Response to Reviewer 1

Thanks to the reviewer for thorough reading and thoughtful points. I've endeavoured to address the issues raised in the revised paper as detailed below.

Major Issues

1)... the model consists of Eqs. B1 and B2, together with B3 for including a transient forcing. An optimization procedure is applied to estimate the parameters based on a given data set (HadCRUT) and cost functions (B4-B6). I do not completely understand how this optimization defines the parameter distribution (the model ensemble),

I have expanded the discussion of what the MCMC optimization does and how the parameter distribution is derived from the cost function

i.e. how is the distribution exactly derived from the optimization, and how do(es) the distribution(s) look(s) like (a Figures of the pdfs may be helpful in this respect)?

The relevant figure is in the supplemental material, Figure S3 - which shows the individual and pairwise parameter distributions of the posterior ensemble.

Furthermore (random order):

(i) in L210 it is stated that CO2 concentrations enter the cost functions, but H(t) and D(t) seem to be heat fluxes (L215)?

Sorry - that was a typo, the treatment of CO2 is specific to the companion paper to this one, which included carbon cycle feedbacks. This paper only considers the thermal part of the model. Typo corrected.

(ii) How do H and D relate to the parameters r1,r2 in B2 (are they the same)?

Now included specific equations for D and H

(iii) Are T_p (B3) and P (B1) the same?

Yes, now T_p throughout

(iv) How does F (B3) relate to R (B2) (or how are eq. B1 and eq. B2 coupled in the model).

The coupling is in the original multi-timescale energy balance model as detailed in Millar et al 2017 (now written explicitly). The particular solutions of the temperature and radiative response to a step change in forcing can be written as a sum of exponential decays (again, now shown explicitly)

(v) How (where) does the non CO2 forcing factor f_r (L208) enters the equations.

Now included an explicit expansion of the historical forcing timeseries F(t) which defines f_r

2) I'm wondering how important the non-CO2 forcing agents (L207) and the factor f_r are for the results. How much of the variability of the control (present day) climate is explained by the non-CO2 forcing, and how is the non-CO2 forcing prescribed in the scenarios (it seems that all is represented by a constant f_r)?

The constant f_r is a scaling factor (as now made clear by equation B5), so the non-CO2 forcing is not constant over time - but you are correct that we are assuming that there is only a single degree of freedom in optimization. Though we could break down this forcing further - our primary goal is not to attribute the response to different forcings, and this formulation allows conceptually for uncertainty in the historical forcing timeseries while minimizing the number of degrees of freedom in the optimization.

3) Not much attempt is made to evaluate/validate the models behaviour under RCP scenarios. So far (as far as I can see) it is only shown that the model reasonably reproduces the HadCRUT data (where it is constrained to), and gives a response within the CMIP range. It would be useful to show that the model can reasonably reproduce the RCP8.5/RCP2.6 response of one particular model if the parameters are constrained by the present day simulation of the same model. This would give more confidence to the obtained results.

I've included an additional supplementary plot S4 to fit the pulse-response model to historical simulations in the CMIP archive with future ensemble projections for RCP2.6 and 8.5, with some discussion in the methods. The technique generally performs well (i.e. future projections fall within the distribution), with the exception of a couple of models which show little or no long term warming response (CCSM4, FGoals, FIO-ESM - models which share some fraction of their codebase). I've added a caveat to this effect, but I'm broadly happy that the technique is producing reasonable probabilistic fits to historical CMIP data.

4) One main result is that residual drift may explain 'surprising' results regarding EffCS and TCR in CMIP. From Fig. 3 we see that different CMIP models seem to exhibit different magnitudes of residual drift. I'm wondering whether the simple model result regarding the effect of drift can be qualitatively checked by comparing respective simulations.

I've attempted to show this qualitatively in a new Figure 5, and supplementary Figure S5. The former is an attempt to 'correct' the control drift uncertainty in the estimation of TCR by estimating baseline temperatures from the 1pctCO2 simulation itself, and background trends from the control simulation. The plots show - for all 3 metrics, but particularly for TCR - that the correlation with 21st century warming under RCP8.5 can be improved a little using this baseline correction, supporting the hypothesis that control drift is an issue for the estimation of sensitivity metrics. Of course this is just an estimate - itself noisy given the relatively short regression,

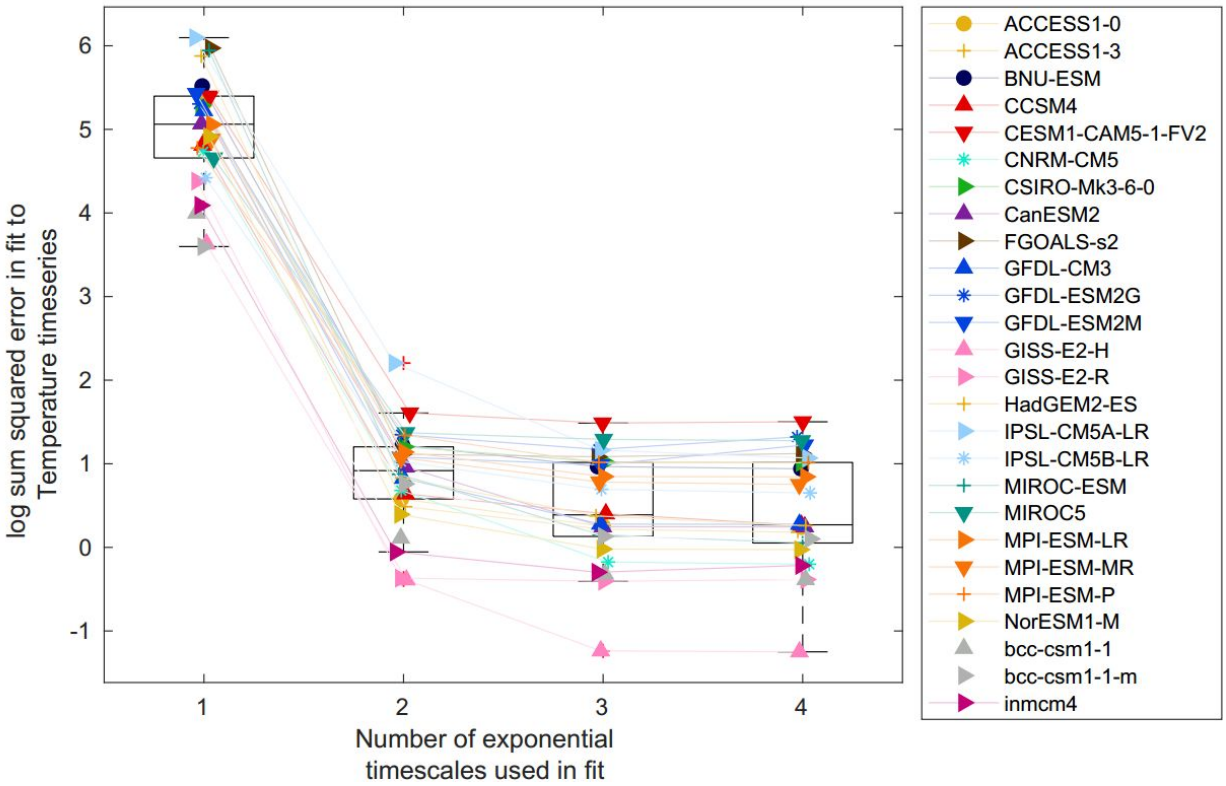
which is noted in the text at the end of the results section, but the improvement over the PIControl average is notable.

Some Minor

1) A common question concerning studies utilizing such simplified models is the sen- sitivity to the particular choice of the model setup. In this respect, the author may like to comment on the sensitivity of the results with respect to the particular choice of the number of timescales (n=2 in B1 & B2). How different would be the results for n=1 (or n=3)?

Repeating the entire analysis with a different model dimensionality is beyond scope, but during development, I experimented with different timescales dimensions - 1 timescale can be trivially dismissed as unable to represent the temporal evolution of the models in response to 4xCO2 forcing. Beyond two timescales, only slight improvement is seen in the fitting error - so two timescales was chosen for this study to be (a) consistent with existing literature (i.e. within the framework of FAIR, which is in common usage), (b) lower dimensional so easier to interpret in terms of slow/deep ocean and fast/shallow ocean response and (c) applicable to CMIP models in.
Fundamentally - only some models show a slightly improved fit with an extra allowed timescale (see GISS-H, for example on the below plot). Other models are adequately described with 2, and adding a 3rd results in a degenerate fit. Thus - a cross ensemble analysis of the additioanl degree of freedom is not clearly defined. Other studies have arrived at the same conclusion (see Proistosescu and Huybers http://doi.org/10.1126/sciadv.1602821, Smith 2018 http://dx.doi.org/10.5194/gmd-11-2273-2018 , Geoffroy 2012 http://doi.org/10.1175/JCLI-D-12-00196.1).

