

Standardising the “Gregory method” for calculating equilibrium climate sensitivity

Anna Zehrung¹, Andrew D. King^{1,2}, Zebedee Nicholls^{1,3,4}, Mark D. Zelinka⁵, 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)

5 **Abstract.** The equilibrium climate sensitivity (ECS) – the equilibrium global mean temperature response to a doubling of atmospheric CO₂ – 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, *N*, and surface air temperature anomalies, *ΔT*, from a 150-year experiment in which atmospheric CO₂ concentrations are quadrupled. The ECS is determined by extrapolating the linear fit to

10 *N*=0, i.e. the *ΔT*-intercept, indicating the point at which 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 32 alternative data processing pathways, varying by differences in global mean weighting, net radiative flux variable, anomaly calculation method, and linear

15 regression fit. Using 44 CMIP6 models, we systematically assess the impact of these choices on ECS estimates and calculate uncertainty ranges using two bootstrap approaches. While the inter-model ECS range is insensitive to the data processing pathway, individual outlier models exhibit notable differences. Approximating a model’s native grid cell area (if irregular) with cosine of the latitude can decrease the ECS by 11%, the choice of *N*-variable can change the ECS by 6%, and some anomaly calculation methods can introduce spurious temporal correlations in the processed data. Beyond data processing

20 choices, we also evaluate an alternative linear regression method – total least squares (TLS) – which has a more statistically robust basis than OLS. However, for consistency with previous literature, and given TLS may reduce the ECS compared to OLS (by up to 24%), thereby making 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, using the top of the atmosphere (as opposed to the top

25 of model) *N*-variable, and calculating anomalies by first applying a rolling average to the preindustrial control timeseries then subtracting from the raw CO₂ quadrupling experiment. This approach 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 insensitive to these processing choices, similar assumptions may not hold for CMIP7, underscoring the need for standardised data preparation in future climate sensitivity assessments.

30

35

Deleted:

Formatted: Font: Italic, Font colour: Auto

Formatted: Font: Italic, Font colour: Auto

Deleted: The ECS is determined at the point where net radiative flux reaches zero i.e. the system is back in equilibrium.

Formatted: Font: Italic, Font colour: Auto

Deleted: 20

Deleted: t choices

Deleted: annual mean weighting

Deleted: 41

Deleted: appears to have

Deleted: physical reasoning suggests that

Deleted: further

Deleted: i.e.

Deleted: make

Deleted: weighting the annual mean by number of days per month

Deleted: piControl

Deleted: implicitly

Deleted: robust

1. Introduction

The equilibrium climate sensitivity (ECS) – the steady state global mean surface temperature response to a doubling of atmospheric CO₂ 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 a range of 1.5 to 4.5 K, based primarily on a three dimensional atmospheric circulation model. The most recent climate model-based estimate uses the model range of the coupled model intercomparison project phase six (CMIP6), placing the ECS between 1.8 to 5.6 K (Zelinka et al., 2020). Meanwhile, the Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (AR6) uses multiple lines of evidence to arrive at the conclusion the ECS is between 2 to 5 K with 95% confidence (Forster et al., 2021; Sherwood et al., 2020).

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 CO₂. 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; Danabasoglu and Gent, 2009).

Since 2004, coupled atmosphere-ocean 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 CO₂ perturbation simulations that are centuries rather than millennia in duration. Hereafter we use the term ECS, noting 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 and leave it up to the reader to decide whether this calculation results in the equilibrium or effective climate sensitivity. Our conclusions are independent of this choice.

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

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

Deleted: [1]

Deleted: n uncertainty

Deleted: places this range at

Deleted: , 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

Deleted: fully

Deleted: We acknowledge

Deleted: because this study does not consider the potential non-linearities within this method (such as an inconstant feedback parameter)...

Formatted: Font: Italic

Formatted: Font: Italic

To calculate the ECS using a coupled climate model, Gregory et al. (2004) take the first 90 years – standard practice has since become 150 years – of an abrupt CO₂ quadrupling experiment (abrupt-4xCO₂) 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 ΔT -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 CO₂ rather than a quadrupling, the ΔT - and N -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 CO₂ is twice that of a CO₂ doubling, which does not exactly hold if the relationship between forcing and CO₂ concentrations is not logarithmic (Byrne and Goldblatt, 2014; Etmann et al., 2016; Meinshausen et al., 2020).

Deleted: 150 years

Deleted: n-linear

The popularity of the GM is likely due to its relative simplicity, offering a linear relationship that allows for a single calculation to estimate the ECS, radiative forcing, and feedback parameter. Moreover, the GM does not require highly specific experiment configurations often needed for estimating the forcing term, such as those with fixed sea surface temperatures (SSTs) or atmospheric model intercomparison project (AMIP)-style setups (such as using SST or sea ice observations). The accuracy of the GM in estimating the three variables of interest is subject to debate (e.g. Andrews et al., 2012; Forster et al., 2016; Rugenstein et al., 2020; Rugenstein and Armour, 2021; Smith et al., 2020), particularly regarding the extent of the linear assumptions and the interpretation of the forcing term. For example, in radiative forcing specific studies, the forcing term is usually estimated from the first 20 or 30 years of data (Forster et al., 2016), rather than the full 150 years more commonly used in climate sensitivity studies. These uncertainties are why we concentrate here primarily on the ECS and feedback parameter. This study focuses on the practical application of the GM, leaving discussions about its widespread use in literature, as well as its strengths and weaknesses, to other work.

Formatted: Default Paragraph Font, Font colour: Auto

The GM is extensively used and cited across literature. It has been applied to assess CMIP5 and CMIP6 (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 climate sensitivity estimates based on alternate lines of evidence, such as observations, historical simulations, or palaeoclimate data (Chao and Dessler, 2021; Sherwood et al., 2020).

Deleted: the fifth and sixth phases of the coupled model intercomparison projects (CMIP)

Deleted: other climate sensitivity estimation approaches

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, amplified by the fact that Gregory et al. (2004) included limited descriptions of data preparation steps in their study. To our knowledge, no study has to date systematically assessed how different data preparation methods may influence ECS results.

Deleted: in approach.

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 (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; Nijssen et al., 2020; Ringer et al., 2014; Zhou et al., 2021). Others provide only limited details, such as specifying the model ensemble member used (Wang et al., 2025; Zelinka et al., 2013).

Deleted: Dai et al., 2020;

Among studies that address N and ΔT data preparation, the focus typically centres on anomaly calculations and methods to account for model drift. In its simplest form, the term 'anomaly' refers to the difference between the corresponding abrupt-4xCO₂ and piControl timeseries. However, methods for calculating anomalies vary widely, including applying a rolling mean (Caldwell et al., 2016; Eiselt and Graversen, 2023; Po-Chedley et al., 2018; Qu et al., 2018; Zelinka et al., 2020), linear trend (Andrews et al., 2012; Armour, 2017; Bloch-Johnson et al., 2021; Dong et al., 2020; Flynn and Mauritsen, 2020; Forster et al., 2013), or long-term average (Chao and Dessler, 2021; Jain et al., 2021; Rugenstein and Armour, 2021) to the piControl prior to subtracting from the abrupt-4xCO₂ experiment.

Deleted: For

Deleted: do

Deleted: is

Deleted: how to account

Deleted: Here, the term anomaly refers to – in its simplest form –

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 32 paths, split into 16 data processing choices and two linear regression methods (Fig. 1); 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. The key difference between the two methods is that OLS requires the choice of an independent variable, and TLS does not assume independence in either variable.

Deleted: :

Deleted: 1

... [2]

Deleted: 10

Deleted: alternative

Deleted: based on the various methods discussed in literature

Deleted: . Each choice ultimately leads to two ECS estimates, given we also compare the application of two different linear regression fits...

