To the Editors of GMD,

Attached is our point-by-point response and manuscript updates to the reviewer reports of our article, *Standardising the "Gregory Method" for equilibrium climate sensitivity (egusphere-2025-2252)*. The difference document (with author tracked changes) is uploaded alongside these responses. We would like to thank the reviewers for the time taken to review our paper.

We are glad that you and the reviewers recognise the benefits of our analysis and recommendations of standardising the Gregory method. The reviewers recommended minor revisions and minor further analysis to address areas of the manuscript that lacked clarity or required further explanation. We have addressed this in the revision: adding an additional anomaly calculation method, adding a data preparation step (*N*-variable), providing a checklist in the conclusion, and enhancing clarity on the aims of this study (especially in the abstract and introduction). As documented in detail in our responses, we hope that we thereby address the reviewer comments.

In the responses below, the original reviewer reports are in black, while all our comments are in blue. We have also numbered all the reviewer comments and our replies for clarity. We have *quoted text from the manuscript in grey italics*.

We thank you and the reviewers for the time invested into our manuscript and hope that it will reach the high standards of *Geoscientific Model Development* upon revision.

Best regards,

Anna Zehrung (corresponding author on behalf of all authors)

Reviewer comments and replies

Topic Editor comment

I think the abbreviation "r1i1p1f1" in the table and in the text requires explanation that I ask to include in the revised version together with the response to the upcoming reviews. An iteration prior to the review process would unnecessarily delay the process, I think.

We thank the topic editor for this comment and have defined our meaning of model variant, including a description of what is meant by “r1i1p1f1”.

Relevant new text in (difference document) lines 1303-1313:

> *The model ensemble member describes 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.*

Reviewer 1

Comment 1

Given that the choice of annual averaging method (weighting all months equally versus weighting by number of days) makes essentially no difference to the calculation of ECS (at most 0.02 K), I found it strange that this choice made its way into your recommendation for standardization. My takeaway is that this choice doesn’t matter, so my suggestion would be to note that and remove it from the recommended standardization list (e.g,. Lines 22-24 in the abstract) so that there is one fewer thing readers will have to keep track of, making it more likely that they will actually follow your recommendations as well.

We thank the reviewer for this comment and appreciate the insight that reducing complexity in recommendations could increase the adoption of our standardisation in future studies. We have removed the annual mean weighting ‘choice’ from the steps we formally investigate (including in Fig. 1 and 3), and have instead included a short paragraph in the methods detailing our findings that the annual mean weighting method makes almost no difference, so we do not include it as part of the steps.

Relevant new text in (difference document) lines 1316-1321:

> *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

Comment 2

Regarding the area weighting, my takeaway from your findings is that for models on a regular latitude-longitude grid, weighting by grid cell area (areacella) and weighting by cosine(latitude) makes no difference. This is of course expected. It's really only those models with output on an irregular grid where weighting by cosine(latitude) produces a different global average than weighting by grid cell area. This is of course also expected, because for an irregular grid, the grid area does not scale like cosine(latitude), and to weight by cosine(latitude) is simply an error. I think it would be much more clear to frame it this way, rather than as a "choice" of area weighting method, which makes it seem like both are acceptable options. Your recommendation to always weight by grid cell area is good since that removes the need to check whether the output is on a regular grid or not, and because weighting by areacella is easier (with less to go wrong) than weighting by cos(latitude) anyway.

Thank you for this comment and recommendation to reframe our analysis and discussion of global mean weighting "choices". We now include specific information on each model's grid and resolution in supplementary (see Reviewer 2, Comment 3). We see that the 'outlier' models have a native grid (gn), meaning that your expectation that cell area and cos(lat) are more likely to differ for these models compared to those with a regular grid. We have updated our abstract and discussion accordingly and reframe our conclusions.

Relevant new discussion in (difference document) lines 1880-1908

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.

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.

Comment 3

Another common choice is calculating anomalies with respect to the long-term average of the piControl simulation. You note this in several places (Lines 101-104) and I thought this was what you mean when you describe taking a climatology (Lines 292, 349, 464. But your results seem to only compare using anomalies relative to the raw piControl (including interannual variability), a 21-year rolling mean, and linear trend (Fig. 3). You should add an analysis using long-term average of the piControl simulation. This choice is far more common than subtracting the raw (annually varying) piControl data, I think.

We thank the reviewer for this comment and have updated our study to include a fourth anomaly calculation method which subtracts the abrupt-4xCO2 experiment from a long-term average of the piControl. This change is reflected in the decision tree (Fig. 1), the ECS comparison figure (Fig. 3), methods, and anomaly discussion (Section 3.3).

We also agree that it may not be common to calculate the anomalies using the raw piControl, however given the number of studies which do not describe how they calculate anomalies. We make this clear explicitly in the text.

Relevant new text in (difference document) lines 1334-1335:

> *In these studies the piControl may not have been pre processed before performing the anomaly calculation.*

Comment 4

Lines 171-176: It is unclear what you mean by performing “branch alignment” or “correction” here, so please elaborate on what exactly you did and how much it matters. Overall, the need to correctly identify the branch point in order to accurately perform the drift correction you propose (subtracting 21-year rolling averages) should be emphasized more. It’s a step that needs accurate metadata or additional effort to make sure the assumed branch point is correct, and this should be noted and perhaps even made as a concrete recommendation for standardization. Alternatively, many authors just use the long-term average over the piControl in an attempt to avoid having to pay attention to the branch point.

We thank the reviewer for this comment and agree that “branch alignment” or “correction” may be unfamiliar terms for readers who have not downloaded and analysed CMIP6 experiments. We now include an explanation of branch alignment and correction and note the reviewer’s suggestion regarding the long-term average over the piControl as a method of reducing the need to align the piControl and abrupt-4xCO2.