Ultimately, for this study, the aim is to reproduce the basic features of CMIP ensemble diversity in response to different types of forcing with the minimum possible complexity of model - and I felt that this was both possible and easier to explain with the two timescale model. Clearly, the real world could have the capacity to respond to forcing on a range of timescales, but two timescales adequately describe the response to forcing on the century timescale in the CMIP ensemble.



I have added a paragraph in the conclusions on how the structural assumption of two timescales may impact results. Primarily - I think the current method is support the primary conclusions that ECS and TCR are insufficient to constrain some future warming trends (i.e. ECS or TCR are insufficient to describe RCP2.6 warming), and that non-equilibration is a problem for measurement. But I do agree that the structural assumption of 2 timescales might impact the constraint, for example, of ECS from historical temperatures - and I've added this caveat.

2) The author introduces a new metric (A140) as an alternative. It would be useful if the author could illustrate the behaviour of A140 (in contrast to EFFCS) for a CMIP data set.

A140 is included in new Figure 5, and estimated values are included in new Table 2.

3) In the abstract, the author quantifies the relative errors for T140 and EFFCS in the simple model framework. As these numbers may certainly not be the same for CMIP model, the may not be part of the abstract.

I have removed these quantitative results from the abstract accordingly

4) f_r appears twice in Table B1

Response to Reviewer 2

Overall, I think that the manuscript is well written, the issue has a great scientific rele- vance, and the arguments here shown provide significant advancement to the discus- sion on the topic. Thus, I appreciate that the author addresses them critically, empha- sizing that their adoption is conditioned to the problem that one needs to focus on. This is in line with previous works having evidenced the limitations of these metrics for the study of the climate response, especially from a modelling perspective.

Many thanks for the positive evaluation and careful reading.

I am a bit skeptical about the effectiveness of the impulse-response model, given that it is a purely linear context. The addition of the noise+drift, though, is convincing in explaining part of the discrepancy between the simple model and CMIP5 outputs. The arguments about the applicability of the metrics are thus promising also in a "real- world" context (using the notation adopted by the author), although with some caveats. For this reason, I think it is important that the author puts more emphasis on the nature of the impulse-response model, in the framework of linear response theory (LRT) and Hasselmann-type response (see my specific comments), and evidences its limits.

These points are well taken - and thanks to the reviewer for the additional literary context, which I've endeavoured to include. I've tried to put the two timescale model in appropriate context - the primary defense for this application being that it is already sufficiently complex to show that TCR and ECS do not constrain future warming under strong mitigation, and that non-equilibration is a potential issue for TCR estimation. I believe that these points, which are statements of lack of confidence, are robust to the consideration of a wider set of models with additional response timescales.

I do however agree that the 2-timescale structural assumption is strong - and any constrained distribution (of future warming, EffCS or TCR) need to be considered in the context of these caveats. For this reason, I do not highlight the actual constrained ranges here - and I have added an additional paragraph to the conclusions to explain this.

I think some improvements can be made in terms of how the methodology and results are described. It would be useful to have the "Methods" section in the main part of the manuscript, instead of as an appendix.

I have restructured the document to have the methods in line.

The notation is not always consistent across the methods.

I've worked to reformat the methods extensively following the comments by both reviewers

The MCMC procedure should be explicitly described, not only by mentioning the original reference.

I've included an extended description of the algorithm and the reasons for using it.

Specific comments:

- ll. 90-92: I do not have clear how the normalized regression coefficients shown in Figure 1f support this argument. Can the author better clarify it?

I've deleted this paragraph - as I think the point is overly subtle.

- ll. 94-96: I do not understand this sentence for a few reasons. Firstly, is the author referring to any specific forcing, when he says that the rate of change in the forcing is approximately constant? In the case of the mitigation scenario (RCP2.6), this is obviously not the case.

Apologies - this paragraph was talking explicitly about RCP8.5, in which total radiative forcing increases broadly linearly throughout the 21st century. I've rewritten this section.

Secondly, I do not have clear in mind what the author means by "saturation" of the fast feedback response, and if this refers to the whole period 2000 to 2100 or to the end of the period.

Section now deleted.

- l. 132: the author suggests that the contributions of the two factors are separately addressed in the following, but, in the end, only the overall effect of the bias is taken into account in the following. -

Thanks - corrected. I now come back to the unknown baseline factor in the CMIP detrending exercise at the end of the results section.

ll. 161-162: the author seems to imply that "real-world applications" are prone to the existence of drifts. But this is rather a model issue, as the unforced "real-world" climate system should not have any drift.

Corrected.

- l. 166: the author did not specify anywhere else in the text what is the length of the abrupt 4xCO2 simulation. As a consequence, "end" of the simulation does not seem to have a specific meaning.

Replaced by “years 121-140”

- Appendix B: according to ESD standards, I think that it would be more appropriate if the Methods section are moved in the main text after the Introduction.

Done - methods are now inline in the text

Moreover, a description of the data that have been used is lacking, especially for what concerns the observational-based datasets used for model optimization.

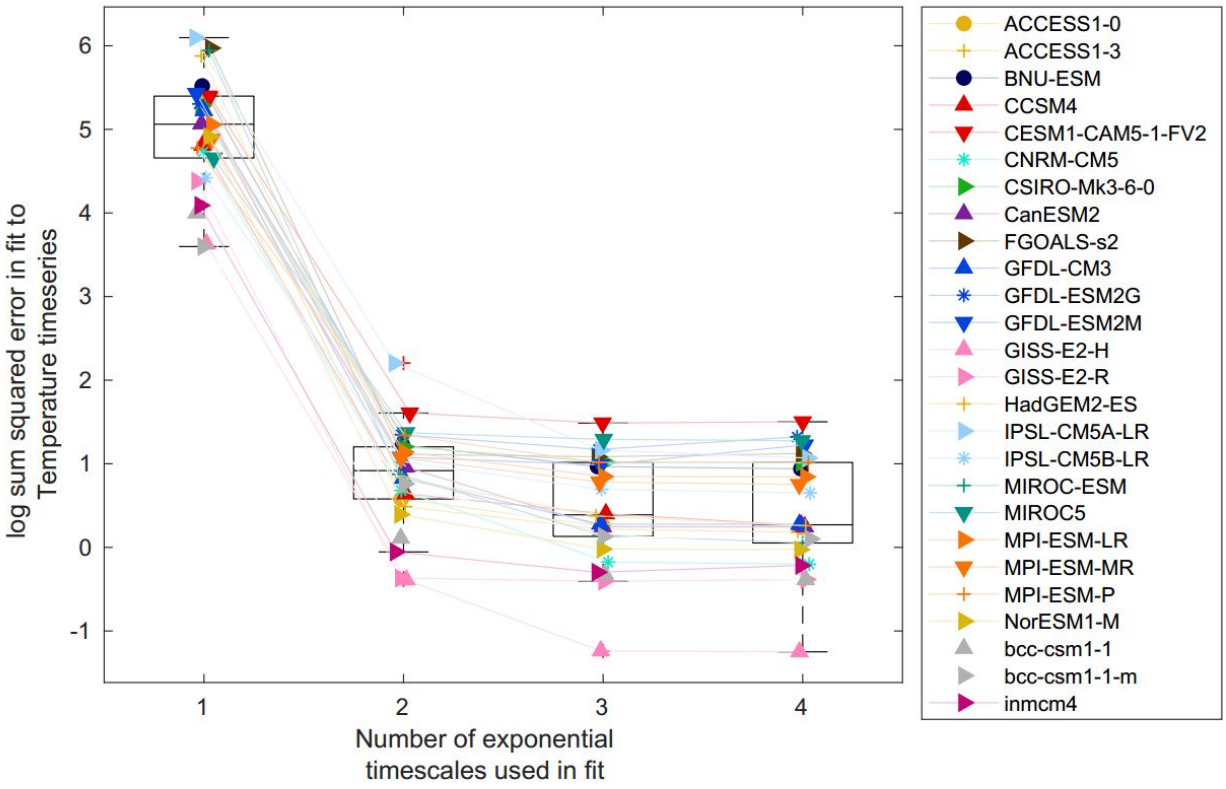
All relevant citations are now included.

- Eqs. B1-B2: the impulse response model here adopted requires using only two timescales. Is it sufficient to describe the response? The FAIR impulse-response model here mentioned includes a set of four simple feedback equations (cfr. Hassel- mann et al. 1993) differing on the magnitude of the feedback parameter (i.e. on the timescale of the response). What happens if one includes more than the two timescales considered in this analysis, given that similar strategies applied to geoengineering sce- narios have used, for instance, three exponentials (cfr. Aengenheyster et al. 2018)?