Notwithstanding the linear fit method, we do not include modifications to the regression itself. While we assess the exclusion of early years of the experiment as a further analysis in investigating ECS uncertainty (see Section 3.5), we do not include this as a formal data preparation step. 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-4xCO₂ and -2xCO₂ experiments (Dai et al., 2020), are already well-documented and these studies are widely cited across the literature.

This study does not aim to constrain the ECS ensemble range. Instead, our focus is on comparing differences in data preparation and linear regression methods, exploring uncertainty, and establishing a standardised GM analysis approach. This approach aims to promote transparency in methods for future research. These objectives are particularly relevant with the upcoming release of CMIP7 data (Dunne et al., 2025), 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.

Deleted: or address potential non-linearities within the GM calculation. ...

Deleted: for future research.

Deleted: This is

Deleted: 2024

2. Methods

For our analysis, we compare the effects of data preparation choices and linear regression methods across 44 CMIP6 models. The resolution and grids of the models vary (see detailed descriptions in Supplementary Table 1). The grid spacing is between 100km and 500km, and the grids are either a regular latitude and longitude or a more complicated irregular (native) grid. These differences between models motivates the need to assess different global mean weighting methods.

To calculate the ECS based on the steps we investigate, the GM requires six variables, the 2m surface air temperature (tas), top of model (TOM) net radiative flux (rtmt) and - for comparison to rtmt - top of the atmosphere (TOA) reflected shortwave radiation (rsut), TOA outgoing longwave radiation (rlut), and TOA downward shortwave radiation (rsdt). Those variables are at monthly timescales for both the abrupt-4xCO₂ and piControl experiments, and in addition, the atmospheric cell area spatial variable (areacella) is needed.

It is essential for studies using CMIP6 data to be explicit about which variables are being used in their methods. This is especially necessary for climate sensitivity research to clarify whether the ECS is an estimate of the global mean surface or global mean surface air temperature – GMST or GSAT, respectively. GSAT refers to the global 2m air temperature, whereas GMST is a combination of 2m air temperature over land, and SSTs over the ocean (Forster et al., 2021), which requires three variables in addition to tas to account for SSTs and sea ice concentrations (Cowtan et al., 2015). Some climate sensitivity studies are explicit about calculating the GSAT for ECS, e.g. (Andrews et al., 2015; Dai et al., 2020; Eiselt and Graversen, 2023; Gregory et al., 2004; Jain et al., 2021; Rugenstein et al., 2020; Zelinka et al., 2020), while others make the distinction between GMST and GSAT explicitly (Armour et al., 2013; Ceppi and Gregory, 2019; Geoffroy et al., 2013; Nijse et al., 2020; Po-Chedley et al., 2018; Zhou et al., 2021). However, many (Caldwell et al., 2016; Flynn and Mauritsen, 2020; Forster et al., 2013; Klocke et al., 2013; Mitevski et al., 2021; Ringer et al., 2014; Rugenstein and Armour, 2021) refer to the ECS as a measure of GMST without describing the variables or methods used to calculate the global mean. Different methods exist to calculate the GMST from climate model data (Cowtan et al., 2015), generally diverging in their treatment of sea ice, with each method introducing potential biases (Richardson et al., 2016, 2018). It would be a step forward if studies that base their ECS derivations on GMST were explicit with their methods of global mean calculation. Given that Gregory et al. (2004) use GSAT and the IPCC recommends model-based estimates use GSAT (Forster et al., 2021), we recommend calculating the ECS using GSAT rather than GMST.

For this study, we investigate 16 data preparation paths based on choices of global mean weighting, net radiative flux variable, and anomaly calculation method (Fig. 1). These paths lead to two ECS estimates based on either OLS or TLS, which we also use to assess uncertainty in ECS for individual models. While we compare all 16 paths, for simplicity we label only four of them according to their anomaly calculation methods (Fig. 1).

Deleted: 41

Deleted:

Deleted: five

Deleted: TAS

Deleted: , and the atmospheric cell area spatial variable (areacella),

Deleted: 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.

Deleted: 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

Deleted: 20

Deleted: three

Deleted: , as the Baseline, Standard, and Alternative paths. These respectively aim to replicate – to the best of our knowledge – the data 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.

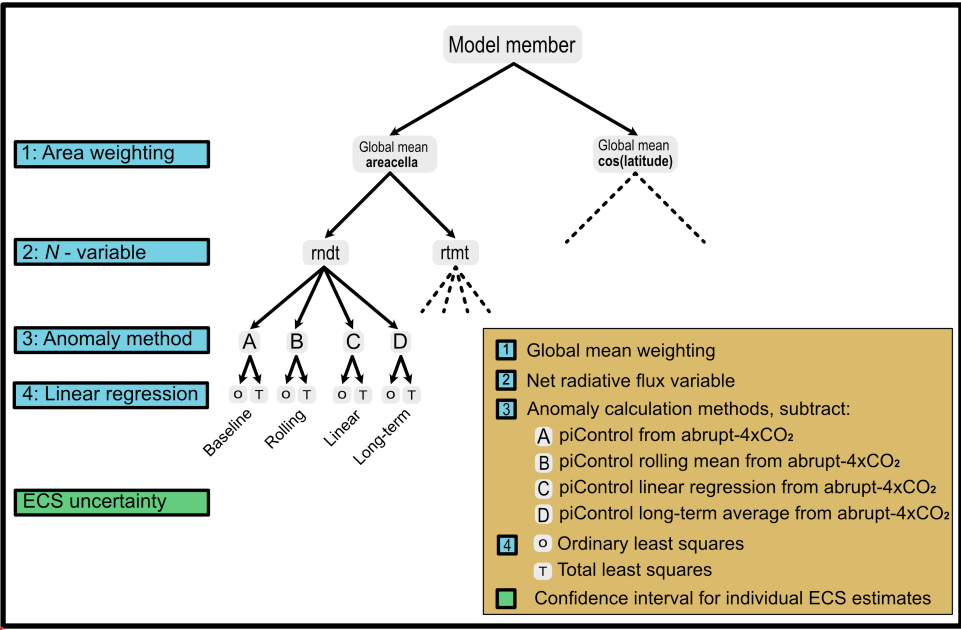
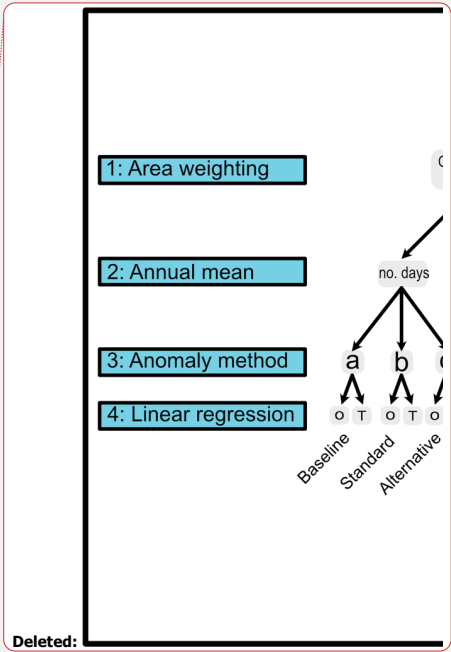


Figure 1. Decision tree illustrating the four steps and possible choices that we compare in this study in addition to the ECS uncertainty calculation. For simplicity, we have not shown all paths, although these are indicated by the dashed lines. The Baseline, Rolling, Linear and Long-term paths form 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. However, 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, 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 ensemble member where possible (Wang et al., 2025; Zelinka et al., 2013). The model ensemble member describes



Deleted:

Deleted: 20

Deleted: Standard, and Alternative

Deleted: G

Deleted: we aim to investigate

Deleted: member, e.g. "r1i1..."

the attributes for each experiment’s specific run. The attributes relate to the realisation (r), initialisation (i), physics (p), and forcing (f) indices. Most models have at least one ensemble member called “r1i1p1f1”, whereas a model which runs two experiments of the same scenario with the same initial conditions, physics, and forcing, would then also have, in theory, an ensemble member called “r2i1p1f1”. The attributes change depending on the indices of the specific run.