See our response to Comment 3 for further discussion regarding the long-term piControl average and additional discussion.

Relevant new text in (difference document) lines 1367-1370:

> *The branch time is the point at which an experiment – in this case the abrupt-4xCO_2 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-4xCO_2 experiment*

(although we note that branch alignment is redundant for the long-term average piControl anomaly method).

Comment 5

Does the choice of temperature variable matter? While most studies use near-surface air temperature ("tas"), others use surface temperature ("ts") which represents the "skin temperature" (i.e., sea-surface temperature over open ocean, surface of sea ice, and surface of land). Does this choice make a difference to ECS, and if so can you make a recommendation to use "tas" rather than "ts"?

We agree that the choice of temperature variable matters, and that not all climate sensitivity studies are clear in the temperature variable they are calculating for the ECS. We are unfamiliar with a global temperature calculated using the surface (or "skin") temperature globally, as suggested by the reviewer, however there is a clear distinction between GMST and GSAT – where GMST uses 2m air temperature over land and ice and SSTs over ocean, while GSAT uses 2m air temperature, "tas" across all surfaces. While we choose not to investigate these differences in this study, given the IPCC AR6 WG1 (chapter 7) discusses this distinction at length and recommends that model-based estimates of the ECS be calculated using GSAT, we now include a paragraph in the methods describing the importance of explicitly describing which temperature variable is being calculated. We find that while some studies make a clear distinction between GMST and GSAT, other studies refer to the ECS as GMST, but do neither describe the variables nor methods used for calculating the global mean temperature, thus leaving ambiguity in the variable they are calculating.

Relevant new text in (difference document) lines 1249-1264:

> *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; Nijsse 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) used GSAT and the IPCC recommends model-based estimates using GSAT (Forster et al., 2021), we recommend calculating the ECS using GSAT rather than GMST.

Comment 6

I understand the focus on using the Gregory method applied to years 1-150 of 4xCO2 simulations, which has traditionally been the way people have estimated ECS in GCMs. However, it's becoming increasingly common to estimate ECS using regression over different set of years, for example (i) using years 21-150 which is thought to provide more accurate estimates of equilibrium warming by avoiding some of the initial curvature in the Gregory plot, or (ii) using years 1-300 when longer output is available (e.g., in LongRunMIP or, hopefully, in CMIP7). It would be good to comment on whether the choices you evaluate here also make a difference for ECS estimates using those different choices of years. You could use the available LongRunMIP simulations to test using years 1-300, for example. I imagine that the difference in OLS vs TLS regression methods might matter more when using years 21-150, but might matter less when using years 1-300. I'm not sure about the other choices you explore. But this analysis and associated set of recommendations will become important as alternative regression periods are chosen for evaluation of ECS in CMIP7 models.

We thank the reviewer for the suggestions to investigate the ECS using a regression set over a different number of years. We now include a discussion regarding the differences between OLS and TLS whether using 1-150 or years 21-150.

Relevant new text in (difference document) lines 2778 - 2783:

> *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.*

Regarding the LongRunMIP experiments, we agree that increasing the number of simulation years from 150 to 300 could impact ECS estimates. However, we choose not to pursue this investigation for the following reasons:

> (i) We are explicit in our experimental design that we do not explore alternate years as a formal Gregory method choice (line 114).
>
> (ii) LongRunMIP is only available for 15 of the 44 models investigated for this study, thus limiting the sample size of assessing the data processing steps within these models.

(iii) The upcoming CMIP7 model DECK will require abrupt-4xCO2 experiments to run for at least 300 simulation years. This could be an avenue for potential further study applying alternate regression years to the Gregory Method.

We also include further discussion of the implications of the longer abrupt-4xCO2 runs for CMIP7 (see Reviewer 3, Comment 2 & 19). In our conclusions, we now recommend future studies to calculate the ECS using years 1-150 and 1-300 for comparison.

Comment 7
You could also consider testing the available 2xCO2 simulations. Do the same recommendations apply to calculating ECS in those, or do some choices become more, or less, important? This seems less pressing than my recommendations above.

We thank the reviewer for this comment. We agree that differences in recommendation could arise from assessing abrupt-2xCO2 experiments. We have a paragraph in the discussion describing that we explicitly do not assess different CO2 perturbation experiments in this study (see lines 473-482). Given the CO2 perturbation experiment may impact the ECS estimate depending on the model (a potential result of ECS state-dependence), and given only 11 models ran an abrupt-2xCO2 experiment with the relevant variables (since it is not part of the CMIP6 DECK), we decide that ECS estimates are incomparable to those calculated using the abrupt-4xCO2, and that the sample size is too small for meaningful results relating to different data processing choices. Abrupt-2xCO2 experiments will be available through CFMIP in AR7 fast track experiments. We hope that a larger number of institutions run this experiment for comparison with abrupt-4xCO2.

Line specific:

Comment 8
Line 24: You should define piControl

We have updated to *preindustrial control simulation* for clarity.

Comment 9
Lines 40-41: I found this summary of ECS ranges confusing since they are comparing different things. The Charney estimate is an approximate range. The CMIP6 range quoted is simply the range of models. And the AR6 range quoted is the 5-95% range. All measure different things, so it is not correct to say that AR6 narrowed the range relative to the models (it is simply more narrow than the model range). AR6 did narrow the range relative to the 5-95% range reported in AR5 and previous reports. Please reword this to be more clear on these points.

We thank the reviewer for the comment and agree that our description of previous ECS estimates leads to confusion. We have updated the paragraph to be more specific with how each ECS estimate is calculated, and acknowledge that only the Forster et al.'s (2013) estimate is a true uncertainty range.