I fully agree that 2 timescales is a structural assumption, and that additional timescales of response would be likely required for longer periods of response. During development, I experimented with different timescales dimensions - 1 timescale can be trivially dismissed as unable to represent the temporal evolution of the models in response to 4xCO2 forcing. Beyond two timescales, only slight improvement is seen in the fitting error - so two timescales was chosen for this study to be (a) consistent with existing literature (i.e. within the framework of FAIR, which is in common usage), (b) lower dimensional so easier to interpret in terms of slow/deep ocean and fast/shallow ocean response and (c) sufficient for demonstrating the main point that drift and noise impact TCR more than ECS.

For 140 abrupt-4xCO2 response, only some models show an improved fit with an extra allowed timescale (see GISS-H, for example on the below plot), and even then it’s a slight improvement. Most models are adequately described with 2, and adding a 3rd results in a degenerate fit. Other studies have arrived at the same conclusion for summarizing responses on the century timescale (see Proistosescu and Huybers http://doi.org/10.1126/sciadv.1602821, Smith 2018 http://dx.doi.org/10.5194/gmd-11-2273-2018 , Geoffroy 2012 http://doi.org/10.1175/JCLI-D-12-00196.1).

Ultimately, for this study, the aim is to reproduce the basic features of CMIP ensemble diversity in response to different types of forcing with the minimum possible complexity of model - and I felt that this was both possible and easier to explain with the two timescale model. Clearly, the real world could have the capacity to respond to forcing on a range of timescales, but two timescales adequately describe the response to forcing on the century timescale in the CMIP ensemble.



This is particularly relevant, as the impulse-response model can always be expressed as an infinite sum of exponential behaviors, differing in their timescale, but the re- sponse of the real system rarely has the shape of a discrete number of exponential behaviors combined with each other (e.g. Ragone et al. 2016: Lembo et al. 2019). Also, the adoption of the fast-slow scale implies a separation of scale, that is here in- ferred "a posterior" through heuristic arguments. Nevertheless, there is no reason, in principle, to assume that a scale separation exists, and this problem traces back to the very foundations of the theory about climate response and forced-free fluctuations dichotomy (Lorenz 1979). One way to deal with that would be to evaluate the memory term (cfr. Ghil and Lucarini 2019). I understand that this might go beyond the scope of this work, but I wonder if the author might comment on that in the manuscript.

This point is well taken - though to redesign the model as an infinite sum would create a challenge in terms of a low-dimensional parametric definition which could be used in MCMC. However - I recognise that the discrete response assumption is a strong one, and I've added a paragraph in the discussion to outline this caveat in the interpretation of the results.

> *" These conclusions are derived from the consideration of a relatively simple two-timescale pulse response model which is sufficient to show that constraining certain types of sensitivity metric is insufficient to*

constrain future projections, and that non-equilibration may confound measurement, however, the constrained distributions for the metrics are subject to the structural assumptions of the model used. The real world may have more than two response timescales (Aengenheyster 2018), or may be better described as a continuous sum (Ragone 2016, Lembo 2019). Further work should identify how such complexity impacts uncertainty in relevant climate metrics."

- Eq. B3: according to the convolution properties, this operation is by all means equiv- alent to the application of the Ruelle Response Theory (RRT) (Ruelle 1998a; Ruelle 1998b) when a hypothetical impulse perturbation is applied, allowing for a particularly simple derivation of the linear Green function (cfr. Hasselmann et al. 1993). This has found several applications in the context of climate prediction (cfr. Ragone et al. 2016; Lucarini et al. 2017; Ghil and Lucarini, 2019 for a review), not only constraining to the temperature response, but also to a wide range of climatic variables (e.g. Helwegen et al. 2019; Lembo et al. 2019). These arguments provide a rigorous mathematical framework to the experimental protocol here described.

Thanks for these. I've included the references when introducing the model.

- Sect. B1.1: I believe that a complete description of the model is here lacking and should be included. Referring to the model settings, in particular, it is not clear to me how the ensemble is generated and how many members are taken into account.

This section has been significantly expanded, and now includes a perfect model demonstration fitting the model to CMIP members.

- Table B1: is it possible to have a range for rn as well? Also, where does the fr parameter enter the mode? This goes back to my minor comment about consistent notation.

This is now clarified in the text. r_1 is varied (r_2 is (1-r_1) due to the initial boundary condition). F_r is now explicitly detailed in Eqn. 5.

- l. 226: I think that it is important to notice here that in the forcing scenario 1pctCO2 the CO2 concentration reaches doubling after 70 years, as I presume that this motivates the choice of the 61-80 and 131-150 20-years averages.

Now noted explicitly, thanks.

- Figure A1: the caption does not contain an explanation of the panel b content. Partic- ularly, the author might want to explain the meaning of the red shading, and the range encompassed by the dotted lines.

Expanded.

- Figure A2: the author does not explain why the choice of a single member from each CMIP model ensemble is reasonable in this context.

I've now noted that the plot is subject to internal variability, but this is a central point which is being made. I am not trying to assess what is the most robust sensitivity metric given a situation where there is noise and potentially drift in the simulations. To have a subset of models with large ensemble averages (and others without) would confuse that assessment.

- Figure A3: it appears that the distributions of fast-scale parameters are much more similar to a Gaussian distribution, compared to the slow-scale parameters. I am surprised that the author does not refer to that explicitly and comments on it. Could it be an evidence that the scale separation that is a priori assumed for parameter model optimization is such that the fast-scale system approaches a stochastic process, in the context of the response of the system to the impulse forcing? This would be certainly reasonable, in an "Hawkins and Sutton, 2009 context" (signal-to-noise ratio approach), but the author might want to justify it in a more rigorous way.

I've expanded this discussion a little - though I'm not sure that we can infer any dynamical separation of timescales from the differences in distribution. My interpretation is that the fast timescales are simply more strongly constrained by the observations, whereas there are solutions with a wide range of slow timescale responses.

Minor comments:

- l. 19: in this sentence there is a repetition ("range" and "ranging"). Consider rearrang- ing the sentence;

Thanks, corrected.

- l. 31-32; I found this part of the sentence a bit difficult to read. A suggestion might be to replace it with "a complication has arisen due to the fact that EffCS seems to be better correlated than TCR with 21st Century warming from present day levels under a business-as-usual scenario."

Thanks - corrected as suggested.

- l. 37: replace "have" with "of".

Thanks, corrected.
- l. 60: remove "to"

Sentence removed
- l. 66: it is not clear whether the author refers here to the Appendix A, Appendix B or both.

Methods are now inline with the paper.

- l. 69: replace “and” with “to”.

Thanks, done

- l. 91: either a sentence breaking is needed here (after the brackets), or “suggest” has to be replaced by “suggesting”.

Sentence removed.
- l. 125: Replace “of CMIP5” with “for CMIP5”.

done
- l. 138: if the “Methods” section is in the appendix, they have to be referred to more appropriately as “Appendix (B)”.

Methods now inline.
- l. 147: replace “the both” with “both”.

Thanks, corrected
- l. 151: replace “Supplemental” with “Supplementary”.

done
- l. 165: replace “an” with “that a”.

done
-Eq. B3: I noticed a potential mismatch in the notation, compared to eq. B1. The author may consider adopting the same notation for the temperature evolution in both equations.

This is now consistent throughout.

- Figure A1: replace “senstivity” with “sensitivity” in the caption.

List of Changes in Manuscript

- Methodology moved to main section and expanded to provide more complete description of the model used
- Expanded discussion of Markov Chain optimization
- Added validation of technique using historical CMIP simulations and RCP2.6/8.5 projections (results shown in Figure S4)
- Tested effect of residual drift on TCR estimation (latter part of results section and Figure 5, and Figure S5)
- Added discussion of number of exponential modes in model
- Added new Table 2 illustrating sensitivity metrics for CMIP models
- Expanded literature review of pulse-response formulation & linear response theory

Relating Climate Sensitivity Indices to projection uncertainty

Benjamin Sanderson[1,2]

[1]CERFACS, Toulouse, France
[2]NCAR, Boulder CO, USA

Correspondence: Benjamin Sanderson (sanderson@cerfacs.fr)

Abstract. Can we summarize uncertainties in global response to greenhouse gas forcing with a single number? Here we assess the degree to which traditional metrics are related to future warming indices using an ensemble of simple climate models together with results from CMIP5 and CMIP6. We consider Effective Climate Sensitivity (EffCS), Transient Climate Response at CO_2 quadrupling (T140) and a proposed simple metric of temperature change 140 years after a quadrupling of carbon dioxide (A140). In a perfectly equilibrated model, future temperatures under RCP(Representative Concentration Pathway)8.5 are almost perfectly described by T140, whereas in a mitigation scenario such as RCP2.6, both ECS and T140 are found to be poor predictors of 21st century warming, and future temperatures are better correlated with A140. However, we show that T140 and EffCS calculated in full CMIP simulations are subject to errors arising from control model drift and internal variability. Simulating these factors in the simple model leads ~~to 30% relative~~ greater error in the measured value of T140 ~~, but only a 10% error~~ than for EffCS. As such, if starting from a non-equilibrated state, measured values of Effective Climate Sensitivity can be better correlated with true TCR than measured values of TCR itself. We propose that this could be an explanatory factor in the previously noted surprising result that EffCS is a better predictor than TCR of future transient warming under RCP8.5.

