Response to Referee Comments on "Constraining net long term climate feedback from satellite observed internal variability possible by mid 2030s"

We appreciate the referees' constructive feedbacks. Below, we provide our detailed responses and the modifications made in response to their comments and suggestions.

- **Referee 1.**

Using CMIP6 model simulations, the authors derive an emergent constraint that relates feedback from internal variability (IV) to forced feedback. They show that there are statistically significant relationships across models between components of IV feedback and forced feedback: a strong relationship for SW, a weaker relationship for LW, and an even weaker, but still significant and meaningful relationship for the net feedback. Using this relationship and combining it with observed internal variability, they show that more observations are needed in order to use this finding to actually constrain ECS. As an alternative to waiting for more satellite data, the authors extend the satellite record back in time by applying a correction model to reanalysis radiative fluxes, and thus constrain ECS to 2.5 K [90 % CI: 2.14 – 3.07 K], which is lower than current estimates from models, the IPCC, or Sherwood et al. 2020. Future satellite observations could be integrated into their method to update the estimates, and the authors quantify the quality of the constraint as a function of the number of observed years. The paper is well-written, well-presented, clearly states its goal and provides evidence to support the claims, guiding the reader through the argumentation. The statistical methods are sound and used in an appropriate way. While I do have a long list of comments and questions, I want to stress that I enjoyed reading the paper and consider it a beneficial addition to the research on feedbacks and climate sensitivity. I have one main point to raise in criticism of this paper, which I will present in the following.

My main comment can be summarized as "What about the pattern effect?"

From the methods section, I understand that the feedback parameters are calculated as differential feedback parameters (referring to Rugenstein and Armour 2021, `https://doi.org/10.1029/2021GL092983`, please confirm if this interpretation is correct). All feedback parameters are estimated as the slope of N(T). For λ_{ab}, which time period is used for the regression? The full 150 years? We know that λ_{ab} changes considerably over time, both over the 150 years period (which is accounted for if the full 150 years are used for the regression), but also after this (e.g. Rugenstein et al. 2020, `https://doi.org/10.1029/2019GL083898`). According to that paper, ECS estimated from the 150-year span is an underestimate of the true ECS by 17 % in models. Would this affect the ECS estimate that the paper gives?

Further uncertainties may arise when leaving the model world. The historical simulations which are used to compute λ_{it}, are not capable of reproducing the observed SST patterns (e.g. Wills et al. 2022, `https://doi.org/10.1029/2022GL100011`). It is currently debated if the observed pattern of strong Western Pacific warming will continue or switch to stronger warming in the Eastern Pacific. This uncertainty implies enormous uncertainty for ECS (Alessi and Rugenstein 2023, `https://doi.org/10.1029/2023GL105795`). The point that I'm trying to make with these explanations is that it may very well be that the connection between λ_{it} and λ_{ab} is very different in the real world and models. While models produce El-Nino like patterns both in the present and future, the real world has warmed more La-Nina like until now, and we don't know how it will continue. Since these patterns are tightly linked to λ_{ab}, the model results may not be applicable to the real world. This would be a major problem for the emergent constraint that the paper develops, because an implicit assumption of the emergent constraint approach is that the statistical relationship that is found in the models is applicable to reality.

I would like to ask the authors to discuss this uncertainty. In particular, do you think it affects the ECS range that is determined? If yes, how? If no, why not? If the authors agree that this could add substantial uncertainty, I propose mentioning this also in the last part of the abstract, which currently suggests that all uncertainties (except for the biogeochemical feedback) are accounted for in the 5 – 95 % CI.

Your interpretation regarding the feedback parameters calculated as differential feedback is accurate, and we derived them using 150 years of simulations.

While we recognize that ECS estimates derived from 150-year experiments may underestimate the true ECS due to time-dependent feedbacks, as shown by Rugenstein et al. (2020), it's equally important to consider that ECS estimates from 4XCO2 experiments frequently overestimate ECS compared to 2XCO2 experiments due to nonlogarithmic forcing, feedback CO2 dependence, and feedback temperature dependence, as demonstrated by Bloch-Johnson et al. 2021. Since the ongoing debate on the interaction of these two effects lies beyond the scope of our study, we chose the standard method as it represents the most straightforward approach and serves as a basis for comparison and analysis. This observation has been incorporated

into the Materials and Methods section, line 85.