Relevant new text in (difference document) lines 56-60:

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).

Comment 10
Line 49: Another paper to cite here is doi: 10.1175/2008JCLI2596.1

We have included this citation in the text.

Comment 11
Line 94-104: This felt out of place here since its not really needed here and you discuss the anomaly calculation in more detail below. Later in the paper I found myself scanning back up to this section to see these details. I suggest moving this to where you discuss the anomaly method in more detail below.

Thank you for your feedback on the progression of logic in the introduction and methods. We have updated the text such that the detailed description of the anomaly calculations now appears in the methods and include only a brief description in the introduction.

Relevant new text in (difference document) lines 145-149:

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-4xCO_2 experiment.

Comment 12
Lines 113-117: That's reasonable not to evaluate how ECS values change when using different regression periods, as many other papers look into that already. However, as I noted above, it would be good to check whether your recommendations still apply when using different choices for regression period. The choice of OLS vs TLS regression in particular could matter more or less.

We thank the reviewer for this comment and advise them to see our response to Comment 6.

Comment 13
Lines 134, 202: Baseline, Standard, and Alternative pathways are mentioned here, but not yet described. Only later on Lines 262-264 are they described.

Thank you for this comment. We now introduce the labelled paths in the methods, along with their descriptions. We also renamed these paths to Baseline, Rolling, Linear, and Long-term, to reflect the anomaly calculation and reduce confusion. This was a necessary step given we added a fourth anomaly method (see response to Comment 3).

Comment 14
Lines 157-170: this feels redundant with the text on Lines 94-104. Also, you should include a fourth choice here (which is common in the literature): calculating anomalies with respect to the long-term climatological mean of the piControl simulation (either over the full length of that simulation, or over the century or so leading up to the branch point).

Thank you for this comment. See our Comment 11 response.

Comment 15
Lines 179-181: This is framed as if it needed investigation. But I think it simply has to be true that calculating anomalies before or after summing the variables makes zero difference since these are linear operations.

We thank the reviewer for this comment. We no longer include the timing of rndt as a 'choice', although given we now include two N-variables (see response to Reviewer 3, Comment 11), we still include the rndt calculation.

Relevant new text in (difference document) lines 1343-1325:

> *In our analysis, we explore approaches which either define N as a measure of the TOA* $rndt = rsdt - rsut - rlut$ *(Lewis and Curry, 2018), or as the explicit TOM radiative flux variable (rtmt).*

Comment 16
Lines 213-244: This should make more clear that using cosine(latitude) when the grid is irregular is simply an error (not a valid different choice).

Thank you for this comment. See our response to Comment 2.

Comment 17
Figure 2: Do you have a sense of why OLS vs TLS regression matters so much for some models, but not for others? Can you comment on under what conditions the choice matters?

Thank you for this comment. See Section 3.4 which describes the reasons for these differences. We find that the models with the most scatter have the lowest absolute correlations between RNDT and TAS, which will have a larger impact on the regression method choice. In addition, see our response to Comment 6 where we will be including further analysis on the differences between OLS and TLS when excluding the first 20 years of the regression.

Comment 18
Lines 274-282: I found this discussion confusing. Non-closure of the global energy budget does not necessarily cause model drift if the model is fully spun up. It instead just means that

there will be a top-of-atmosphere energy imbalance maintained in the piControl state, which balances that non-closure within the model.

Thank you for this comment. We have updated our discussion of model drift to reflect these insights.

Relevant new text in (difference document) line 2004:

> *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), which refers to a long-term unforced trend in state variables.*

Comment 19
Line 307: You mention calculating anomalies relative to the climatological mean here, but as I note above I don't think you've tested that case?

See our response to Comment 3, we have included an anomaly calculation using a long-term average and include this in all anomaly method analysis.

Comment 20
Line 307-309: I think aiming for consistency in how ECS was calculated in Zelinka et al. (2020) is a pretty strong argument. Could you expand this to explain what choices were made in Zelinka et al. with respect to global area and annual averaging methods?

We thank the reviewer for this comment. The Zelinka et al. (2020) study calculated the global mean using a function that computes area weights based on lat/lon bounds. We acknowledge that their study differs in some of the data processing steps compared to some of the choices that we make in our study. However, we recognise this more broadly in the methods where we identify that our order of steps may be different to other studies but that the lack of information in methods for many studies inhibits their replicability.

We now include a table in supplementary which includes ECS values for each path that we analyse in Figure 3, which also includes the Zelinka et al. (2020) values.

Relevant new text in (difference document) lines 1835-1836:

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

Comment 21
Lines 348-350: I do not follow how removing a climatological average (constant values) or linear trend would change the variability or the correlation between variables.

Thank you for this comment. Firstly, we note that here when we use 'climatology', we are referring to the rolling climatology of the 21-year rolling average. We have updated this to reduce confusion, especially since we now include an additional anomaly method calculated using a long-term average over the piControl.

Secondly, we find that using the N and T anomalies calculated using the 21-year rolling average or linear trend over the piControl increases the absolute correlation between the two variables. We would expect this compared to using the raw piControl because the alternative anomaly methods remove much, if not all, of the interannual variability from the piControl prior to subtracting from the abrupt-4xCO2 experiment. While the interannual variability within the abrupt-4xCO2 experiment remains, there is potential that 'noise' is reduced by excluding the variability from the piControl in the anomaly calculation method, thus increasing the absolute correlation between the variables.

Comment 22
You should cite doi: 10.1175/JCLI-D-17-0667.1, who discuss using Deming regression instead of OLS.

Thank you for this relevant suggestion. We have included this citation in the text.

