Comment on egusphere-2025-2252

This paper provides practical guidance for standardizing the calculation of Equilibrium Climate Sensitivity (ECS) using the Gregory regression method, based on abrupt-4xCO_2 simulations. Although the method is widely adopted in studies, its implementation has varied. The authors argue that small choices (e.g., area weighting, drift correction) can impact reproducibility and comparability. Their goal is not to redefine ECS estimation, but to provide a common framework for processing data, enabling consistent use in future model inter-comparisons. Overall, I recommend this paper for publication after revision. Below, I provide detailed comments and suggestions for improvement, both general and line-specific.

General Comments

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.

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.
- 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.

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.

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 slap 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.

Specific Comments

Line 37: global mean **surface** temperature

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

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

Line 59: global mean **net** radiative flux

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

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

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

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

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

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

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

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

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

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

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

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

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?

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

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.

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

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

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

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

Line 314: ΔT instead of temperature.

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

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

Line 335-337: can you explain?

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

Line 359: ESGF

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

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

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

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

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

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

Line 418-422: Quantify the confidence intervals derived

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