Introduction

Summarizing the response of the Earth System to anthropogenic forcers with ~~a~~ metrics has long been practised as a way to illustrate uncertainty in Earth system response to greenhouse gases. For example, the concept of the Equilibrium Climate Sensitivity (ECS), the equilibrium global mean temperature increase which would be observed in response to a doubling of atmospheric carbon dioxide concentrations (Hansen et al., 1984) has existed for over 50 years (Charney et al., 1979) and significant amount of literature has been devoted to constraining its value (Knutti et al., 2017).

The Earth ~~System~~ system responds to a step-change in forcing on ~~a range of~~ timescales ranging from days to millennia (Knutti and Rugenstein, 2015), so an ~~'~~'Effective Climate Sensitivity' (EffCS hereon) is often used as a proxy for decadal to centennial feedbacks. EffCS is generally calculated in a coupled atmosphere-ocean model from the output of the ~~'~~'abrupt4xCO2' simulation, a standard experiment in which CO2 concentrations are quadrupled instantaneously from pre-industrial levels and the model is allowed to evolve (Gregory et al., 2004).

EffCS is calculated by assuming that a model is associated with a single feedback parameter (i.e. a rate of change of top of atmosphere radiative flux per unit surface temperature increase), allowing the equilibrium temperature response to a step change forcing to be predicted by linear extrapolation (we refer to this approach henceforth as the Constant Feedback (CF)

approximation, with EffCS referring to the estimate of ECS made using this approach). Another metric, the Transient Climate Response at CO_2 doubling (TCR) or quadrupling (T140) is calculated from an ~~'~~'1pctCO2' idealized experiment in which CO_2 concentrations are increased by 1 percent each year, starting from a pre-industrial state, resulting in linearly increasing forcing.

Although it was generally assumed that TCR would be a better predictor of transient warming under a high emissions scenario such as RCP8.5 (Riahi et al., 2011), a complication has arisen due to the fact that EffCS seems to be better correlated than TCR with 21st ~~century~~ Century warming from present day levels under a business-as-usual scenario ~~than TCR in the CMIP5 ensemble~~ (Grose et al., 2018). The reason for this is not yet well understood given the radiative pathway in RCP8.5 leading up to 2100 is relatively similar to that of the 1 percent annual increase experiment used to measure T140. ~~Another pressing concern is that~~ Furthermore, neither EffCS nor TCR is well correlated ~~to~~ with end of century temperatures in a mitigation scenario (Grose et al., 2018) such as RCP2.6 (Van Vuuren et al., 2011), which calls in to question the relevance of such summary metrics in the discussion of mitigation adaptions.

Similarly, a number ~~have~~ of studies have shown that the EffCS approximation does not well describe the true equilibrium behaviour of most models(Knutti et al., 2017). When GCM abrupt-4xCO2 simulations are continued for thousands of years, many are found to deviate significantly from the linear trend-line one would fit to a 150 year simulation (Andrews et al., 2015; Knutti et al., 2017; Senior and Mitchell, 2000; Rugenstein et al., 2016).

The conceptual models representing the evolving feedbacks as a function of timescales vary slightly between studies - either modulating the efficacy of deep ocean heat uptake (Geoffroy et al., 2013; Winton et al., 2010; Held et al., 2010) or by representing the climate system as sum of warming patterns which emerge on different adjustment timescales (Armour et al., 2013; Rugenstein et al., 2016), each associated with their own feedback parameter. However, the analytical set of solutions for the temperature response to a step change in forcing is the same in either case - a superposition of decaying exponential modes with different timescales varying between a few years and a few centuries (Proistosescu and Huybers, 2017). It has been shown that the implications of these additional degrees of freedom, and ambiguity over contributions from different timescales of response might imply that ~~equilibrium climate sensitivity~~ EffCS may not be strongly constrained by temperature change over the last century (Proistosescu and Huybers, 2017; Andrews et al., 2018), and that the Long Term Equilibrium (LTE) sensitivity may be greater than that implied by estimates which use the CF framework (Otto et al., 2013; Lewis, 2013).

This state of understanding leads to a number of emerging critical questions which we discuss in this paper - can we explain the non-intuitive result that EffCS is a better predictor then T140 of end-of-century temperatures under RCP8.5~~, which~~ ? Which summary metrics of global sensitivity to greenhouse gas forcing are most useful for effective policy decisions~~, and~~ ? Finally, do the implicit structural assumptions underpinning the very existence or applicability of these metrics to the real world cause us to mis-categorize and potentially underestimate future warming risk?

1 A simple model example

We begin by considering an idealized ensemble of climate model simulations. We use a two timescale thermal response model, conceptually representing the deep ocean (with a response timescale of a century or more) and shallow ocean response

timescales (with a response timescale of 10 to 50 years). Such a model, although simple, is capable of resolving evolving feedback amplitudes and can emulate the climatological responses of complex Earth System Models ~~to on a range of timescales (Proistosescu and Huybers, 2017; Geoffroy et al., 2013)~~on two timescales. Such a model makes a reasonably strong structural assumption that the Earth can be modelled as a discrete sum of linear decaying exponential responses to forcing, but this model has been found to well describe GCM evolution on a century timescale (Proistosescu and Huybers, 2017; Geoffroy et al., 2013) and is sufficiently complex to illustrate the limitations of defining system sensitivity through TCR or EffCS.

The two-timescale impulse response model follows the thermal feedback-timescale implementation from the FAIR simple climate model(Smith et al., 2018; Millar et al., 2017), which follows Hasselmann et al. (1993):

$$\frac{dT_n}{dt} = \frac{q_n F - T_n}{d_n}; T = \sum_n T_n; n = 1, 2, \tag{1}$$

where T_n is global mean temperature and for each timescale n. T_n is the component of warming associated with that timescale, q_n is the feedback parameter and d_n is the response timescale. We consider the heat flux into the shallow and deep ocean to be functions of the same timescale:

$$R_n = r_n(F - T_n/q_n); R = \sum_n R_n; \sum_n r_n = 1; n = 1, 2 \tag{2}$$

where r_n is an efficacy factor for heat absorbed by the deep ($n = 1$) or shallow ($n = 2$) ocean, which sum to unity given the boundary condition that $R(0) = F(0) = F_{4xCO2}$ at $t = 0$ (allowing just one degree of freedom r_1 - the fraction of heat which is allocated to deep ocean storage).

The particular solutions for temperature and radiation response to a step change in forcing F_{4xCO2} at time $t = 0$ can be expressed as a sum of exponential decay functions:

$$T_p(t) = F_{4xCO2} \sum_{n=1}^{2} q_n (1 - exp(-t/d_n)) \tag{3}$$

$$R_p(t) = F_{4xCO2} \sum_{n=1}^{2} r_n (exp(-t/d_n)), \tag{4}$$

where $T_P(t)$ is the annual global mean temperature and $R_p(t)$ is the net top-of atmosphere radiative imbalance at time t, and F_{4xCO2} is the instantaneous global mean radiative forcing associated with a quadrupling of CO_2, taken here to be $3.7 W m^{-2}$(Myhre et al., 2013).

To efficiently describe the response of the system to a generic forcing, this study employs a linear Green's function which describes the forcing by convolution with an impulse response Ruelle (1998) (in this case, the step change in CO2 forcing).

This approach can be applied to global climate dynamics Ragone et al. (2016); Lucarini et al. (2017), and its computational efficiency allows Markov-Chain Monte Carlo parameter estimation for the physical parameters.