To calculate the global mean, we compare two common approaches, weighting by grid-cell area or by cosine of the latitude, $\cos(\text{lat})$. After this step we also calculate the annual mean, although this is not included as a formal step in our investigation. We choose to use an annual mean (rather than the mean of a longer time period), which is consistent with much of the literature including the original Gregory et al. (2004) study. We analyse two annual mean weighting choices: weighting each month equally or each month by the number of days. However, we find the median multi-model ECS difference between these two choices is 0.005 K, and the maximum difference is 0.023 K for CESM2-FV2. Given the ECS appears almost entirely insensitive to the annual mean weighting across all models, we do not include this as a distinct comparison in our analysis.

For the N -variable, most studies lack detail on how they calculate the net radiative flux. In our analysis, we explore approaches which either define N as a measure of the TOA $r_{\text{ndt}} = r_{\text{sd}} - r_{\text{st}} - r_{\text{lt}}$ (Lewis and Curry, 2018), or as the explicit TOM radiative flux variable (rtmt). While we are unfamiliar with the rtmt variable’s use in climate sensitivity literature, it is worthwhile to investigate especially if there are large differences between a model’s explicit top and the TOA.

To calculate the anomalies, we compare four approaches which reflect the methods used across the literature, which we label as:

A. **Baseline:** Subtract each year of the piControl from the contemporaneous abrupt-4xCO₂ timeseries. Despite this method not explicitly appearing in the literature, we include it here given the number of papers which cite anomalies with no method described (Dessler and Forster, 2018; Klocke et al., 2013; Lutsko et al., 2022; Meehl et al., 2020; Mitevski et al., 2021, 2023; Nijssen et al., 2020; Ringer et al., 2014; Zhou et al., 2021). In these studies the piControl may not have been pre processed before performing the anomaly calculation.

B. **Rolling:** Calculate a 21-year rolling average over the piControl and subtract the resulting timeseries from the contemporaneous abrupt-4xCO₂ simulation (Caldwell et al., 2016; Eiselt and Graversen, 2023; Po-Chedley et al., 2018; Qu et al., 2018; Zelinka et al., 2020). 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. 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.

Deleted: where possible.

Deleted: 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:

Deleted: A.

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: A, B, C, ... + Start at: 1 + Alignment: Left + Aligned at: 0 cm + Indent at: 0.74 cm

Formatted: Font colour: Black

Formatted: Font colour: Black

Deleted: 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,

Deleted: w

- 1360 C. Linear: 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-4xCO₂ timeseries ([Andrews et al., 2012](#); [Armour, 2017](#); [Bloch-Johnson et al., 2021](#); [Dong et al., 2020](#); [Flynn and Mauritsen, 2020](#); [Forster et al., 2013](#); [Lewis and Curry, 2018](#)).
- 1370 D. Long-term: 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-4xCO₂ experiment ([Chao and Dessler, 2021](#); [Jain et al., 2021](#); [Rugenstein and Armour, 2021](#)).

In addition to the steps described above, it is necessary to manually align the abrupt-4xCO₂ experiment with the piControl at the prescribed branch time. The branch time is the point at which an experiment – in this case the abrupt-4xCO₂ experiment – diverges from the piControl following an initial piControl spin up ([Eyring et al., 2016](#)). Branch alignment is important for the anomaly calculation, so that the correct part of the piControl is being subtracted from the abrupt-4xCO₂ experiment (although we note that branch alignment is redundant for the long-term average piControl anomaly method). 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 provided branch times in the model attributes are not always reliable. 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.

Following the data processing, we fit a linear regression over the first 150 years of the N and ΔT anomalies using two methods. First, for consistency with previous literature, we perform an OLS regression with ΔT as the independent variable. Additionally, we fit a TLS – alternatively called ‘orthogonal regression’ – line to the data. The key differences 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 ΔT -intercept (divided by two) as the ECS, the N -intercept (divided by two) as the radiative forcing due to doubling CO₂, and the slope as the feedback parameter.

1385 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 N and ΔT 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.

Deleted: .

Formatted: Font colour: Black

Formatted: Indent: Left: 1.27 cm, No bullets or numbering

Deleted: are not always reliable

Deleted: and for some models the correction may not be accurate

Deleted: ¶

... [3]

Deleted: line

Deleted: $RNDT$

Deleted: TAS

Deleted: $RNDT$

Deleted: TAS

3. Comparing the Gregory method data processing choices

We calculate 32 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 $N-\Delta T$ 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 Kolmogorov-Smirnov test to compare the inter-model ECS distributions between the remaining 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. We note here that our significance testing does not consider the shared code bases between some models (for a full model code genealogy see Figure 2 of Kuma et al. (2023)).

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. For individual ECS estimates across different paths (including a comparison to the Zelinka et al. (2020) calculated values) see Supplementary Table 2.

3.1 Global mean weighting

We compare two global mean weighting methods: by grid cell area and cosine of the latitude (Fig. 3a). To ensure a valid comparison, we keep the other data preparation choices constant by following the Baseline pathway: i.e. using rndt as N and the raw piControl for anomalies. Between the two global mean weighting methods, the median [min, max] multi-model ECS range of 3.88 [1.84, 5.67] does not change. For most models, the method of global mean weighting has little to no impact. 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%).

Deleted: 20

Deleted: 20

Deleted: choice of

Deleted: method

Deleted: (likely because these models have regular grids, Fig. 3a), as the median ECS difference across the ensemble when comparing weighting methods is effectively zero.

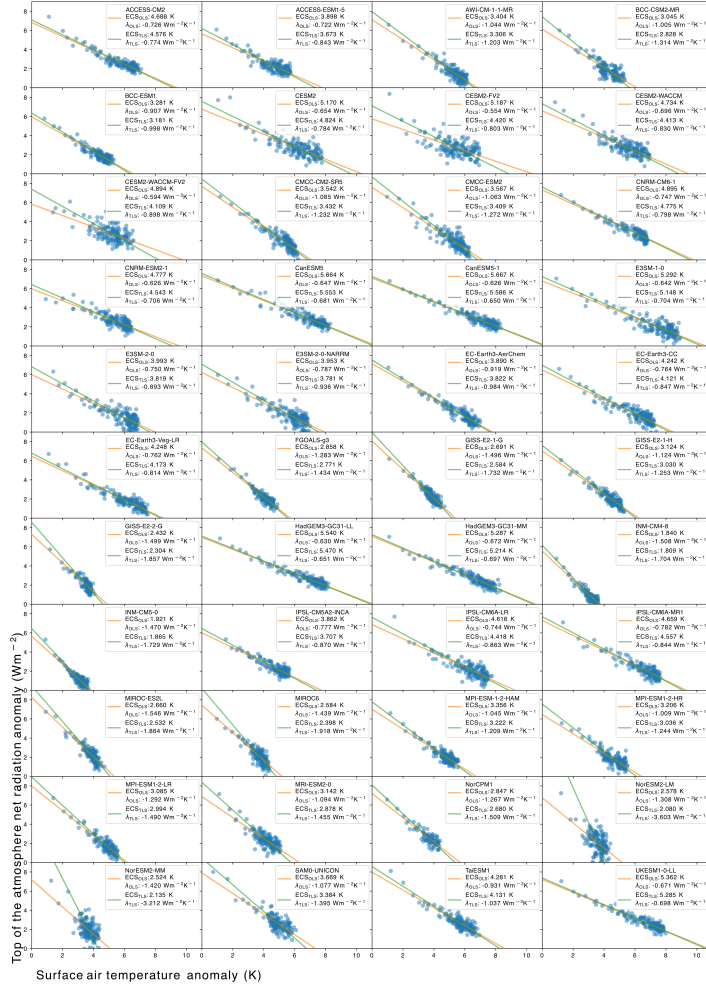
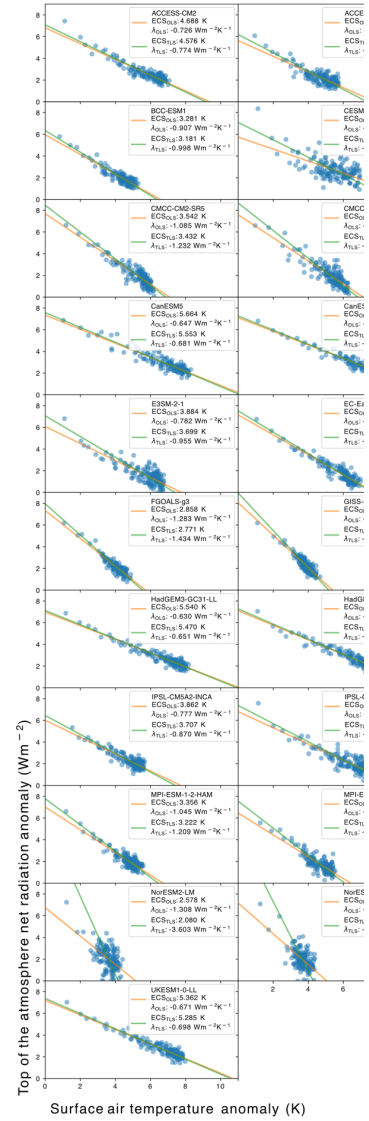


