, N., Cairns, B., Canuto, V., Chandler, M., Cheng, Y., Del Genio, A., Faluvegi, G., Fleming, E., Friend, A., Hall, T., Jackman, C., Kelley, M., Kiang, N., Koch, D., Lean, J., Lerner, J., Lo, K., Menon, S., Miller, R., Minnis, P., Novakov, T., Oinas, V., Perlwitz, Ja., Perlwitz, Ju., Rind, D., Romanou, A., Shindell, D., Stone, P., Sun, S., Tausnev, N., Thresher, D., Wielicki, B., Wong, T., Yao, M., and Zhang, S.: Efficacy of climate forcings, J. Geophys. Res. Atmospheres, 110, https://doi.org/10.1029/2005JD005776, 2005.

Him, W. (Kinen) K. and Pendergrass, A. G.: Timescale Dependence of the Precipitation Response to CO2-Induced Warming in Millennial-Length Climate Simulations, Geophys. Res. Lett., 51, e2024GL111609, https://doi.org/10.1029/2024GL111609, 2024.

Hobbs, W., Palmer, M. D., and Monselesan, D.: An Energy Conservation Analysis of Ocean Drift in the CMIP5 Global Coupled Models, https://doi.org/10.1175/JCLI-D-15-0477.1, 2016.

Irving, D., Hobbs, W., Church, J., and Zika, J.: A Mass and Energy Conservation Analysis of Drift in the CMIP6 Ensemble, https://doi.org/10.1175/JCLI-D-20-0281.1, 2021.

Isobe, T., Feigelson, E. D., Akritas, M. G., and Babu, G. J.: Linear regression in astronomy. I., 364, 104, https://doi.org/10.1086/169390, 1990.

Jain, S., Chhin, R., Doherty, R. M., Mishra, S. K., and Yoden, S.: A New Graphical Method to Diagnose the Impacts of Model Changes on Climate Sensitivity, J. Meteorol. Soc. Jpn. Ser II, 99, 437–448, https://doi.org/10.2151/jmsj.2021-021, 2021.

Klocke, D., Quaas, J., and Stevens, B.: Assessment of different metrics for physical climate feedbacks, Clim. Dyn., 41, 1173–1185, https://doi.org/10.1007/s00382-013-1757-1, 2013.

Li, C., von Storch, J.-S., and Marotzke, J.: Deep-ocean heat uptake and equilibrium climate response, Clim. Dyn., 40, 1071–1086, https://doi.org/10.1007/s00382-012-1350-z, 2013.

Lutsko, N. J., Luongo, M. T., Wall, C. J., and Myers, T. A.: Correlation Between Cloud Adjustments and Cloud Feedbacks Responsible for Larger Range of Climate Sensitivities in CMIP6, J. Geophys. Res. Atmospheres, 127, e2022JD037486, https://doi.org/10.1029/2022JD037486, 2022.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., and Schlund, M.: Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models, Sci. Adv., 6, eaba1981, https://doi.org/10.1126/sciadv.aba1981, 2020.

Meinshausen, M., Nicholls, Z. R. J., Lewis, J., Gidden, M. J., Vogel, E., Freund, M., Beyerle, U., Gessner, C., Nauels, A., Bauer, N., Canadell, J. G., Daniel, J. S., John, A., Krummel, P. B., Luderer, G., Meinshausen, N., Montzka, S. A., Rayner, P. J., Reimann, S., Smith, S. J., van den Berg, M., Velders, G. J. M., Vollmer, M. K., and Wang, R. H. J.: The shared socio-

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500, Geosci. Model Dev., 13, 3571–3605, https://doi.org/10.5194/gmd-13-3571-2020, 2020.

Meraner, K., Mauritsen, T., and Voigt, A.: Robust increase in equilibrium climate sensitivity under global warming, Geophys. Res. Lett., 40, 5944–5948, https://doi.org/10.1002/2013GL058118, 2013.

Mitevski, I., Orbe, C., Chemke, R., Nazarenko, L., and Polvani, L. M.: Non-Monotonic Response of the Climate System to Abrupt CO2 Forcing, Geophys. Res. Lett., 48, e2020GL090861, https://doi.org/10.1029/2020GL090861, 2021.

Mitevski, I., Polvani, L. M., and Orbe, C.: Asymmetric Warming/Cooling Response to CO2 Increase/Decrease Mainly Due To Non-Logarithmic Forcing, Not Feedbacks, Geophys. Res. Lett., 49, e2021GL097133, https://doi.org/10.1029/2021GL097133, 2022.

Mitevski, I., Dong, Y., Polvani, L. M., Rugenstein, M., and Orbe, C.: Non-Monotonic Feedback Dependence Under Abrupt CO2 Forcing Due To a North Atlantic Pattern Effect, Geophys. Res. Lett., 50, e2023GL103617, https://doi.org/10.1029/2023GL103617, 2023.