We define a historical forcing timeseries as a function of CO_2 concentrations $C(t)$ and a non-CO2 forcing timeseries $F_{nonCO2}(t)$ (both taken from (Meinshausen et al., 2011)):

$$F(t) = \frac{F_{4xCO2}}{ln(4)} ln\left(\frac{C(t)}{C_0}\right) + f_r F_{aer} + F_{other}, \tag{5}$$

where f_r is a free parameter to allow scaling of aerosol forcing (conceptually allowing for forcing uncertainty in the historical timeseries), and $F_{otherAnt}$ is all other anthropogenic and and natural forcers (summed from (Meinshausen et al., 2011)). The thermal response is calculated by expressing the numerical time derivative of the forcing timeseries $F(t)$ where the change in forcing in a given time-step in a given year $\Delta F(t')$ is $[F(t') - F(t'-1)]$. The forcing timeseries can thus be expressed a series of step functions, and T_p from equation 12 can be used to calculate the integrated thermal response.

$$T(t) = \sum_{t'=0}^{t} \Delta F(t') \sum_{n=1}^{2} q_n \left(1 - exp\left(\frac{-(t-t')}{d_n}\right)\right), \tag{6}$$

Heat fluxes into the deep ($D(t)$) and shallow ($H(t)$) ocean components are represented by numerical integration of the slow (n=1) and fast (n=2) pulse response components of $R_p(t)$ in Equation 4:

$$D(t) = r_1 \sum_{t'=0}^{t} \Delta F(t') exp\left(\frac{-(t-t')}{d_1}\right), \tag{7}$$

$$H(t) = (1 - r_1) \sum_{t'=0}^{t} \Delta F(t') exp\left(\frac{-(t-t')}{d_2}\right), \tag{8}$$

1.0.1 Model Optimization

The model input time-series for calibration are observed CO_2 concentrations, along with radiative estimates from Meinshausen et al. (2011) of non-CO2 forcing agents. We optimize the thermal model parameters for 2 timescales and the non-CO2 forcing factor (see Table 1).

A Markov-Chain Monte-Carlo (MCMC) optimization procedure produces an ensemble of parameter configurations such that the density of the simulations in parameter space reflects the likelihood as reflected in a cost function (as represented by a number of pre-defined likelihood metrics). MCMC algorithms employ a random walk in parameter space which ultimately seeks to produce a representative sample of the distribution.

The classical approach to this random walk is the Metropolis Hastings algorithmMacKay and Mac Kay (2003), which iteratively moves a set of 'walkers' or sample points throughout the parameter space. This approach, however is computationally inefficient requires the specification of the transition distribution with a large number of degrees of freedom. Here, we follow

the (Goodman and Weare, 2010) MCMC implementation which updates a walker position using a vector defined stochastically from the remaining ensemble of walkers. This approach has fewer degrees of freedom and is a well-tested approach for multidimensional optimization problems Foreman-Mackey et al. (2013). We use flat initial parameter distributions as shown in Table 1, 200 walkers and 50,000 iterations for each optimization.

Cost functions are computed for global mean temperature, shallow and deep ocean content:

$$E_T = \sum_t \left(\frac{(T(t) - T_{obs}(t))}{\sqrt{2}\sigma_T} \right)^2 \tag{9}$$

$$E_H = \sum_t \left(\frac{(H(t) - H_{obs}(t))}{\sqrt{2}\sigma_H} \right)^2, \tag{10}$$

$$E_D = \sum_t \left(\frac{(D(t) - D_{obs}(t))}{\sqrt{2}\sigma_D} \right)^2, \tag{11}$$

where T_{obs} are HadCRUT 4.6 ensemble median global mean temperature anomalies Morice et al. (2012) relative to a 1850-1900 baseline and σ_T is defined as the standard deviation of HadCRUT 1850-1900 values. Shallow and Deep Ocean heat fluxes are taken as the 0-300m and 300m+ heat content derivatives respectively in (Zanna et al., 2019), with σ_H and σ_D taken as 1850-1900 standard deviations from the same dataset.

Flat priors are used for all parameters, with an additional prior on true equilibrium climate sensitivity using the likely value and upper bound on Equilibrium Climate Sensitivity from Goodman and Weare (2010) to specify the median and 90th percentile of a gamma distribution for equilibrium sensitivity (i.e. warming as $t \rightarrow \infty$).

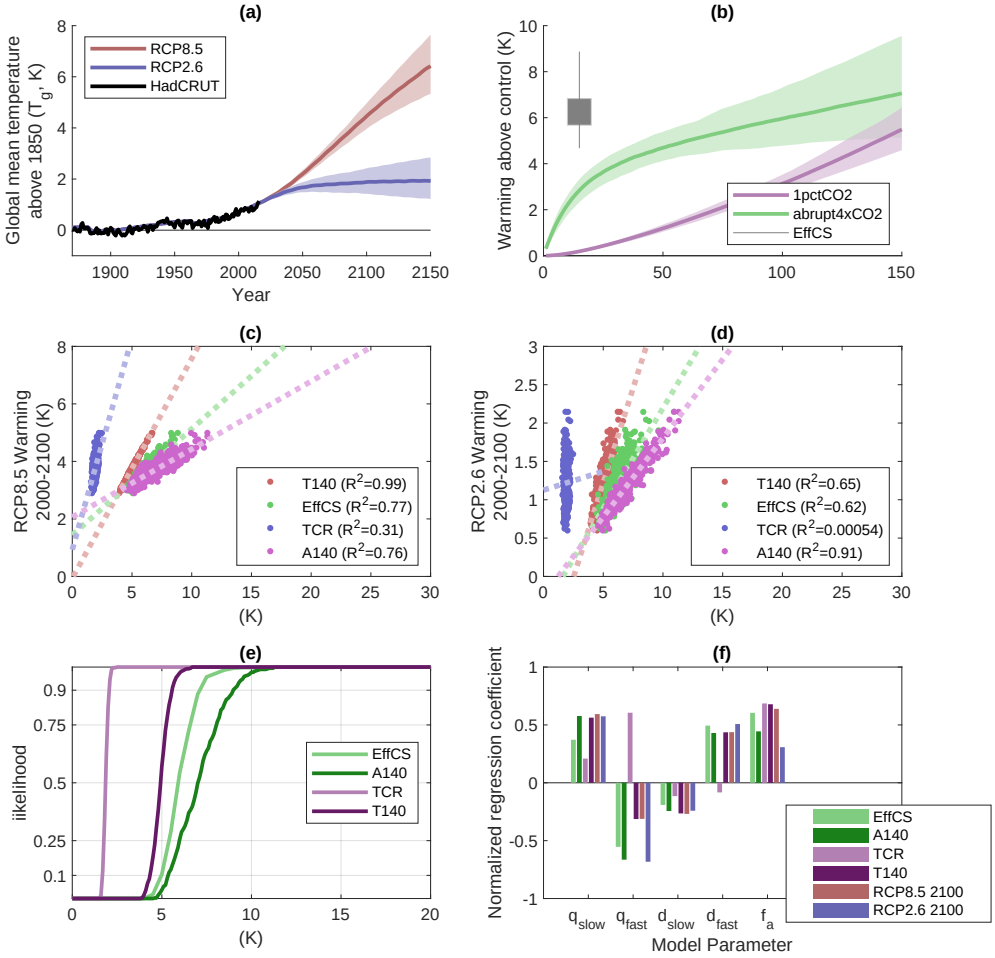
We demonstrate that this technique is able to capture the broad uncertainty associated with future projections of CMIP models by using pre-2020 temperatures in RCP8.5 to calibrate the simple model outlined above (Figure S4). In most cases, the future projection for each scenario falls within the distribution arising from the MCMC ensemble fit, with some specific exceptions - FIO-ESM, FGoals-G2, CCSM4 (which share some common heritage) and the GISS models. As such, the observationally fitted MCMC ensemble explores broadly comparable uncertainty to that seen in the bulk of the CMIP ensemble, with the caveat that the ensemble tends to under-sample cases where there is little or no long term warming response to emissions.

The physical parameters of this simple model are constrained by historical carbon dioxide concentrations together with observed global mean temperatures from 1870 to present day (together with aggregate forcing estimates representing other anthropogenic emissions (Meinshausen et al., 2011), which are not the focus of this study). ~~A Markov-Chain Monte Carlo algorithm is used to produce a~~ The posterior parameter distribution for the model ~~which~~ can then be used to project the corresponding range of response in probabilistic projections of the future scenarios or in idealized experiments ~~(see additional material)~~ which simulate a range of self-consistent values for various climate sensitivity metrics.

~~The resulting ensemble produces model variants with Effective Climate Sensitivities~~

1.0.2 Idealized Simulations

Figure 1. An observationally constrained ensemble of simple models. (a) shows the global mean temperature both historically and under the RCP2.6 and RCP8.5 scenarios. Black lines show the HadCRUT data used in calibration, whereas shaded regions show the 10-90% range of scenario projections in the posterior simple model ensemble distribution. (b) shows the corresponding time-series posterior distributions for the abrupt4xCO2 and 1pctCO2 simulated experiments, with grey errorbars showing range of EffCS for CO_2 quadrupling (boxes and whiskers show 25-75th and 1-99th percentiles respectively) . (c/d) show relationships between different sensitivity indicators and 2000-2100 temperature changes under RCP8.5/RCP2.6 respectively (e) shows the posterior cumulative probability density functions for the 4 sensitivity variables considered and (f) shows the parameter regression coefficients relating the 5 normalized model input parameters to the 4 normalized sensitivity metrics.