Figure 2. The Gregory plots calculated from the Baseline pathway for each model. The blue scatter plot represents the anomalies over time in the surface air temperature and radiative flux **anomaly** timeseries. The orange and green lines show linear fits calculated using ordinary and total least squares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



Deleted:

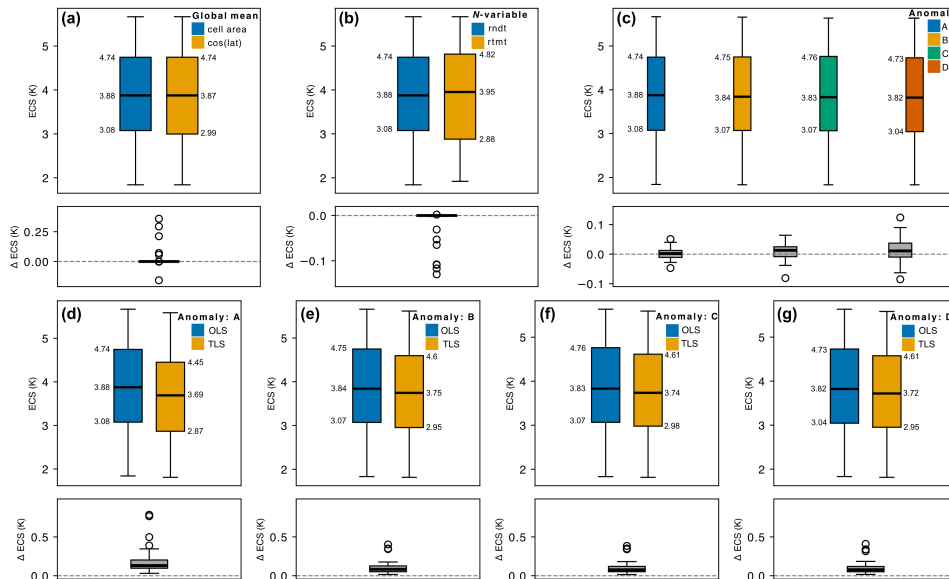
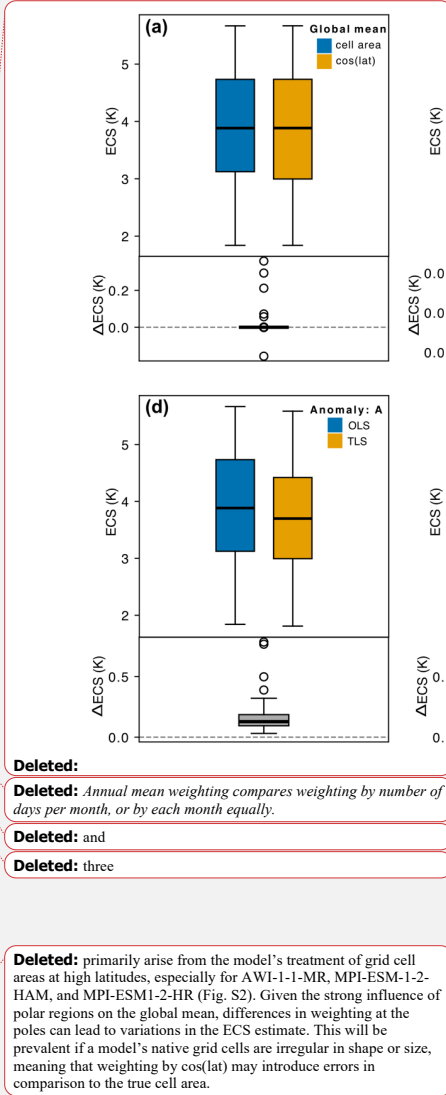


Figure 3. Each subplot shows the inter-model ECS range (upper) and differences between these ranges (lower) comparing the choices at each of the data preparation steps. **Boxplots show median first/third interquartile ranges (with ECS labelled in units of K), with whiskers showing the min/max excluding outliers, which are shown as hollow circles.** **a)** Global mean weighting comparing cell area and cosine of the latitude. **b)** **N-variable** compares the ECS calculated using **rmdt** or **rmtt**. **c)** Anomaly calculation method, with uppercase letters denoting the raw piControl, A, rolling mean, B, linear trend, C, and long-term average. **D.** **d), e), f), g)** OLS compared to TLS regression for the **four** anomaly methods. Note that the differences in range are always calculated as orange subtracted from blue (or green and dark orange subtracted from blue, in the case of plot c)). Additionally, note that the difference in ECS range for plots **d), e), f), g)** share a y-axis.

The differences in ECS for global mean weighting methods arise due to each model's grid cell configuration (grid information for each model can be found in Supplementary Table 1). Each outlier model uses native grid cells that are irregular in shape or size and thus cannot be approximated by **cos(lat)**. Our results suggest that, for these models, it would be an error to use the **cos(lat)** approximation instead of the native grid cell area variable to calculate the global mean.



Deleted:
Deleted: Annual mean weighting compares weighting by number of days per month, or by each month equally.
Deleted: and
Deleted: three

Deleted: 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.

900 In comparison to the two weighting methods we explore, many researchers may use various regridding techniques to calculate the global mean, which we do not consider in this study. Although regridding may be necessary for certain types of studies, we recommend weighting by the model's native grid and using the cell area when calculating the global mean for ECS preparation. This approach eliminates the need to verify if the model's grid is regular and is simpler than the cos(lat) approximation. In cases where cell area data is unavailable, cos(lat) can serve as an approximation, but it may introduce minor errors depending on the model's grid cell configuration. This is a clear demonstration of the importance of the cell area variable in CMIP submissions.

905 **3.2 Net radiative flux variable**

To compare the two net radiative flux variables, we again fix the remaining data processing choices as per the Baseline pathway. Of the 44 models in this study, only 35 have the rtmt variable available for both experiments, thus reducing the sample size for this comparison. We note, however, that all 44 models have the required TOA radiation variables meaning they are included for analysing the remaining data processing steps.