Comment 23
It may be worth mentioning that choices of drift correction method will likely make much more of a difference for the calculation of anomalies in historical simulations (as a percent change), even if they don't matter for ECS calculation. This of course could use further study to compare those choices.

We thank the reviewer for this comment and we include this suggestion in our discussion.

Relevant new text in (difference document) lines 2036-2037:

> *We note here that choices in drift correction method may have a larger impact on anomalies calculated over historical simulations relative to abrupt-4xCO_2 experiments, which may warrant further study.*

Reviewer 2

Comment 1: Background
The introduction would benefit from a more thorough background on ECS estimation and the Gregory method. For example:

- The Gregory method was originally designed for slab-ocean models, using short (e.g., 20- year) spin-up periods.
- Clarify the concept of radiative forcing in the Gregory framework, especially the distinction between instantaneous radiative forcing and the effective forcing derived from regression.
- Include the rationale behind separating fast and slow feedbacks and how this influences the interpretation of the forcing term.
- Also note that other ECS estimation methods exist, such as the Fixed Sea Surface Temperature (FSST) or AMIP-style configurations, and briefly position the Gregory method in this broader context.

We thank the reviewer for this comment. We now include a paragraph in the introduction to contextualise the Gregory method in the broader literature context and provide reasoning as to why we primarily focus on the ECS compared to forcing in our study. More specifically, we

describe how the Gregory method is likely popular due to the relative simplicity of the linear relationship which allows for a single calculation to estimate the ECS, ERF, and feedback. Additionally, the method does not require the highly specific experiment configurations (like fixed SSTs or AMIP-style configurations) which are generally used to calculate radiative forcing. We note that the accuracy of the Gregory method is subject to debate, but that our study focuses on the practical application of this method and will leave discussion of its strengths and weaknesses to other works.

Relevant new text in (difference document) lines 107-117:

> *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.*

In response to the reviewer's first dotpoint, we also would like to emphasize that the Gregory et al. (2004) study included multiple models and experiments. As the reviewer notes, the study compared an atmosphere-only model simulated for 20 years following an abrupt-4xCO2 increase. However, the Gregory et al. (2004) study also analyses an AOGCM simulated for 90 years for both abrupt-2x and abrupt-4xCO2, and 1200 years for abrupt-4xCO2. The linear relationship underpinning the Gregory method has multiple uses across the literature, and we focus almost solely on the climate sensitivity aspect, as this has become a key comparative metric for CMIP ensembles and other ECS estimation approaches. The 20-year spin up slab ocean configuration is more consistent for studies which investigate radiative forcing (see also our response to Reviewer 3, Comment 18 for further radiative forcing detail).

Comment 2: Inconsistent variable naming

There are multiple inconsistencies in the use of variable names, which undermine clarity: • Sometimes "temperature" refers to ΔT (temperature anomaly), but this should be clearly defined.

> • The paper mixes generally used (e.g., N, T) and CMIP6-specific (e.g., TAS, RNDT) variable names. These should either be standardized throughout or clearly defined at first use.

Thank you for this comment and we have updated the text to use N and ΔT, for consistency throughout.

• The symbol λ is typically used in the literature for the climate sensitivity parameter, whereas α is often used for the feedback parameter. This distinction should be respected throughout the manuscript to avoid confusion.

We agree that in previous literature (e.g. Gregory et al., 2004) α was used as the standard naming convention for the climate feedback parameter. However, we choose to use λ for the feedback parameter to reflect more recent literature – much of which we cite in the manuscript – such as the Zelinka et al. (2020) study which calculates the ECS range for CMIP6 models, or the Sherwood et al. (2020) paper, which assesses climate sensitivity based on multiple lines of evidence. We feel that continuing with the more recent naming convention reduces confusion and makes our study comparable to others. Additionally, given the more recent use of λ as the feedback parameter, it could be unclear to readers if we re-define the ECS as λ.

Comment 3: Description of Model Data and Experimental Setup
A centralized and detailed description of the CMIP6 model data (resolution, grid , …) and experiments (4xCo2 and pi-Control setup) used in the study is currently missing in the methods section. I strongly recommend adding this.

We thank the reviewer for this comment and have updated the methods with a cursory description of the model resolutions and grids. Since the models vary widely in grids, we now also include a table displaying each model's specific grid and resolution, which we display in the supplementary material.

Relevant new text in (difference document) lines 1239-1241:

> *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.*

Comment 4: Extension of discussion section
Following points in the discussion would be beneficial:

• Although applying the Gregory method to fully coupled models is standard practice today (e.g., CMIP6), this is a methodological shift from the original approach by Gregory et al. (2004), who used a slab ocean. The linearity assumption may break down over long timescales due to deep ocean heat uptake and evolving feedbacks.
• Is standardization worth the complexity, if the impact on ECS is so small? How large is the effect compared to the uncertainty of ECS due to Gregory approximation method?
• Including recommendations for calculating uncertainty ranges in ECS estimates would also be valuable, as it would support standardization in future analyses.

Thank you for these detailed comments. We note here that the Gregory method uses both a slab ocean and fully coupled model, with the original aim of the Gregory method being to use a fully coupled model which does not have to be run to a full equilibrium. We describe this in the introduction (lines 51-57). We agree, however, that the Gregory method's linear assumptions may break over long timescales, with studies exploring altering the Gregory method by, e.g., excluding earlier years of the regression to account for inconstant feedbacks (lines 113-117).