Long name	Symbol	Min	Max
Thermal equilibration of deep ocean Sensitivity (KWm^{-2})	q_1	0	10*
Thermal adjustment of upper ocean Sensitivity (KWm^{-2})	q_2	0	10
Thermal equilibration of deep ocean timescale ($years$)	d_1	100	4000
Thermal adjustment of upper ocean timescale ($years$)	d_2	10	100
Fraction of forcing in deep ocean response	r_1	0.	1
Non-CO2 Forcing ratio	f_x	.7	1.3

Table 1. A table showing model parameter values and minimum and maximum values allowed in model optimization.

Effective Climate Sensitivity is measured by implementing a step-change abrupt CO_2 quadrupling, and following (Gregory et al., 2004) to assess the linear extrapolation of warming at the point of net top of atmosphere energetic balance. A140 is calculated as the average of year 131-150 of the abrupt4xCO2 simulation. TCR and T140 are calculated as the average of years 61-80 and 131-150 respectively of the 1pctCO2 simulation (during which the CO2 concentrations are doubled and quadrupled, respectively), where CO_2 concentrations are increased annually by 1pct resulting in a linear increase in climate forcing. RCP scenario temperature trajectories are calculated for each parameter set using concentration and forcing timeseries from (Meinshausen et al., 2011) from 1850 until 2300.

Resulting EffCS values (to a doubling of CO2) ~~ranging~~ range from 2.4 to 4.6K (5th and 95th percentiles), and values of TCR from 1.6 to 2.2K (Figure 1(b,e)). This results in a range of 21st century warming under two scenarios considered, RCP2.6(RCP8.5) 2100 warming ranges from 1.4 ~~and~~ to 2.4 K (3.8 to 5.1K) respectively (5th and 95th percentiles, see Figure 1(a)).

We then consider in the context of this observationally constrained ensemble of simple models, what idealized metrics of system response are most informative for describing 21st century warming. We consider four metrics: the EffCS, TCR/T140 (transient warming under an annual compounded 1 percent increase in CO_2 concentrations at time of CO_2 doubling/quadrupling, corresponding to years 70 and 140 of the simulation). We also introduce A140 as a possible metric for consideration, defined as the global mean warming above pre-emission levels in the abrupt4xCO2 simulation calculated 140 years after time of CO_2 quadrupling (here and throughout estimated as the mean from years 131-150). Figure 2 illustrates how ensemble spread would be impacted for a set of different scenarios if each of these metrics were constrained to lie within a narrow range (nominally the 45-55th percentile range of values present in the entire observationally constrained ensemble).

In the high emissions, RCP8.5 scenario (Riahi et al., 2011), 2000-2100 warming is nearly perfectly described ($R^2 = 0.99$) by T140, the transient climate response after 140 years in a 1 percent CO2 simulation (Figure 1(c) and Figure 2(k)). The corresponding response after only 70 years, TCR, is a much poorer predictor at $R^2 = 0.31$).

These results are physically intuitive. The climate forcing and rate of change of forcing in RCP8.5 at the end of the 21st century are of similar magnitude to those in year 140 of the 1 percent CO2 simulation, and so it is unsurprising that T140 is an efficient predictor for RCP8.5. TCR is a poor predictor in the simple model ensemble largely because TCR itself is already

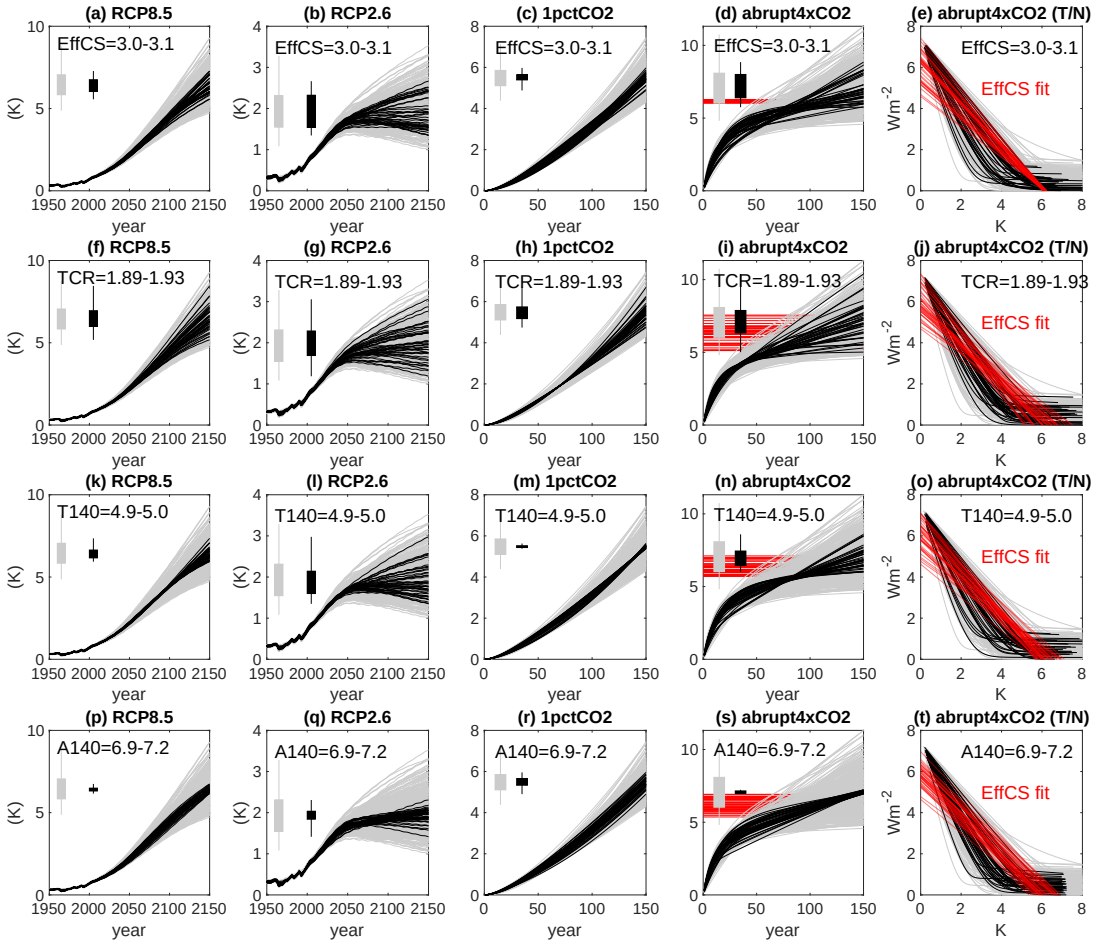


Figure 2. An illustration of how constraining different types of global sensitivity metric impact the idealized spread of global mean temperature evolution under different scenarios. Each row illustrates one constraint, Effective Climate Sensitivity to CO_2 doubling (EffCS), TCR (70 year, CO_2 doubling), T140 (140 year, CO_2 quadrupling) and A140. Lines in grey show the entire posterior distribution of models from Figure 1, while lines in black show the 45-55th percentiles of the distribution of the respective quantity. The first four columns show global mean temperature time-series of a scenario or idealized experiment - RCP8.5, RCP2.6, 1 percent ramping CO2, abrupt CO2 quadrupling (the 5th column shows energetic imbalance as a function of surface temperature in the abrupt4xCO2 experiment). Histograms show the resulting distribution of temperature in 2150 (RCP8.5/2.6) or year 140 (1pctCO2, abrupt4xCO2) for the complete distribution (grey) and 45-55th percentile range (black). Red lines show the distribution of values of effective climate sensitivity (4th column) and the trend lines used to compute it (5th column).

highly constrained by historical warming (Figure 1(e)), and thus the ensemble is effectively conditioned on a value of TCR and it has little additional explanatory value in explaining the ensemble variance in the RCP projections (Figure 2(f,g)).

EffCS and A140 are also well correlated with the RCP8.5 warming ($R^2 = 0.77$ and 0.76 respectively), but less so than T140. For the mitigation scenario RCP2.6, the most effective predictor of 2000-2100 warming is A140 ($R^2 = 0.91$). Both EffCS and T140 are weakly correlated ($R^2 = 0.62$ and 0.65 respectively), and TCR shows no significant correlation.

To help understand these relationships, we can perform a regression analysis of the metrics as a function of model ensemble parameters (Figure 1(f)) ~~suggests that variance in both RCP8.5 warming and T140 are strongly controlled by the slow climate feedback parameter.~~

~~In a pulse response formulation, the response of the global temperature to forcing can be understood as a sum of a fast- and slow-equilibrating responses to the *change* in forcing in each timestep. Because the rate of change of forcing remains broadly constant from 2000 until 2100, the fast feedback response associated with the shallow ocean is already saturated and the linear growth in temperature in the transient regime is governed mainly by the slow response (the rate of warming of the deep ocean).~~

which suggests A140 and RCP2.6 warming from 2000 to 2100, ~~however is broadly defined~~ are controlled by the *difference* between the slow and fast components of sensitivity. We can understand this in the context of the way the model is constrained by historical temperatures.