910 The median ECS for models using rndt and rtmt, respectively, is 3.88 [1.84, 5.67] and 3.96 [1.92, 5.67] (Fig. 3b). The choice of *N*-variable makes no difference for most models, except for, most notably, BCC-CMS2-MR, CESM2, and FGOALS-g3 with an ECS increase of 2% when using rtmt instead of rndt, BCC-ESM1 and CESM2-FV2 with an ECS increase of 3%, and INM-CM4-8 with an ECS increase of 6%.

915 The differences in ECS between rndt and rtmt are unexpected. A similarity between each of the above models is that they all have a low model top relative to the TOA, however not all models with a low top have a difference in ECS between *N* variables (for a list of all model tops see Supplementary Table 3). From an energy balance perspective, calculating the net radiative flux at different points in the atmosphere is unlikely to result in large changes in flux, given most of the Earth's energy imbalance is taken up by the ocean and land surface, with a common approximation of radiative flux being ocean heat uptake (Forster et al., 2021).

920 To investigate the differences in rndt and rtmt, we calculate the global annual average over 150 years of the piControl for both variables (see Supplementary Table 3 for all models). The models with differences in ECS between rndt and rtmt are the only models (apart from SAM0-UNICON) to have notable differences between rndt and rtmt (Table 1), with the largest absolute difference observed for INM-CM4-8 being 3.08 Wm⁻². Notably, many models have non-zero differences between rndt and rtmt values – even if these values are equivalent. In theory the piControl should have zero net radiative flux because it is at equilibrium, thus non-zero net radiative flux values are likely a result/indicator of accounting for model drift.

Deleted: 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.

Deleted: "areacella"

Deleted: Annual mean weighting

Formatted: Justified

Formatted: Font: 10 pt

While in theory the ECS should not change between the *rndt* and *rtmt* variable, we show that the variables can differ for some models. Given *rtmt* availability is limited depending on the model, our default suggestion is to use *rndt* for *N*.

Table 1. Global annual mean *N* averaged over 150 years of the piControl for *rndt*, *rtmt*, and the difference between the two variables (Wm^{-2}). Only the models with a change in ECS between variables are shown. For the rest of the models see Supplementary Table 3.

Model	<i>rndt</i>	<i>rtmt</i>	Absolute difference
BCC-CMS2-MR	-0.54	0.36	0.90
BCC-ESM1	-0.39	0.70	1.09
CESM2	0.70	0.09	0.61
CESM2-FV2	0.58	-0.03	0.61
FGOALS-g3	1.81	-0.12	1.93
INM-CM4-8	4.19	1.11	3.08

945

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 four methods that broadly reflect the different approaches between studies. These approaches form the basis for the labelled paths in Fig. 1: the Baseline, Rolling, Linear, and Long-term paths, which all use the cell area to calculate the global mean, *rndt* as the *N* variable, and differ only in their treatment of the piControl for the anomaly calculation.

The multi-model ECS ranges for the Baseline, Rolling, Linear, and Long-term paths are, respectively, 3.88 [1.84, 5.67], 3.84 [1.83, 5.66], 3.83 [1.83, 5.63], and 3.82 [1.83, 5.63] (Fig. 3c). To evaluate the impact of the different anomaly methods on individual models, we calculate the differences between the ECS of each model using different anomaly methods. We subtract from the Baseline path the Rolling, Linear, and Long-term paths (Fig. 3c). We observe a wider spread in the differences in ECS between the Baseline and Long-term paths compared to the Rolling and Linear paths. The largest percent difference for individual models is for NorESM2-MM which reduces by 3.4% (0.09 K) between the Baseline and Long-term paths. In comparison, the largest percent difference between the both the Rolling and Linear paths and the Baseline is 1.6% (0.05 K for MPI-ESM1-2-HR for Linear, and 0.04 K for NorESM2-MM for the Rolling path).

Studies which compute anomalies relative to a smoothed, averaged, or linear piControl 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).

Formatted: Font: 10 pt

Formatted: Indent: Left: 0 cm

Formatted: Font: 10 pt

Formatted: Indent: Left: 0 cm

Deleted: 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.

Deleted: three

Deleted: methods

Deleted: 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 (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.

Deleted: Previous

Deleted: s

Deleted: climatological mean or

Deleted: trend

2000 which refers to a long-term unforced trend in state variables. Since these anomaly 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).

2005 Unforced experiments, like the piControl, are typically used to diagnose model drift (Gupta et al., 2012, 2013; Irving et al., 2021). However, 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-4xCO₂ 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-4xCO₂ and piControl experiment, we would expect that the Baseline, Rolling, and Linear paths implicitly removes model drift following the subtraction. Calculating the anomaly relative to the piControl long-term average, however, does not account for biases that may be introduced by model drift.

2010 In addition to model drift, the correlation between N and AT is another approach of comparing the anomaly calculation methods. The median absolute correlations across all models for the Baseline, Rolling, Linear, and Long-term paths are respectively 0.88 [0.57, 0.95], 0.93 [0.64, 0.97], 0.93 [0.65, 0.98], and 0.93 [0.65, 0.98]. The differences in correlation likely results from a reduction in variance for the Rolling, Linear, and Long-term paths in comparison to the Baseline. For AT , the variance is less sensitive to the anomaly calculation method, with median variances across all models being 0.77, 0.76, and 0.73, and 0.70 for the Baseline, Rolling, Linear, and Long-term paths, respectively. However, for N , the median variances show a more substantial difference: 0.81, 0.70, 0.71, and 0.70 for each respective path.

2020 While the differences in correlation and variance between anomaly methods has minimal impact on the ECS estimates for an OLS fit, we observe more notable differences when comparing an OLS and TLS fit (Fig. 3d,e,f,g). The median differences between OLS and TLS for the Baseline, Rolling, Linear, and Long-term paths are 0.13 K [0.03, 0.79], 0.08 K [0.02, 0.4], 0.08 [0.02, 0.39], and 0.08 K [0.02, 0.41], respectively. Applying a trend or climatology to the piControl prior to the anomaly calculation reduces scatter between variables, thus increasing the absolute correlation compared to the Baseline pathway.

2025 Based on our anomaly method analysis we recommend that future climate sensitivity studies apply either a rolling average or linear trend to the piControl. We favour these two methods due to their implicit treatment of model drift (in comparison to the long-term average method), and due to their larger absolute correlation and avoided artificially inflated variance (in comparison to the raw piControl method) which provides improved alignment with the assumptions that underpin the linear regression. We note here that choices in drift correction method may have a larger impact on anomalies calculated over historical simulations relative to abrupt-4xCO₂ experiments, which may warrant further study. When choosing more specifically between the rolling average and the linear trend method, we recommend the 21-year rolling average. This

Deleted: 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

Deleted: each of the anomaly calculation methods we investigate will implicitly remove the model drift following the subtraction

Deleted: 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.

Deleted: Standard and Alternative

Deleted: which retains the raw piControl for the anomaly calculation method...

Deleted: TAS

Deleted: 80

Deleted: 78

Deleted: 73

Deleted: Standard, and Alternative paths,

Deleted: RNDT

Deleted: 83

Deleted: 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).

Formatted: Subscript

1080 method has been used to compare both CMIP5 and CMIP6 model ensembles (Caldwell et al., 2016; Zelinka et al., 2020),
1090 providing consistency with existing literature.

Deleted: ¶

... [4]

3.4 Linear regression method

2085 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 ΔT 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,
1090 across CMIP6 models, this choice is not arbitrary, given the median slope (λ) across models is affected by the choice of
independent variable; 0.88 W/m²/K when using ΔT , and 0.74 W/m²/K when using N (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.