The Gregory method remains widely applied across literature, and will likely be a key initial diagnostic in comparing the upcoming CMIP7 models. For these reasons, the Gregory method will likely remain a useful tool for calculating the ECS, and thus benefits from standardisation. Even though standardisation may make little difference in ECS estimates for some models (although it can make a larger difference for others), another key aspect of the study is calling for transparency for future studies which apply the Gregory method. Currently, a number of key climate sensitivity studies have limited details in methods reducing replicability.

Regarding the uncertainty calculation, we include this in the checklist in the conclusion to support standardisation. We also make uncertainty calculations more explicit in the abstract, introduction, and figure 1.

Line specific

Comment 5
Line 37: global mean surface temperature

We have updated this.

Comment 6
Line 44: Clarify that ESMs require a coupled ocean to simulate climate feedbacks and energy balance properly.

Thank you for your comment, we have updated the text to define ESMs as fully coupled atmosphere-ocean (see Comment 7). By definition an ESM has a coupled ocean.

Comment 7
Line 51: Define what is meant by a "fully coupled ESM."

We have updated to *coupled atmosphere-ocean ESM.*

Comment 8
Line 59: global mean net radiative flux

We have updated this.

Comment 9
Line 60: Define effective radiative forcing and distinguish it from instantaneous RF.

See our response to Comment 1 and our response to Reviewer 3, Comment 18.

Comment 10
Line 60–61: Use λ for ECS and α for the feedback parameter, as per standard usage in literature.

Thank you. See our response to Comment 2.

Comment 11
Line 61: …. is the global mean surface air temperature change …

We have updated this. Thank you for this comment, we have also added discussion clarifying temperature differences between GMST and GSAT as per Comment 5 from Reviewer 1.

Comment 12
Line 65: Gregory (2004) did not use a 150-year simulation—please clarify

We thank the reviewer for directing our attention to this oversight. Please see our response to Comment 1 for further detail on what experiments were included in the original study.

Relevant new text in (difference document) lines 107-108:

> *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 CO_2 quadrupling experiment (abrupt-4xCO_2) relative to the model's preindustrial control experiment (piControl) and calculate an ordinary least squares (OLS) linear regression of annual mean values of N against ΔT.*

Comment 13
Line 68: Explain that the Gregory method includes fast feedbacks (e.g., water vapor, clouds) in the forcing term.

See our response to Comment 1. Note that we almost exclusively focus on the ECS in this study.

Comment 14
Line 78: When referencing "other climate sensitivity estimates," specify which ones

We agree that we should be specific in this paragraph. We have updated the line so that we now describe the Gregory method as a reference method for comparing climate sensitivity based on alternate lines of evidence, such as observations, historical simulations, or palaeoclimate data.

Relevant new text in (difference document) lines 122-123:

…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).

Comment 15
Line 85–86: "… many …" Be precise: Eg. Did Gregory et al. (2004) describe their data processing in detail?

We thank the reviewer for this comment but are unsure how to increase specificity given we cite and separate the papers with varying degrees of data preparation in their methods. We are now explicit in the paragraph that the Gregory et al. (2004) paper did not describe many of the processing choices that we investigate.

Relevant new text in (difference document) lines 122-123:

Many studies lack transparency regarding these preparatory steps, leading to potential inconsistencies, amplified by the fact that Gregory et al. (2004) included limited description of data preparation steps in their study.

Comment 16
Line86: these 150 years are not used in all studies (e.g Gregory et 2004 used 20 years spin-up)

Thank you for the comment. See our response to Comment 1. We cite the studies here which do use the full 150 years, a standard in ECS literature.

Comment 17
Line 110: Could you explain differences in OLS and TLS here?

Thank you for the comment. We have updated the text to explain the key difference in OLS and TLS as the choice (or lack thereof) of an independent variable. We only briefly introduce the concepts here given we further explain them in the methods.

Relevant new text in (difference document) lines 154-155:

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.

Comment 18
Line 117: " …across literature" -> add references

We thank the reviewer for this suggestion and realise our wording causes confusion. We have updated the sentence from *… are already well-documented and widely cited across literature* to *…. are already well-documented and these studies are widely cited across literature*. With this change we show that the studies we cite in the previous sentence are the ones which are widely known and cited amongst ECS literature given their explorations of altering the Gregory method through, e.g., excluding earlier years of the regression.

Comment 19
Line 123: Emphasize that the paper also evaluates regression methods and uncertainty, not just data processing workflows.

We thank the reviewer for this comment and have updated the sentence to reflect each investigation we explore. We also have updated Figure 1 (the decision tree) to show that ECS uncertainty is included as part of the study, as well as the abstract, introduction, and conclusion.

Comment 20
Line 128: how is TAS defined (1,5m temperature?)

We thank the reviewer for this comment and have included the height (2m for CMIP6 models). See line 1243 in the difference document. As per a recommendation from Reviewer 1, we also include a short discussion of the necessity of explicitly stating the variables used, especially in relation to the common misinterpretations between GMST and GSAT.

Comment 21
Line 154: "… to use annual (rather than longer) time period mean" -> Discuss consequences of using longer (e.g., decadal) time means—does it reduce noise or bias estimates?

Thank you for this comment. Given we choose to use only annual means, and do not analyse the potential impacts of using a longer time period mean, we therefore cannot accurately discuss the consequences on the ECS given we have not included it in our study. We acknowledge that most, but not all (e.g. 10.1029/2020GL088852), climate sensitivity studies use the annual mean

Comment 22
Line 208: "…we find that the preparation choices matter for a subset of individual models"-> Be specific: Which models are affected?

Thank you for your comment. Given we discuss in detail the models which are affected in the following subsections, and given each subsection generally affects different models, we leave it to the reader rather than listing each model individually here.