There is a weak trade-off between fast and slow components of climate sensitivity in the posterior parameter distribution of the ensemble (see ~~additional~~ supplementary figure S3), which broadly determines the fraction of equilibrium warming associated with current forcing levels that has already been experienced.

If a greater fraction of today's observed warming is explained with the faster component of model response, there is less unrealized warming in a mitigation scenario later in the century. ~~A140 shows similar parameter correlations and thus is well correlated to RCP2.6 end of century temperatures. Although the Effective Climate Sensitivity is a moderately good predictor of warming in both RCP2.6 and RCP8.5 in the simple model, A140 is more effective for predicting~~ This causes large uncertainties in RCP2.6 ~~temperatures due to its greater sensitivity to the~~ evolution, even if EffCS, TCR or T140 are known (Figure 2b,g,l).

The constrained distribution for fast-timescale sensitivity is near-Gaussian, and non-zero in all ensemble members, whereas slow-timescale sensitivity is more weakly constrained by the observations ranging from near-zero to large (20K/Wm^{-2}) long term equilibrium responses. The slow feedback component strongly controls A140 and RCP2.6 warming (Figure 1(d,f)~~)~~ , Figure 2q).

RCP8.5 warming and T140, however are associated with a near-linear increase in forcing throughout the simulation which results in a near-linear temperature increase. The relative fraction of warming associated with fast- and slow-timescale feedbacks remains constant over time, and thus warming to date (effectively fixing TCR, subject to aerosol forcing uncertainty) better constrains relative error in future response in a non-mitigation scenario (Figure 2f).

2 Considering the multi-model ensemble

But how do the findings in the simple model framework reconcile with findings in the CMIP5 and CMIP6 multi-model ensembles? Firstly, it is plausible that there is some commonality in the lack of skill of TCR (the transient response after 70 years) in our simple model ensemble and in the CMIP ensembles. In our simple model case, the ensemble members were explicitly calibrated to reproduce the 20th and early 21st century warming - which is a very strong constraint on the value of TCR in this idealized setup.

Earth System Model calibration is conducted in a much larger parameter space by groups with a wide range of objectives which complicate interpretation (Mauritsen et al., 2012; Sanderson and Knutti, 2012), but simulations are generally only published using models which are able to adequately describe the 20th century and thus might be subject to a similar effective constraint on TCR which renders the metric ineffective for describing variance in the future evolution of the model. But there remains a direct contradiction for T140, where the simple model suggests T140 should be a better predictor than EffCS for non-mitigation warming in the 21st century whereas the opposite was found in the CMIP correlations (see ~~additional~~ Supplementary material, Figure S2 and (Grose et al., 2018)).

To understand this, we need to consider how the properties of the simple model ensemble differ from the CMIP archive. Although the thermal response of the simple model is broadly able to represent the climatological response of CMIP models to step forcing and transient forcing in CO_2 over a century timescale ((Geoffroy et al., 2013; Proistosescu and Huybers, 2017)), it contains no internal climate variability and all experiments in ~~section~~ Section 1 are conducted from an idealized, perfectly spun up state.

Both of these assumptions are not true ~~of~~ for CMIP5 or CMIP6. Measurement of EffCS and TCR are complicated by internal variability (Knutti and Rugenstein, 2015), and many models still exhibit some temperature drift in the control simulation from which the ~~'~~'1pctCO2' simulations and ~~'~~'abrupt4xCO2' simulations are branched (Figure 3). This creates uncertainty from two sources - firstly, it is not always apparent at what point during the control simulations the 1pctCO2 simulation has been branched, thus there is uncertainty in how the anomaly should be measured. Secondly, there is the potential for an unknown contribution of control drift to be erroneously included in the temperature evolution of the 1pctCO2 and abrupt4xCO2 simulations.

To assess the contribution of ~~these two factors in metrics of climate sensitivity~~ control drift bias in sensitivity metrics, we implement idealized representations of ~~these sources of measurement error~~ non-equilibration into our simple model from Section 1. We then create an idealized distribution of drift similar to that seen in the CMIP ensembles in the simple model ensemble by initializing the model 500 years before the experiment begins, defining an effective 'baseline' period from which anomalies are measured to be the average temperature between years 400 and 500. Climate internal variability is represented by a 2nd order autoregressive model, which is fitted to each CMIP model in turn. The ensemble-mean autoregressive parameters are used to create artificial ~~'~~'noisy' simulations by linearly adding noise generated from the autoregressive model to the output of the simple model~~(see methods.~~

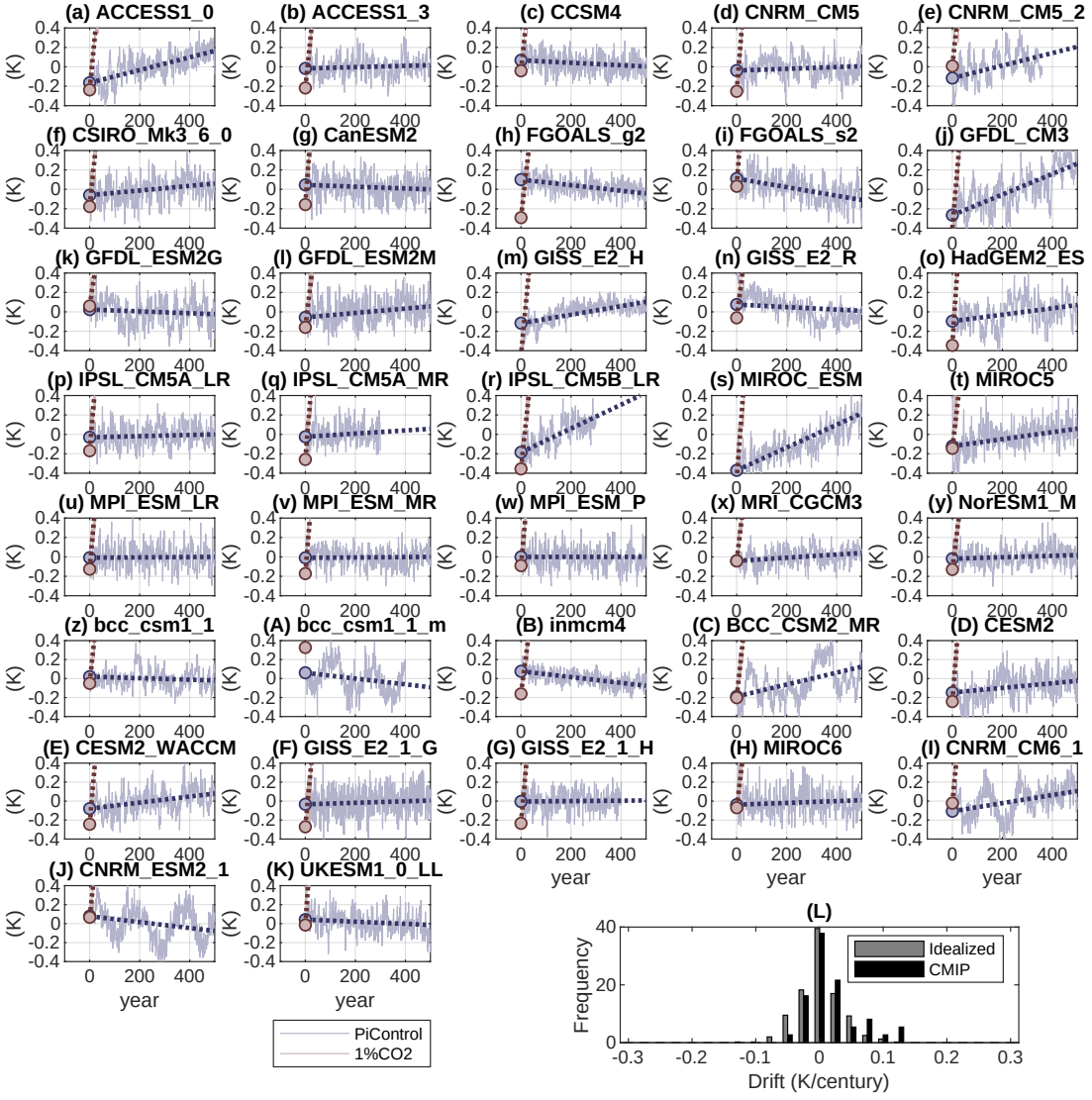


Figure 3. (a-K) Control simulation global mean temperatures from a selection of models in the CMIP5 and CMIP6 ensembles. Control simulations (blue) and initial years of 1pctCO2 simulations (pink) are plotted. Dotted lines show linear fit to the available timeseries. Blue and pink circles show the intersection of the linear temperature fit at the start of the simulation. (L) histogram showing the distribution of control model trend in CMIP (black) and in idealized ensemble of non-equilibrated simple models considered in Figure 4 (grey).