Indeed, an essential assumption in our study is that the relationship between λ_{it} and λ_{ab} observed in models also holds true in the real world. You rightly highlight that, so far, the observed warming pattern may differ from model projections, with the real world displaying more La Niña-like trends while models generally exhibit El Niño-like patterns. This discrepancy introduces significant uncertainty for our λ_{ab} and ECS estimates. Given the importance of this caveat, we have include this potential discrepancy between models and observations in the abstract, discussion section and final notes of caution in our manuscript.

In addition, I have other comments:

- l. 12, 21, 335 – 337: What biogeochemical processes does this refer to? Can you specify? I wonder if they are relevant for ECS, as the carbon-cycle does not matter for this concept of fixed CO2 concentration, and vegetation changes are not included in the definition of ECS.

  While several definitions of ECS do not include biogeochemical processes, we align with the IPCC's definition as detailed in Forster et al. 2021:

  "Feedbacks in the Earth system are numerous, and it can be helpful to categorize them into three groups: (i) physical feedbacks; (ii) biogeophysical and biogeochemical feedbacks; and (iii) long-term feedbacks associated with ice sheets. ... biogeophysical/biogeochemical feedbacks (e.g., those associated with changes in methane, aerosols, ozone, or vegetation; Section 7.4.2.5) act both on time scales that are used to estimate the equilibrium climate sensitivity (ECS)."

  The implementation of these biogeophysical and biogeochemical processes varies among models. For instance, some models use static vegetation, some implement aerosol indirect effects, and others prescribe ozone and/or methane concentrations. Given this variability, we consider important to highlight this restriction in our manuscript.

- l. 82 paragraph: As mentioned before, please state which years are used for the regression of λ_{ab};

  We used 150 years for the regression of λ_{ab}. This information has now been added to our manuscript on lines 73 and 85.

- l. 83 – 84: Is there a particular reason for subtracting the control state? I wonder, because a constant shouldn't affect the slope estimate. It wouldn't hurt the calculation, but I'm curious.

  There is no specific rationale for subtracting the control state, as a constant should not influence the slope estimate. To address

this, we have revised the sentence in line 84 to: "We calculate forced climate feedbacks using linear Ordinary Least Squares (OLS) regression coefficients derived from 150 years of annual global averages of R and T from abrupt4xCO2 simulations."

- l. 125: It is not immediately clear to me what was done here by "randomly permuting". Were the R and T time series randomly matched (e.g. R from model 1 realization 1 and T from model 2 realization 1), and were the feedback parameters subsequently computed from these randomly matched time series? Am I right in assuming that only complete time series were permuted, not individual values in the time series?

  We acknowledge that our initial description of the method lacked detailed explanation. To clarify, our approach did not involve randomly pairing the R and T time series to compute feedbacks from these pairs. Instead, the method estimates the likelihood of obtaining a correlation as high as, or higher than, the observed correlation between internal variability and forced climate feedbacks in climate models. Specifically, we randomly permuted the feedback datasets, disrupting the correspondence between models for internal variability and forced climate feedbacks (for example, by pairing the internal variability feedback from one model with the forced climate feedback from another). We then recalculated the correlation coefficient using this shuffled data. This procedure was repeated 10^5 times, generating a null distribution of correlation coefficients that reflects the range of values expected if no real relationship exists. Finally, we compared the observed correlation to this null distribution to estimate how often a correlation of equal or greater magnitude could arise by chance, providing a p-value as a measure of statistical significance. We have added more details to this method in line 135 in the manuscript to clarify the procedure.

- Fig. 2: I am not sure that Fig. 2 is really needed. To me as a reader, the only relevant information is the likelihood of obtaining the correlations by chance, which is mentioned in the text; the full distribution is not so interesting, and the differences between the blue, red, and black lines are anyway hard to grasp. While I take no issue with this figure, I believe that it could be removed without loss of information; however, I would like to see the likelihood to obtain the correlations for the net feedback parameter by chance in the text, I only found this information for LW and SW.

  We agree that the figure is not essential and have decided to remove it. We have added the likelihood of obtaining the correlations for the feedback parameters by chance in lines 144 and 155.