Deleted: . The

Deleted: 89

Deleted: TAS

Deleted: RNDT

For OLS to provide a reasonable fit, the data must meet two key conditions: there should be a clear dependent variable, and
2095 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 as an error in
this study.

2100 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-4xCO₂
experiment from the HadSM3 slab ocean model. In comparison, we observe substantial scatter across a range of CMIP6 models
(Fig. 2), indicating that the original assumption of minimal scatter does not hold for the more complex fully coupled ESMs
2105 developed since 2004. This suggests that the original justification of OLS is worth reconsidering.

Deleted: 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),

Deleted: ing

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
2110 model simulations, the leading relationship between radiative flux and temperature is not always evident from the timeseries
alone. This is particularly true for the strongly perturbed abrupt-4xCO₂ experiments, where the climate system is responding

Deleted: ,

Deleted: especially

to an imposed radiative forcing that is far more extreme than anything observable in the real world, making it difficult to identify a relationship with N lagging ΔT .

Deleted: .

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 avoid inflating the variability in the ΔT and N timeseries through the anomaly calculation method. The methods which apply a rolling mean or linear fit to the piControl experiment are suitable, for example. Otherwise, subtracting raw piControl runs would inflate the variability and decrease the absolute 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 (simulations of the recent past from approximately 1850 to 2014 (Eyring et al., 2016)) 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 N or ΔT as the perfect independent variable, OLS may not be the most robust regression method in this context.

Deleted: remove some of

Deleted: TAS

Deleted: RNDT

Deleted: climatology

Deleted: removes some of the variability from the timeseries and increases the correlation between the two variables.

Deleted: RNDT

Deleted: TAS

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 archive and access to data provided by the Earth System Grid Federation). Secondly, physical reasoning regarding ECS bias supports OLS. The climate sensitivity estimated as the ΔT -intercept from the GM is biased relative to the 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). Comparing an OLS and TLS fit, the median ECS reduces from 3.9 K to 3.7 K, with the percentage difference for individual models ranging from 1.4 % (0.08 K) for HadGEM3-GC31-LL to 24% (0.65 K) for NorESM2-LM. The reduction between linear fits is consistent with findings of Forster and Gregory (2006), who deliberately chose the regression method which gave the largest sensitivity estimate. The low bias of TLS likely arises given

Deleted: archival

Deleted: ESGF

Deleted: ECS

Deleted:

Deleted: s

Deleted: from

Deleted: have a known

Deleted: low bias compared

Deleted: This

Deleted: from the

TLS ~~weights the~~ earlier years of the regression ~~more heavily~~ compared to OLS. While TLS may introduce a low bias in ECS estimates, it is worth noting that this method could potentially reduce the low bias in effective radiative forcing (ERF) observed in studies that calculate ERF using OLS over the full 150-year simulation period (Forster et al., 2016; He et al., 2025; Lutsko et al., 2022; Smith et al., 2020).

Deleted: favouring

Deleted: , 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; Lewis and Curry, 2018; 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.

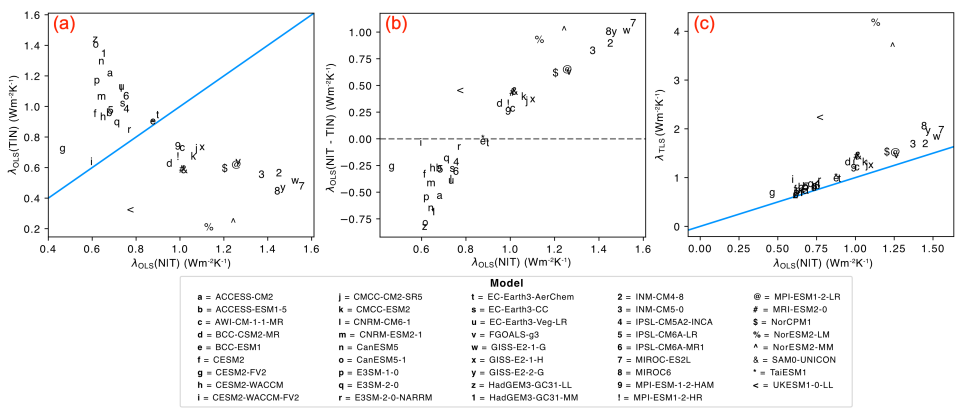


Figure 4. **a)** The slope (λ) of each CMIP6 model calculated using ordinary least squares (OLS) regression with ΔT as the independent variable (x-axis) and N_{λ} as the independent variable (y-axis). Blue line shows the linear relationship required for the choice of independent variable to make no difference. **b)** y-axis showing the difference in slope for each CMIP6 model between the OLS regression based on ΔT or N_{λ} as the independent variable. x-axis is the same as (a). Dashed line at $y=0$. **c)** The slope of the linear regression fit for each 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-4xCO₂ experiment here instead of the historical simulation. Each axis has units of $Wm^{-2}K^{-1}$.

Deleted:

Deleted: RNDT

Deleted: TAS

Deleted: RNDT

Deleted: CMIP6

Formatted: Font: 10 pt

3.5 Uncertainty range for individual ECS estimates

Calculating uncertainty over ECS estimates is an important step that is lacking from most of the climate sensitivity studies we cite in this paper. In the original study, Gregory et al. (2004) calculate uncertainty as the root mean square deviation from the OLS regression fit. More recent studies that 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 (identified in the following discussion).

To assess the level of inter-annual dependence across models, we calculate the autocorrelation function of the AT timeseries following the removal of a quadratic fit for the four different anomaly method pathways (Fig. S2). The autocorrelation function plots the correlation between a time series and its lagged versions, with particular focus on the correlation between adjacent timepoints. This analysis reveals two common temporal relationships exhibited by the models: an exponential decaying decorrelation, where the relationship between years decreases as more time passes, and an oscillating relationship, indicating that a periodic cycle is influencing the climate system.

While most models exhibit the exponential decaying decorrelation, the models which show an oscillating behaviour include 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, which have periods of between 3-6 years. For some of these models the process displayed depends on the anomaly calculation method, for example CMCC-CM2-SR5 shows an oscillating process for anomaly methods (B), (C) and (D), whereas when using the raw piControl for anomalies it shows an exponentially decaying process.

The oscillating behaviour 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 an irregular 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-4xCO2 time series also exhibit similar periodic behaviour for the same models (Fig. S3).

Deleted: While c
Deleted: has not been included in the steps we analyse, we feel this
Deleted: some
Deleted: The
Deleted: that do

Deleted: three
Deleted: S3

Deleted: of an autoregressive 1 (AR1) process
Deleted: some models exhibit oscillating behaviour consistent with an AR2 process.
Deleted: In particular,
Deleted: show oscillations, with
Deleted: AR
Deleted: AR2
Deleted: and
Deleted: AR1
Deleted: AR2 characteristics

Deleted: Thus

Deleted: AR2
Deleted: processes
Deleted: 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 ΔT_e and N_e 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.

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; [see Supplementary Table 4 for the confidence intervals calculated for each model using both bootstrap approaches](#)). 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. [S4](#)).

For most models the median ECS calculated using both the bootstrap approaches are larger than the original ECS estimate – for [40](#) models using the standard bootstrap, and [37](#) models using the block bootstrap. Additionally, for [27](#) models the median ECS calculated using the block bootstrap is larger than the median ECS calculated from the standard bootstrap. Most notably, [however, we find that](#) 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 [for an equilibrium climate](#).

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 ([thus replicating the method of Bloch-Johnson et al. \(2021\)](#)). 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

Deleted: TAS

Deleted: RNDT

Deleted: S5

Deleted: 38

Deleted: 35

Deleted: 31

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.