Murphy, D. M., Solomon, S., Portmann, R. W., Rosenlof, K. H., Forster, P. M., and Wong, T.: An observationally based energy balance for the Earth since 1950, J. Geophys. Res. Atmospheres, 114, https://doi.org/10.1029/2009JD012105, 2009.

National Research Council: Carbon dioxide and climate: A scientific assessment, The National Academies Press, Washington, DC, https://doi.org/10.17226/12181, 1979.

Po-Chedley, S., Armour, K. C., Bitz, C. M., Zelinka, M. D., Santer, B. D., and Fu, Q.: Sources of Intermodel Spread in the Lapse Rate and Water Vapor Feedbacks, https://doi.org/10.1175/JCLI-D-17-0674.1, 2018.

Qu, X., Hall, A., DeAngelis, A. M., Zelinka, M. D., Klein, S. A., Su, H., Tian, B., and Zhai, C.: On the Emergent Constraints of Climate Sensitivity, https://doi.org/10.1175/JCLI-D-17-0482.1, 2018.

Ringer, M. A., Andrews, T., and Webb, M. J.: Global-mean radiative feedbacks and forcing in atmosphere-only and coupled atmosphere-ocean climate change experiments, Geophys. Res. Lett., 41, 4035–4042, https://doi.org/10.1002/2014GL060347, 2014.

Rugenstein, M. and Armour, K. C.: Three Flavors of Radiative Feedbacks and Their Implications for Estimating Equilibrium Climate Sensitivity, Geophys. Res. Lett., 48, e2021GL092983, https://doi.org/10.1029/2021GL092983, 2021.

Rugenstein, M., Bloch-Johnson, J., Gregory, J., Andrews, T., Mauritsen, T., Li, C., Frölicher, T. L., Paynter, D., Danabasoglu, G., Yang, S., Dufresne, J.-L., Cao, L., Schmidt, G. A., Abe-Ouchi, A., Geoffroy, O., and Knutti, R.: Equilibrium Climate Sensitivity Estimated by Equilibrating Climate Models, Geophys. Res. Lett., 47, e2019GL083898, https://doi.org/10.1029/2019GL083898, 2020.

Russell, G. L., Lacis, A. A., Rind, D. H., Colose, C., and Opstbaum, R. F.: Fast atmosphere-ocean model runs with large changes in CO2, Geophys. Res. Lett., 40, 5787–5792, https://doi.org/10.1002/2013GL056755, 2013.

Sanderson, B. M. and Rugenstein, M.: Potential for bias in effective climate sensitivity from state-dependent energetic imbalance, Earth Syst. Dyn., 13, 1715–1736, https://doi.org/10.5194/esd-13-1715-2022, 2022.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B., and

https://doi.org/10.5194/egusphere-2025-2252
Preprint. Discussion started: 6 June 2025

Zelinka, M. D.: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence, Rev. Geophys., 58, e2019RG000678, https://doi.org/10.1029/2019RG000678, 2020.

Tang, Y., Zhang, R.-H., Liu, T., Duan, W., Yang, D., Zheng, F., Ren, H., Lian, T., Gao, C., Chen, D., and Mu, M.: Progress in ENSO prediction and predictability study, Natl. Sci. Rev., 5, 826–839, https://doi.org/10.1093/nsr/nwy105, 2018.

Wang, X., Li, L., Wang, H., Zuo, L., Wang, B., and Xie, F.: Understanding equilibrium climate sensitivity changes from CMIP5 to CMIP6: Feedback, AMOC, and precipitation responses, Atmospheric Res., 315, 107917, https://doi.org/10.1016/j.atmosres.2025.107917, 2025.

Washington, W. M. and Meehl, G. A.: General circulation model experiments on the climatic effects due to a doubling and quadrupling of carbon dioxide concentration, J. Geophys. Res. Oceans, 88, 6600–6610, https://doi.org/10.1029/JC088iC11p06600, 1983.

Wu, C. and Yu, J. Z.: Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting, Atmospheric Meas. Tech., 11, 1233–1250, https://doi.org/10.5194/amt-11-1233-2018, 2018.

Zelinka, M. D., Klein, S. A., Taylor, K. E., Andrews, T., Webb, M. J., Gregory, J. M., and Forster, P. M.: Contributions of Different Cloud Types to Feedbacks and Rapid Adjustments in CMIP5, https://doi.org/10.1175/JCLI-D-12-00555.1, 2013.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., and Taylor, K. E.: Causes of Higher Climate Sensitivity in CMIP6 Models, Geophys. Res. Lett., 47, e2019GL085782, https://doi.org/10.1029/2019GL085782, 2020.

Zhou, C., Zelinka, M. D., Dessler, A. E., and Wang, M.: Greater committed warming after accounting for the pattern effect, Nat. Clim. Change, 11, 132–136, https://doi.org/10.1038/s41558-020-00955-x, 2021.

Zehrung, A., and Nicholls, Z.: ECS Gregory method analysis. Zenodo [code], https://doi.org/10.5281/zenodo.15485520, 2025