- l. 150: Given that the first term is 0.43 and the last one is -0.72, does that mean that the internal SW feedback outperforms the internal LW feedback as a predictor for the forced LW feedback (by having a strong anticorrelation)? I find that interesting.

  That is an interesting observation. However, it is important to clarify that the referred terms are not correlations, but rather the covariance divided by the product of standard deviations of the internal variability and forced climate net feedbacks. The actual correlation between internal variability and forced climate longwave feedbacks is 0.68, and the anticorrelation between shortwave internal variability and longwave forced climate feedback is -0.64. These values indicate that the internal variability shortwave feedback does not outperform the internal variability longwave feedback as a predictor of forced climate longwave feedback.

- l. 174 – 175: So if the SW is the strongest contributor, that means that it comes down to clouds (unsurprisingly). Do you think the poor model representation of clouds is a problem for that?

  Indeed, extending the period for estimating internal variability feedbacks improves the relationship between internal variability and forced climate shortwave feedbacks. Given that shortwave feedbacks are closely linked to clouds, this improvement indicates that clouds respond to natural variations in surface temperature similarly to how they respond to external radiative forcing in models. This consistent misrepresentation of clouds actually benefits our methodology as it allows us to use internal variability observations to constrain uncertainties in forced climate feedbacks. However, if models had more accurate cloud representations, the uncertainties in forced climate feedbacks might be reduced, potentially diminishing the applicability and need for this emergent constraint methodology.

- l. 182: Models have no measurement uncertainty, but EBAF does. Is the uncertainty that arises from the satellite measurements (and also from the temperature data, but I assume that will be less important) taken into account? Would it affect the estimate of ECS or is it too small to make a difference? When combining the measurements from CERES and ERBE, is it problematic that the satellite changes, e.g., are there inconsistencies or steps?

  We acknowledge that our current methodology does not explicitly incorporate satellite measurement uncertainties. In our analysis of internal variability feedbacks, the noise in the TOA fluxes time series arises from both natural variability and measurement errors. Consequently, when we regress TOA fluxes

against surface temperature, the confidence intervals (CIs) inherently account for the total variability in the data, encompassing both natural fluctuations and measurement noise. Explicitly adding measurement uncertainties to the CI calculations could result in double-counting, thereby inflating the confidence intervals unnecessarily. This inflated observed internal variability feedback would, in turn, broaden the estimate of forced climate feedback after applying the emergent constraint. Consequently, using this broader estimate of forced climate feedback would lead to a wider range of ECS uncertainties.

Regarding the combination of satellite datasets, there are challenges involved, such as spatial biases in radiative fluxes, changes in the observing system used in the data assimilation process, and unrealistic variability in radiative fluxes due to the absence of volcanic aerosol effects. While these issues are indeed significant, they fall outside the scope of our study. Instead, we relied on the methodology implemented by Allan et al. (2014), who addressed these challenges and provided the merged dataset we used in our analysis.

- l. 187: The values are almost all well below 1 %. Doesn't that mean that less years might also be enough, if we think that, e.g., 5 % would be sufficient?

  Indeed, fewer years might suffice; however, this is not the case for the most recent potential relationship, as illustrated in now Figures 2a and 2b. It would be interesting to verify whether the relationship holds for periods shorter than 51 years when including more years beyond 2014 if the historical simulations were extended. Nevertheless, we prefer not to include this observation to avoid speculation.

- Fig. 3 caption: Unclear what is meant by "n – 2014", what is n here? Should I read it as "n to 2014" or "n minus 2014"?

  We clarify this now in the manuscript. The term "n – 2014" should be read as "starting year to 2014".

- l. 194 – 195: The suggested approach here is to wait for new satellite observations, but by then we will also have longer historical simulations. Can't we just run your analysis on the historical simulations again in 14 years, circumventing the whole problem of using the emergent relationship from one period with observations from another? It's still an interesting question to ask, but I don't see the practical necessity to use the "old" emergent relationship 14 years from now.

  Yes, if we gather enough observations and extend the historical simulations to cover a 51-year period that includes the observational data, we could potentially apply the method without

relying on the "old" emergent relationship. We have now incorporated this consideration into line 212.