Comment 23
Figure2:

- Improve plot resolution
- Use α instead of λ for the feedback parameter.
- Acknowledge that some models (e.g., ACCESS and WACCM) share codebases and may not be fully independent (in the main text?).
- Include confidence intervals for feedback and ECS.

We thank the reviewer for this comment. Regarding plot resolution, we have all our figures as vector rasterised PDF's which unfortunately reduce in resolution when input into Microsoft Word. Upon request we can supply the high resolution PDFs.

We prefer not to change the λ to α to avoid confusion with recent papers (see our response to Comment 2). We discuss the confidence intervals in Section 3.4, given we observe some confidence intervals are almost outside of the ECS estimate. We also acknowledge that some models share codebases and may not be fully independent, including citing this relevant study 10.1029/2022MS003588.

Relevant new text in (difference document) lines 1835 - 1836:

> *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)).*

Comment 24

Line 215: how can you see from Fig 3a, that this is likely because these models have regular grid

Thank you for this comment. We understand the source of confusion and adjust our sentence and figure citation to reflect that it is the outlier models shown in Fig. 3a (box and whisker plot outliers), which show a native grid.

Relevant new text in (difference document) lines 1847:

We compare two global mean weighting methods: by grid cell area and cosine of the latitude (Fig. 3a).

Comment 25

Line 216: “outliers”: do the outliers have an irregular grid?

Thank you for your comment. Yes, these models have a native (irregular) grid. Note that we include a table in our supplementary material describing each model’s grid and resolution. In addition, see Reviewer 1, Comment 2 response where we have our discussion of global mean weighting.

Comment 26

Fig 3, caption: “range” -> Define what “range” refers to (e.g., min–max, 95% confidence interval). And what is shown median or mean?

Thank you for this comment. We have updated the figure caption to include the specifics of what the boxplot is showing.

Relevant new text in (difference document) lines 1873-1874:

> *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*

Comment 27

Line 284-295: The comparison of OLS and TLS fits better in Section 3.4—consider moving it.

Thank you for your recommendation. We have included this comparison of OLS and TLS in this section because it discusses the results shown in Fig. 3 including the potential differences in anomaly calculation methods. Based on this suggestion, however, we have reframed this part of the discussion to focus on the correlation between N and T between the anomaly methods, and that an impact of a lower correlation arises when comparing OLS and TLS. Discussing these differences here is relevant to be able to conclude the subsection with an anomaly method recommendation.

Relevant new text in (difference document) lines 2026-2030:

> *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). 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.*

Comment 28
Line 314: ΔT instead of temperature.

We have updated. See our response to Comment 2.

Comment 29
Line 318: Clarify what makes temperature choice "not arbitrary" in CMIP6

We thank the reviewer for this comment and clarify that the choice is not arbitrary given the differences in slope depending on whether N or T is the dependent variable. We have updated the text to clarify this.

Relevant new text in (difference document) lines 2109:

> *However, across CMIP6 models, this choice is not arbitrary <u>given the</u> median slope (λ) across models is affected by the choice of independent variable*

Comment 30
Line 330: Explain why this assumption may not hold in fully coupled models.

Thank you for the comment. This comment refers to our assumption that minimal scatter does not hold for many fully coupled CMIP6 models. CMIP6 models include interactive ocean and atmosphere, thus increasing the complexity of feedbacks and interannual variability compared to the single slab ocean model used to assess the “minimal scatter” in 2004. The climate interactions in CMIP6 models can impact the scatter, or correlation, observed between N and T from the abrupt-4xCO2 anomalies.

Relevant new text in (difference document) lines 2122-2124:

> *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 developed since 2004.*

Comment 31
Line 335-337: can you explain?

These lines refer to the inability to observe a lead/lag relationship between N and T, especially for the highly perturbed abrupt-4xCO2 experiment, where the climate is being forced under such a strong scenario. In comparison, identifying a lead/lag relationship between N and T based on observations, with a comparatively much weaker forcing, a relationship may be more likely to be identified.

Relevant new text in (difference document) lines 2130-2132:

> *This is particularly true for the strongly perturbed abrupt-4xCO_2 experiments, where the climate system is responding 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.*

Comment 32
Line 351: What is meant by historical ensemble? And what is the difference to idealized abrupt CO2 simulation setups?

Thank you for your comment. We have clarified these in the discussion.

Relevant new text in (difference document) lines 2628:

> *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.*

Comment 33
Line 359: ESGF

Thank you for this comment but we require clarification because we have written ESGF in the paper. We have expanded this to write the *Earth System Grid Federation.*

Comment 34
Line 361: Provide numbers when stating that TLS provides lower ECS—by how much?

We thank the reviewer for their comment and we are now explicit with the ECS calculated using TLS vs OLS, rather than simply providing the differences in slope (feedbacks).

Relevant new text in (difference document) lines 2639 - 2641:

> *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.*

Comment 35
Fig 4: Legend is very hard to read. Move to the bottom.

We thank the reviewer for the comment and updated the figure legend accordingly.

Comment 36
Line 383: Provide references for "some climate sensitivity studies."

Thank you for this comment. Given most studies do not calculate an uncertainty range around individual ECS values, we have updated the text to *lacking from most of the climate sensitivity studies we cite in this paper*, and cite the papers which do calculate uncertainties in the following sentence.

Comment 37
Line 348: Clarify whether bootstrap is typical—e.g., Gregory et al. (2004) use standard error of the regression slope.

We thank the reviewer for this comment and clarify in the text that Gregory et al. (2004) use RMS deviation, whereas it is the slightly more recent studies which use the bootstrap approach, although any uncertainty estimate around individual ECS values is uncommon.