Despite the benefit of using years 21-150 on the confidence interval calculations, additional factors must be considered. Excluding early years from the regression is a common alteration to the GM (Andrews et al., 2015; Armour, 2017; Bloch-Johnson et al., 2021; Dai et al., 2020; Dunne et al., 2025; Lewis and Curry, 2018). However, the exclusion of the first 20 years results in a reduced absolute correlation between N and ΔT . For years 1-150 and 21-150, respectively, the median absolute correlation is 0.85 [0.49, 0.94] and 0.63 [0.3, 0.86]. The reduction in absolute correlation is most important when considering the choice of linear regression fit, given the difference between the inter-model ECS distribution using OLS and TLS is larger when using years 21-150 compared to years 1-150.

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-4xCO₂ 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 modelling groups, we recommend plotting the autocorrelations of the ΔT and N 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-4xCO₂ experiments are run to 300 simulation years instead of the previously required 150 years (Dunne et al., 2025).

- Deleted: modeling
- Deleted: TAS
- Deleted: RNDT

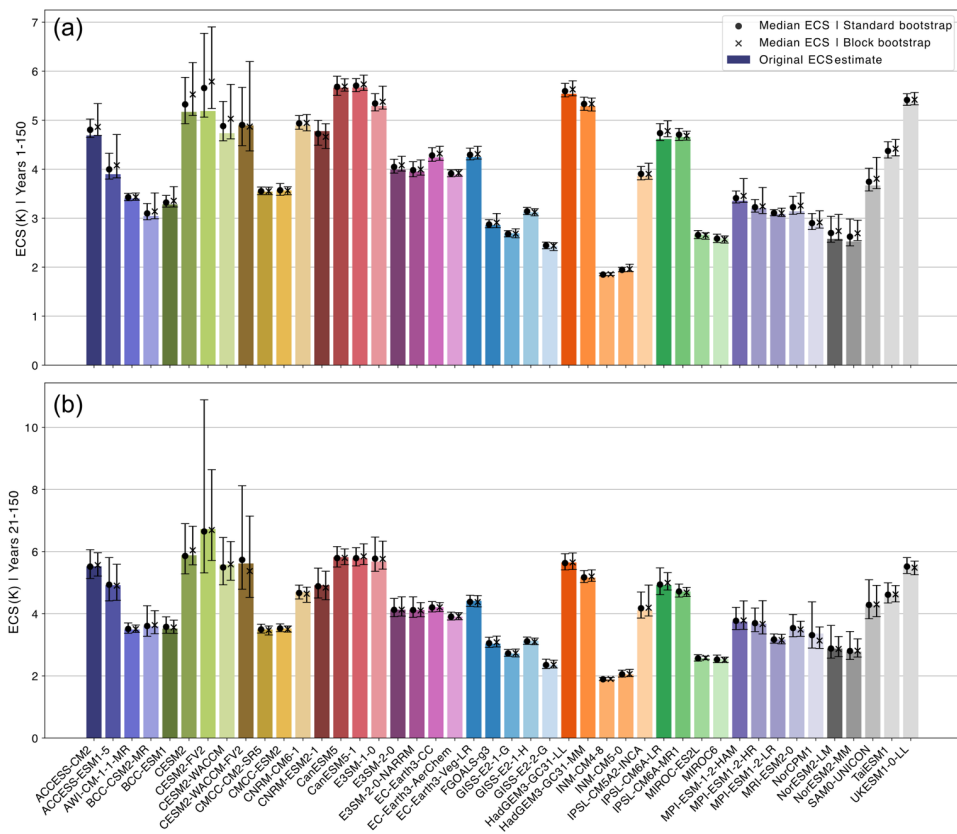
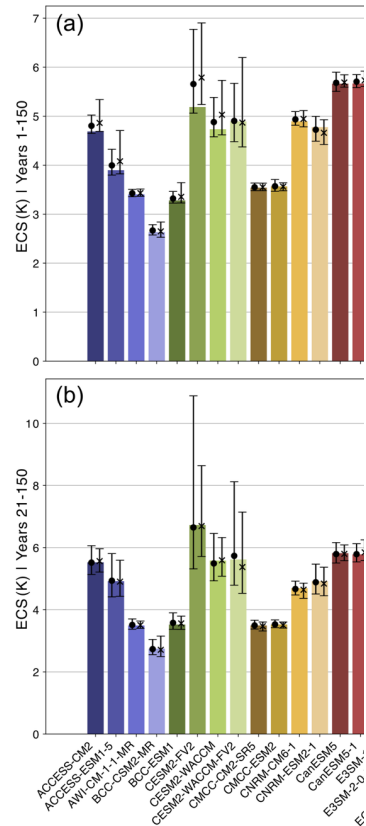


Figure 5. ECS uncertainty using an ordinary least squares fit. **a)** ECS estimates for each model using the **Baseline** Gregory Method, using years 1-150. Bars represent 95% confidence intervals, with medians calculated using a simple bootstrap (solid circle) and a moving block bootstrap with a block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



Deleted:

Deleted: baseline

Deleted: RNDT

Deleted: TAS

4. Discussion and conclusions

For each of the 44 CMIP6 models in this study, we compare 32 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.

Deleted: 41

Deleted: 20

Differences in ECS estimates arise, however, when comparing a subset of CMIP6 models. At each step, the largest individual model ECS differences are 11% for global mean weighting, 6% for N-variable, 3% for anomaly method, and 24% for linear regression method. Additionally, 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 rolling climatology or linear trend applied to the piControl, resolving some of the differences between OLS and TLS, likely due to the increase in absolute correlation compared to the raw piControl.

Deleted: 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

Deleted: .

Deleted: method

OLS has traditionally been the default linear regression fit 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.

Two additional aspects of ECS estimation which we do not investigate in this study are: the choice of CO₂ perturbation experiment, and using different time periods for the regression. Despite the ECS metric being defined as the response to CO₂ doubling, research typically uses CO₂ quadrupling 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 CO₂ 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-2xCO₂ 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 CO₂ 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.

Deleted: One step that we do not include in this study is

Deleted: . Using CO₂ quadrupling intends

2835 The landscape of ECS estimation is set to change for CMIP7, following the recommendation for modelling groups to extend
the abrupt-4xCO₂ experiment requirements from 150 to 300 simulation years (Dunne et al., 2025). This extended simulation
is expected to narrow the gap between GM-estimated ECS and the results from ESMs run to near-equilibrium (Dunne et al.,
2020; Rugenstein et al., 2020). A longer simulation will likely increase the ECS when calculated over the full 1-300 years,
potentially affecting comparability to previous CMIP generations. Given these changes, we recommend that future studies
applying the GM to CMIP7 data calculate the ECS based on both 1-150 years and 1-300 years. Computing these two values
will allow comparison to CMIP5 and CMIP6, provide further evidence of inconstant feedbacks (Rugenstein et al., 2020), and
2840 allow the research community to evaluate more thoroughly the merits and limitations of the linear relationship currently used
for ESC estimation.

Based on our findings, we provide recommendations for standardising the GM (Table 2) and a checklist of what to include in
future climate sensitivity research. Our standardisation framework details the steps involved, the alternative steps we
investigate, our proposed recommendations, and associated caveats. We acknowledge that not all studies applying the Gregory
2845 method have the ECS as their primary focus, and researchers may make alternative choices for their analyses that we have not
explored. We therefore include a checklist to ensure that, at minimum, 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
2850 standardised Gregory method calculation.