- l. 206 - 216: This seems to be in disagreement with the results of Fig. 4 (d). In Fig. 4 (d) you show that when taking at least 40 years, it doesn't matter which period one picks, λ_{it} will always be the same. So λ_{it} does not depend on the chosen period if the period is long enough. λ_{ab} obviously doesn't depend on the chosen period either. So how can the relationship between λ_{it} and λ_{ab} depend on the chosen period (that's what I read from Fig. 4 a and b)? I have a hard time reconciling this. In addition, Gregory and Andrews 2016 (https://doi.org/10.1002/2016GL068406) show that historical feedback has varied quite a bit, although they use shorter than 40-year periods for their regression.

  We appreciate your observations and recognize that Figure 4 (Now Figure 3) may cause some confusion. In now Figure 3, we address two distinct questions with our analysis.

  First, we test the hypothesis: "It is possible to use observations from one 51-year period with the model relationship between λ_{it}, from a different 51-year period, and λ_{ab} to produce the emergent constraint." To validate this hypothesis, we used a proof by contrapositive approach, examining whether all model relationships using 51-year periods to estimate λ_{it} would be statistically similar if the hypothesis were true.

  In Figures 3a and 3b, we present the slopes and intercepts of all potential 51-year model relationships between 1850 to 2014 and compare them with the confidence intervals of their means. The results indicate that the 51-year relationships are indeed statistically different, providing evidence to reject the hypothesis. As the reviewer correctly notes, while λ_{ab} remains unchanged, λ_{it} varies for each 51-year period, indicating that those λ_{it} from models are statistically different.

  Second, we test the hypothesis: "It is possible to use the available 14 years of observations (2001-2014) with the model relationship between λ_{it} (1958-2014) and λ_{ab} to produce the emergent constraint." We again used a proof by contrapositive approach. If the hypothesis were true, then all possible 14-year λ_{it} values within the period 1958-2014 would be statistically similar to the λ_{it} of the full period 1958-2014. Using ERA5 reanalysis, we calculated all possible 14-year λ_{it} values within the range 1958-2014 and compared them to that of the full period 1958-2014 (Figure 3c). The results provide evidence to reject the hypothesis. Additionally, we estimated the length of an observation period that would lead to a λ_{it} statistically similar to that of the period 1958-2014, finding that 40 years are required (Figure 3c).

In summary, the information presented in now Figure 3 addresses different questions and should be read with care. Figures 3a and 3b compare all modeled slopes and intercepts from the linear regressions between λ_{ab} and 51-year λ_{it} within the period 1850 and 2014 with their mean, while Figure 3c determines the observation window size needed to estimate a λ_{it} statistically similar to that from the period 1958-2014 using data within the same period.

We would like to clarify that, upon reviewing the referee's comment, we realized our initial statement may have seemed categorical, implying that all 51-year modeled λ_{it} periods are statistically different. In reality, there is some degree of similarity, as slopes and intercepts from adjacent time periods can be quite similar in certain cases. However, this pattern does not consistently apply across the entire analyzed period. Consequently, the extent of allowable discrepancy between the time periods used to calculate internal variability from observations and from models for generating an emergent constraint on forced climate feedbacks varies depending on the specific periods compared. We have updated the manuscript to reflect this clarification for improved understanding.

- Fig. 4 (a) and (b). How can the starting year be 1980 and higher for 51-year periods?

  As noted by the reviewer, having a starting year of 1980 or later for 51-year periods is indeed inconsistent. We have identified and corrected an error in the computing code that led to this issue. The figures have been updated to reflect the correct starting year.

- Does it surprise you that the relationship between λ_{it} and λ_{ab} varies strongly in time?

  The substantial variation in the relationship between λ_{it} and λ_{ab} can be attributed to the previously mentioned error. Even after correcting for this, some variation persists, likely related to the specific internal variability present in each 51-year period. We chose not to include this in the manuscript as it remains speculative and requires further investigation beyond the scope of our study.

- l. 250 – 252 and Fig. 5 (a): +/- 2 W/m^2 seems not negligible compared to interannual variability of global-mean TOA flux, which I would expect to vary by less than 10 W/m^2. How can it be that the correlation with CERES-ERBE is still so high (0.99)? It means that 98% of the variance of the ERA5 feedback parameter is explained by CERES-ERBE, so only

2 % is left for the error, which seems low given that the error gets up to +/- 2 W/m^2.