Relevant new text in (difference document) lines 2692-2693:

> *In the original study, Gregory et al. (2004) calculate uncertainty as the root mean square deviation from the OLS regression fit.*

Comment 38
Line 387: "…which does not hold for some models" - > Identify which models violate the independence assumption—most fully coupled models do have interannual autocorrelation.

We now include the following for clarity and leave it up to the reader.

Relevant new text in (difference document) lines 2697:

> *(identified in the following discussion).*

Comment 39
Line 391: Define AR(1), AR(2), etc., here or in the methods section for clarity.

Thank you for this comment. We describe autocorrelation more explicitly and remove our descriptions of AR1/2, instead describing the autocorrelation relationships we observe for clarity

Relevant new text in (difference document) lines 2701-2705

> *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.*

Comment 40
Line 418-422: Quantify the confidence intervals derived

Thank you for this comment. We quantify the intervals derived and include these in a supplementary table, as per line 2753: *(Fig. 5a; see Supplementary Table 4 for the confidence intervals calculated for each model).*

Comment 41
Line 463: Why do most models have no error? Are these the models which have a regular grid?

Thank you for your comment. See our response to Comment 3.

Reviewer 3

Comment 1
Lines 9–10: Please clarify that ECS is obtained by extrapolating to N = 0 in the N–ΔT regression, i.e., the ΔT intercept.

We have updated this in the abstract.

Relevant new text in (difference document) lines 9-10:

> *The ECS is determined by extrapolating the linear fit to N=0, i.e. the ΔT-intercept, indicating the point at which the system is back in equilibrium.*

Comment 2
Lines 10–11, 121–123: Given that abrupt-4xCO2 simulations may extend to 300 years, I wonder how the authors envision ECS estimation in CMIP7, and how best to compare results across phases.

Thank you for this comment. In line 442 we acknowledge that CMIP7 abrupt-4xCO2 experiments will be run for 300 years instead of the previously standard 150 years. However, we agree that the longer simulations will have an impact on how we compare CMIP7 to previous model generations. We include a more detailed discussion in the paper (in summary and conclusions) describing these changes in CMIP7 and the potential implications for ECS calculations and for our recommendations and proposed standardisation. We also recommend that future studies calculate the ECS using both 1-150 and 1-300 years to compare (in the concluding checklist).

Relevant new text in (difference document) lines 2841-2860:

> *The landscape of ECS estimation is set to change for CMIP7, following the recommendation for modelling groups to extend the abrupt-4xCO_2 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 allow the research community to evaluate more thoroughly the merits and limitations of the linear relationship currently used for ESC estimation.*

Comment 3
Lines 16–18, 460: Please emphasize that these sensitivities occur only for a very small set of models (outliers), while most models are not affected.

Thank you for your comment. We update the manuscript accordingly.

Relevant new text in (difference document) lines 17

> *individual outlier models exhibit notable differences.*

Comment 4
Lines 22–24, 455–459: While I appreciate the recommendations, I find they may make the ECS estimation unnecessarily complicated, especially given the finding of “no statistically significant difference” and “unlikely to see a meaningful difference in results.” Related concerns include: (1) areacella is not commonly used; (2) leap-year treatment varies across calendars (e.g., noleap, julian, gregorian); (3) checking the branching date

of each simulation is time-consuming—branch time metadata are usually included in CMIP6 but rarely in CMIP5.

Thank you for your comment. Note lines 145-148 in the manuscript, where we acknowledge that our study may include different data processing choices or order of steps compared to other researchers. However, these differences arise because many studies are not transparent in their methods, which inhibits replicability.

While the impact of some of the different steps we investigate do not result in a meaningful impact on the ECS estimate, the priority is understanding whether there *could be* an impact and, if so, what the impact is. For example, while in practice perhaps researchers understand that areacella is not commonly used, in the literature this is unclear given the number of studies which do not describe their global mean weighting methods entirely. It may be useful for future studies to consider the implications of the global mean weighting method they use, even if they choose not to use areacella.

Regarding leap-year treatment between model calendars, because our analysis is based on annual means (rather than daily or monthly means), differences in calendar conventions (e.g., Gregorian vs. no-leap) do not affect comparability. We calculate annual means by weighting each month by its number of days, so the annual value represents the average per day in the model year. For robustness, we also tested the simpler approach of weighting each month equally (1/12 per month) and found the effect on ECS to be negligible. Thus, our results are not sensitive to calendar differences, and models using different calendars remain directly comparable in this context.

While we agree also that branch alignment is time-consuming and challenging, this is not a valid argument to avoid the step in the data processing leading up to the regression. We describe branch alignment more explicitly in methods. Also note lines 174-176 where we state that validating branch information for CMIP7 would reduce the time spent on corrections after the model's original submission.

Our study has two goals: (1) to offer a standardised Gregory method for future studies, and (2) to promote transparency in all methods. This second goal may not be as clearly identified in the study as we would like. As a result, we increase clarity throughout the paper so that these two goals are clear to readers. Additionally, in the conclusion, we include a checklist (along with the standardisation recommendation) as a minimum that researchers should be providing to enhance transparency and replicability (such as describing all methods, sharing code, etc.).

Relevant new text in (difference document) lines 2870-2878:

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*

Comment 5
Lines 26–27: Did the authors mean that the "CMIP6 multi-model mean ECS appears not sensitive to these processing choices"?

Yes, that is correct. We have updated the text.

Comment 6
Line 60: Did you mean "the global mean radiative response $\lambda\Delta T$"?

Yes, thank you for the comment. We have updated this in the text.

Comment 7
Lines 65–67: Please confirm whether Gregory et al. (2004) actually used 150-year simulations.

We thank the reviewer for this comment and update the text to reflect the Gregory et al. (2004) experiment set up of a 90 year simulation (although the standard has since become 150 years). See Reviewer 2, Comment 1 for further details.