Checklist:

- ☐ Provide public access to all code used in the analysis
 - ☐ Clearly describe all data preparation steps in the methods section, including:
 - ☐ All variables used
 - ☐ Any differences from the recommended standardisation
 - ☐ Order of operations
 - ☐ Verify each model's grid configuration (to inform global mean weighting method)
 - ☐ Calculate the ECS based on both an OLS and TLS regression
 - ☐ For CMIP7, calculate the ECS based on both years 1-150 and 1-300
 - ☐ Calculate uncertainty around individual ECS estimates
- 2855
- 2860

- Deleted:** a set of
- Deleted:** for
- Deleted:** (Table 1)
- Deleted:** These detail the steps, choices at each step, our recommendations, and the caveats in those recommendations.
- Deleted:** so there may be alternative choices
- Deleted:** make
- Deleted:** i
- Deleted:** do
- Deleted:** At a minimum, we recommend that

Formatted

Formatted: List Paragraph, Bulleted + Level: 1 + Aligned at: 0.63 cm + Indent at: 1.27 cm

2875

Table 2. 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	rlilplfl	Use the first by default, although ideally calculate the ECS for all available ensemble members to quantify the sensitivity to different realisations, initialisations, forcing, and model physics.
Global mean weighting	Cell area Cosine of latitude	Cell area (areacella)	<u>This step is less of a ‘choice’ and should instead be informed by each model’s grid cell configuration. Using a model’s native cell area variable is the simplest method of avoiding potential errors due to incorrectly approximating with cos(lat).</u>
<u>Net radiative flux variable</u>	<u>rndt (= rsdt – rsut – rlut)</u> <u>rtmt</u>	<u>rndt</u>	<u>There should be little difference between these two variables. Models with a difference between rndt and rtmt could be investigated further or removed from the ensemble.</u>
Anomaly calculation	Subtracting from the abrupt-4xCO ₂ : a. Raw piControl b. 21-year rolling average c. Linear trend <u>d. Long-term average</u>	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 <u>N</u> 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.

Formatted: Justified

Deleted: 1

Deleted: Where cell area is not available, cos(lat) is a useful approximation, although it can affect the ECS by around 10%.

Formatted: Font colour: Black

Deleted: Annual mean weighting

Deleted: Weight each month equally

Deleted: Weight by number of days

Deleted: For precision, although it effectively makes no difference.

Formatted: Font: Not Italic

Deleted: Weight each month by number of days

Deleted: RNDT

2880

Code and data availability

Code required to conduct the analysis is available at <https://doi.org/10.5281/zenodo.15485520> (Zehrung and Nicholls, 2025).

2895 All data used in this study are publicly available. The raw CMIP6 ESM data (Eyring et al., 2016) can be downloaded from the
Earth System Grid Federation (<https://aims2.llnl.gov/search/cmip6>, ESGF LLNL Metagrid, 2025).

Deleted: USA portal of the

Author contributions

2900 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

2905 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,
2910 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

2915 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 FT240100306](#))
and the Australian Government through the National Environmental Science Program. [MM acknowledges funding via the](#)
2920 [Australian National Environmental Science Program – Climate Systems Hub](#). ZN acknowledges funding from the European
Union's Horizon 2020 research and innovation programmes (grant agreement no. 101003536) (ESM2025).

Deleted: ¶

References

Deleted:

... [5]

- Danabasoglu, G. and Gent, P. R.: Equilibrium Climate Sensitivity: Is It Accurate to Use a Slab Ocean Model?, <https://doi.org/10.1175/2008JCLI2596.1>, 2009.

- 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.
- 2970 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.
- 2975 [Dunne, J. P., Hewitt, H. T., Arblaster, J. M., Bonou, F., Boucher, O., Cavazos, T., Dingley, B., Durack, P. J., Hassler, B., Jukes, 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, *Geosci. Model Dev.*, 18, 6671–6700, <https://doi.org/10.5194/gmd-18-6671-2025>, 2025.](#)
- 2980 Eiselt, K.-U. and Graversen, R. G.: Change in Climate Sensitivity and Its Dependence on the Lapse-Rate Feedback in $4 \times \text{CO}_2$ 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
- 2985 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.
- 2990 [Forster, P. M., Richardson, T., Maycock, A. C., Smith, C. J., Samset, B. H., Myhre, G., Andrews, T., Pincus, R., and Schulz, M.: Recommendations for diagnosing effective radiative forcing from climate models for CMIP6, *J. Geophys. Res. Atmospheres*, 121, 12,460–12,475, <https://doi.org/10.1002/2016JD025320>, 2016.](#)
- 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.
- 2995 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.
- 3000

Deleted: Dunne, J. P., Hewitt, H. T., Arblaster, J., Bonou, F., Boucher, O., Cavazos, T., Durack, P. J., Hassler, B., Jukes, 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.

- 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.
- 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, J., Perlwitz, J., 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.
- He, H., Soden, B., and Kramer, R. J.: Improved Estimates of Equilibrium Climate Sensitivity from Non-Equilibrated Climate Simulations, <https://doi.org/10.22541/essoar.175157564.42459435/v1>, 3 July 2025.
- 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.
- Kuma, P., Bender, F. A.-M., and Jönsson, A. R.: Climate Model Code Genealogy and Its Relation to Climate Feedbacks and Sensitivity, *J. Adv. Model. Earth Syst.*, 15, e2022MS003588, <https://doi.org/10.1029/2022MS003588>, 2023.

- 3045 [Lewis, N. and Curry, J.: The Impact of Recent Forcing and Ocean Heat Uptake Data on Estimates of Climate Sensitivity, *https://doi.org/10.1175/JCLI-D-17-0667.1*, 2018.](https://doi.org/10.1175/JCLI-D-17-0667.1)
- 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.
- 3050 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.
- 3055 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-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.
- 3060 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.
- 3065 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.
- 3070 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.
- 3075 [Nijssen, F. J. M. M., Cox, P. M., and Williamson, M. S.: Emergent constraints on transient climate response \(TCR\) and equilibrium climate sensitivity \(ECS\) from historical warming in CMIP5 and CMIP6 models, *Earth Syst. Dyn.*, 11, 737–750, *https://doi.org/10.5194/esd-11-737-2020*, 2020.](https://doi.org/10.5194/esd-11-737-2020)
- 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.
- 3080 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.

- [Richardson, M., Cowtan, K., Hawkins, E., and Stolpe, M. B.: Reconciled climate response estimates from climate models and the energy budget of Earth, *Nat. Clim. Change*, 6, 931–935, <https://doi.org/10.1038/nclimate3066>, 2016.](#)
- [Richardson, M., Cowtan, K., and Millar, R. J.: Global temperature definition affects achievement of long-term climate goals, *Environ. Res. Lett.*, 13, 054004, <https://doi.org/10.1088/1748-9326/aab305>, 2018.](#)
- 3085 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.
- 3090 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.
- 3095 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 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.
- 3100 [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., Olivé, D., Fiedler, S., Lewinschal, A., Mackallah, C., Dix, M., Pincus, R., and Forster, P. M.: Effective radiative forcing and adjustments in CMIP6 models, *Atmospheric Chem. Phys.*, 20, 9591–9618, <https://doi.org/10.5194/acp-20-9591-2020>, 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.
- 3110 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.
- 3115 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.
- 3120 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.
- 3125 Zehrung, A., and Nicholls, Z.: ECS Gregory method analysis. Zenodo [code], <https://doi.org/10.5281/zenodo.15485520>, 2025

Page 3: [1] Deleted	Anna Zehrung	01/10/2025 10:03:00
---------------------	--------------	---------------------

Page 5: [2] Deleted	Anna Zehrung	23/09/2025 11:03:00
---------------------	--------------	---------------------

Page 9: [3] Deleted	Anna Zehrung	23/09/2025 11:27:00
---------------------	--------------	---------------------

Page 16: [4] Deleted	Anna Zehrung	23/09/2025 14:08:00
----------------------	--------------	---------------------

Page 27: [5] Deleted	Anna Zehrung	26/10/2025 10:07:00
----------------------	--------------	---------------------