We believe the referee is referring to the fact that 98% of the variance in the ERA5 TOA fluxes (rather than the feedback parameter) is explained by CERES-ERBE, leaving only 2% for the error. The correlation coefficient of TOA fluxes between ERA5 and CERES-ERBE measures the strength and direction of their linear relationship. A systematic offset between the two datasets does not significantly affect the correlation coefficient because it does not alter the way the datasets co-vary over time. Therefore, it's possible to have a high correlation coefficient (0.99) despite differences in their absolute values. Additionally, it is important to emphasize that a high R^2 value pertains to the variance in ERA5 TOA fluxes, not their exact values. Consequently, even with the absolute differences caused by the error margin, the relative variability and trends of the datasets remain aligned, resulting in a high correlation and coefficient of determination.

- l. 277 – 279: I don't understand the method here. A probability density function of which quantity? What values are sampled from this distribution? I had expected one value for λ_{it} from ERA5, obtained from regressing over the 40-year period, not a whole distribution. What am I missing? This seems like a central point of the paper and maybe deserves another sentence or two to clarify the method.

  We recognize that the initial description was difficult to read and included unnecessary details for deriving the emergent constraint. To clarify, we have simplified the explanation in the manuscript line 300.

  In summary, our method uses a Monte Carlo simulation to estimate forced climate feedbacks. We begin by generating a predictor variable dataset from a truncated normal distribution, based on the 40 years of adjusted ERA5-CERES-ERBE observations and their confidence intervals. Applying a linear model ($\lambda_{ab} = m\lambda_{it} + b$), we calculate the confidence intervals for predicted forced climate feedbacks for each predictor value. To address prediction uncertainties, we sample from these confidence intervals, resulting in a new dataset of predicted forced climate feedback values. The emergent constraint is then characterized by the probability density function of this dataset.

- Is there a reason for presenting the results from this analysis as small insets in Fig. 1? It seems like one of the main outcomes of this paper is hidden in a small inset. If showing it in Fig. 1, I would prefer the y-axes of the main plot and the inset to be aligned.

  After evaluating several alternatives, we determined that retaining the insets offers the clearest representation of feedback distri-

butions. To ensure consistency and improve visibility, we aligned the y-axes of the main plot and the inset.

- l. 296 – 301: The list of limitations seems short. In addition to my questions about the pattern effect potentially limiting the results of this study, I think it may be beneficial to discuss further limitations. In particular, the emergent relationship is obtained from model simulations using models, hoping that this relationship would translate to the real world. However, most models that contribute to this relationship simulate λ_{it} values way outside the observed range (see Fig. 1 f). Could this limit the results?

  We have now expanded our discussion of limitations to include both the pattern effect and the assumption that emergent relationships from models apply to the real world. Concerning the issue of models simulating λ_{it} values significantly outside the observed range, we previously noted that "uncertainties in the model emergent relationship, as illustrated in Figure 1f, reduce confidence in the emergent constraint". We have revised this statement for clarity to: "Uncertainties in the model emergent relationship within the adjusted ERA5-CERES-ERBE period, due to most models simulating λ_{it} values significantly outside the observed range (Figure 1f), reduce confidence in the emergent constraint."

Minor comments:

- l. 72 – 75: the half-sentence "incorporating a more extensive. . . " appears twice

  The error has been corrected.

- l. 161: The use of the word "assuming" makes sense here, but made me stumble, because it sounds like it's a prerequisite to run the hypothesis, when it's actually rather the null hypothesis; "testing for" or something similar would have been clearer to me.

  We changed the word "assuming" to "where the null hypothesis posits" in line 180.

References

Bloch-Johnson, Jonah et al. (2021). "Climate Sensitivity Increases Under Higher CO2 Levels Due to Feedback Temperature Dependence". In: *Geophysical Research Letters* 48.4, e2020GL089074. DOI: `https://doi.org/10.1029/2020GL089074`.

Forster, P. T. et al. (2021). "The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity." In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. Chap. 7. DOI: 10.1017/9781009157896.009.