Comment 8
Lines 154–155, 248: Were leap years accounted for, given the different calendars used by models?

Please see our response to Comment 4.

Comment 9
Lines 157–170, 269–270 & Fig. 1: Why were anomalies relative to climatological monthly means not considered? Including a climatology-based method could show how model drift affects results.

Thank you for your comment. We now include an additional anomaly calculation method based on a long-term average over the piControl. See Reviewer 1, Comment 3 response for further details.

Comment 10
Lines 171–176: While aligning abrupt-4xCO2 and piControl at the prescribed branch time is useful, checking branch dates for each model is time-consuming. Would it be possible for modeling centers to standardize branching dates across experiments (e.g., r1) or to mark them more clearly in their piControl runs to simplify this step? The authors might consider recommending this.

Yes - this would be useful for modelling centers to standardise branching. We already describe this in lines 174-176.

Comment 11
Lines 178–181: Since the variable rtmt (TOA net radiative flux) is available for most models, I suggest computing rtmt for those models without the variable directly before preprocessing.

Thank you for your comment. Given rtmt is available for 35 models used in this study, we investigate the potential differences between using rtmt and our current calculation of rndt. From initial research on this topic, it seems that rtmt is computed as the 'Top of Model' (TOM), whereas our rndt value is calculated as the Top of Atmosphere (TOA). For some models, TOM and TOA can be different, and therefore could impact the energy balance between the variables.

We now include the choice of N-variable (rndt or rtmt) as a formal "choice" that we investigate and provide recommendation on. We have made updates to the abstract, introduction, Figure 1, Figure 3, discussion (section 3.2 now explores N-variable), conclusion, and include additional supplementary information on this choice.

Comment 12
Lines 234–235: Could the authors provide further explanation here, particularly regarding the different distribution of MPI-ESM1-2-HAM compared with the other two models?

Yes, we provide further explanation. Note that we also include a table in supplementary material detailing all model's grid and resolution information. Additionally, see our response to Reviewer 1, Comment 2, for further details.

Comment 13
Lines 250–251: Given the conclusions, is annual-mean weighting strictly necessary?

We thank the reviewer for this comment and note that this is a similar comment to Reviewer 1, Comment 1. See our response to this comment for details.

Comment 14
Lines 265–267 and elsewhere: Why not also report differences in the multi-model mean?

Thank you for your comment. We include the multi-model mean here and elsewhere for clarity in each subsection of the discussion (and label figure 3 for clarity).

Comment 15
Lines 306–309, 463–465: I find it odd to recommend anomalies relative to a climatological mean or linear fit initially and then just apply a 21-year running mean over the piControl, citing Zelinka et al. (2020). Widespread use is not, by itself, a sufficient justification.

Thank you for your comment. Here when we refer to a *climatological mean* we mean the rolling average, given a 'climatology' is often considered a 20-30 year window average. We understand this leads to confusion and will update the text accordingly. See our discussion in section 3.3 where we provide a more clear discussion on why to choose the rolling average, which is related to the treatment of model drift and increased correlation (although we also acknowledge that applying a linear fit to the piControl provides similar benefits).

Comment 16
Lines 335–337: Is this effect due to the imposed radiative forcing in the idealized abrupt-4xCO2 experiment? The argument could be clarified.

Thank you for your comment. That is correct that the lead/lag relationship between N and T is unclear from such a strongly perturbed forcing. We explain this in the text.

Relevant new text in (difference document) lines 2130-2132:

> *This is particularly true for the strongly perturbed abrupt-4xCO_2 experiments, where the climate system is responding 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.*

Comment 17
Line 361: It might be useful to report ECS values explicitly, rather than only feedback values.

Thank you for your comment. We include ECS values in addition to the feedback values.

Relevant new text in (difference document) lines 2639-2641:

> *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.*

Comment 18
Lines 363–364: In fact, previous studies (e.g., Forster et al., 2016; He et al., 2025; Lutsko et al., 2022; Smith et al., 2020) suggest that the Gregory method (OLS) underestimates effective radiative forcing.

Forster, P. M., et al. (2016). JGR: Atmospheres, 121, 12,460–12,475. https://doi.org/10.1002/2016JD025320

He, H., et al. (2025). ESSOAr. DOI: 10.22541/essoar.175157564.42459435/v1

Lutsko, N. J., et al. (2022). JGR: Atmospheres, 127, e2022JD037486. https://doi.org/10.1029/2022JD037486

Smith, C. J., et al. (2020). Atmos. Chem. Phys., 20, 9591–9618. https://doi.org/10.5194/acp-20-9591-2020

Thank you for these relevant citations. It is important to clarify that these studies all find that using the full 150 years of an abrupt-4CO2 experiment applied to the Gregory method underestimate the ERF in comparison to using either years 1-20 or 1-30 in the regression to calculate the ERF. This low bias indicates that it is the number of years used as inputs in the regression, not necessarily the Gregory method itself, which impacts the ERF estimates. It is for this reason that we focus almost explicitly on the ECS (and feedbacks) in this paper rather than publishing also ERF ranges, given it seems well known that there are more confident estimates of ERF compared to the full 1-150 year Gregory method. See also our response to Comment 1 from Reviewer 2.

The low bias of ERF, however, is an interesting consideration for the regression method itself. It may be possible that using TLS instead of OLS could reduce the ERF bias. We now include discussion accordingly.

Relevant new text in (difference document) lines 2643-2646:

> *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).*

Comment 19

Lines 436–438: It seems CMIP7 is prioritizing longer simulations rather than additional ensemble members—could the authors comment?

Thank you. See our response to Comment 2.