We consider the range of control drifts observed in the CMIP5 and CMIP6 ensembles (illustrated in Figure 3(L)) which range from -.3 to +.6K /century in the CMIP5 and CMIP6 models considered in this study. An idealized distribution of drift in the simple model ensemble is created by initializing the model 500 years before the abrupt4xCO2 or 1pctCO2 simulation with a non-zero, constant forcing drawn from a flat distribution ranging from -1 to $+1 W m^{-2}$, which results in a distribution of control drift of -.4K to +.4K per century (i.e. broadly comparable to the CMIP case). For each simulation we consider a baseline for temperature to be defined by the average global mean temperature in years 400-500.

To represent the first order effect of climate noise, we fit a 2nd order autoregressive model to the detrended global mean temperature timeseries in each available model in the CMIP5/6 ensemble. Taking CMIP mean parameters for the variance and autoregressive parameters, we generate noise for each realization of the simple model (though we note, in practise that the noise characteristics vary by CMIP model).

The results are illustrated in Figure 4(a), where the simple model ensemble is initialized in a non-equilibrium state with additive Gaussian noise. With these additional sources of error, both EffCS and A140 are not strongly impacted when measured in the noisy/unequilibrated model variants (Figure 4(b,c)), but the T140 measurement is strongly degraded (Figure 4(d)). Indeed, in this ensemble the biased measurements of EffCS or A140 are slightly better correlated with true T140 than the biased measurement of T140 itself. This provides a possible explanation for why T140 may be a poor predictor of RCP8.5 warming in CMIP.

In our simple framework, the reasons for the more accurate measurement of EffCS are primarily associated with the lack of equilibration. Simply adding noise from the autoregressive model has little effect on the accuracy of EffCS, T140 or A140 (where ~~the~~ both T140 and A140 are estimated using the average of years 131 to 150 in the simulation, see Table 2). ~~However, both A140 and EffCS are less sensitive to non-equilibrated initial states than T140. The former experiences the same variance due to the uncertain climate drift, but the absolute value of A140 tends to be larger than T140, thus there is less relative error in its estimation. The effect on the drift on EffCS is muted because the near-linear climate drift primarily biases the estimation of slow rather than fast feedbacks (see Supplemental Figure S1). Because EffCS is primarily a measure of fast-mode feedback strength (see Figure 1(f)), its value is less impacted if experiments are started from a non-equilibrium state.~~

Both A140 and EffCS are less sensitive to non-equilibrated initial states than T140. The former experiences the same variance due to the uncertain climate drift, but the absolute value of A140 tends to be larger than T140, thus there is less relative error in its estimation. The effect on the drift on EffCS is muted because the near-linear climate drift primarily biases the estimation of slow rather than fast feedbacks (see Supplementary Figure S1). Because EffCS is primarily a measure of fast-mode feedback strength (see Figure 1(f)), its value is less impacted if experiments are started from a non-equilibrium state.

There is some evidence that the lack of equilibration has an outsized effect on the estimation of TCR in the CMIP models. In Figure 5, we attempt to unbias the estimate of TCR in two ways. Firstly, we estimate the baseline temperature by regressing the temperatures in the first 20 years of the 1 percent CO2 ramp experiment as a function of time (see Supplementary Figure S5). Anomalies in temperature (and TOA fluxes for ECS) are measured relative to the corrected baselines derived from the 1pctCO2 simulation, and estimated linear pre-industrial trends are subtracted from the 1pctCO2 and abrupt4xCO2 timeseries. This pre-processing of the temperature timeseries improves the correlation between TCR and 21st century warming under

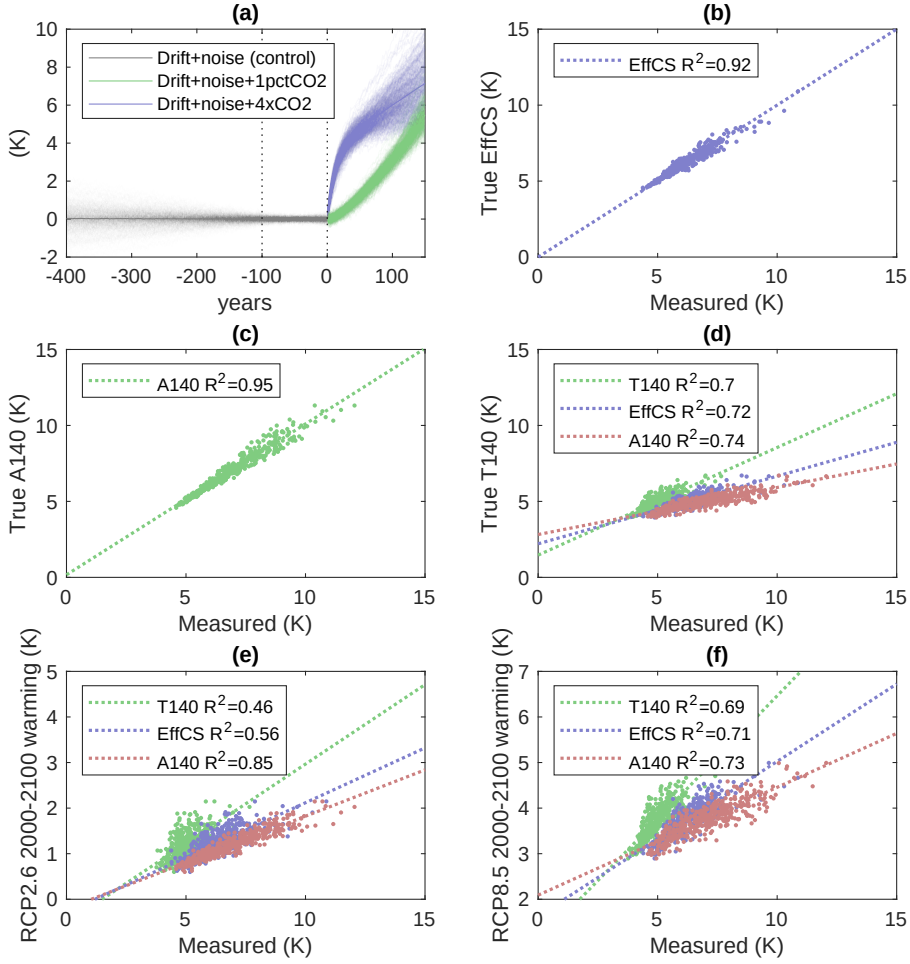
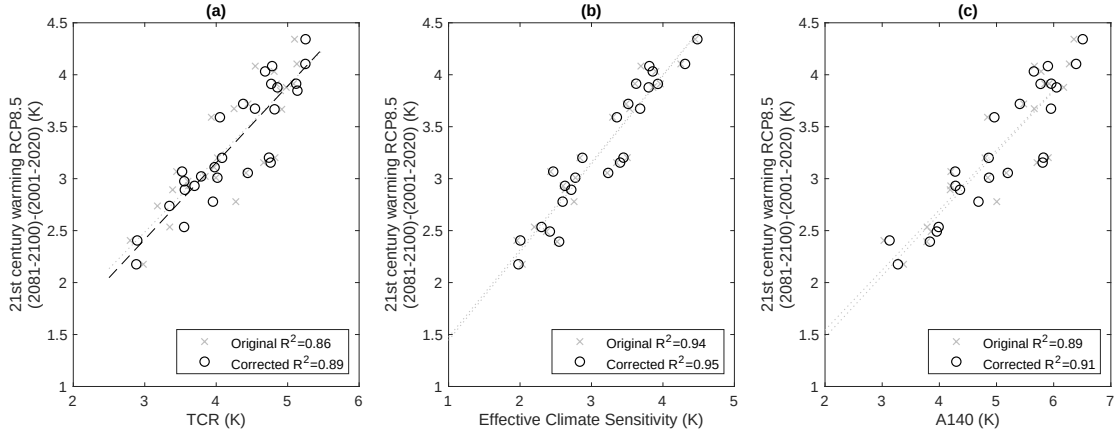


Figure 4. An idealized ensemble of simple models, where model parameters are identical to those considered in Figure 1(b), but models are initialized in a non-equilibrium state such that the baseline period is subject to some control drift, and model output is also subject to interannual variability of a similar magnitude to models in the CMIP archive. (a) shows global mean temperature evolution for the control period (gray), abrupt4xCO2 simulation (blue) and 1pctCO2 simulation (green). (b,c) show the true value of (EffCS,A140) as calculated in the noise-free, equilibrated simulations, plotted as a function of the measured value of (EffCS,A140) in a noisy, non-equilibrated simulations. (d,f,g) shows the true value of (T140,RCP2.6,RCP8.5 2000-2100 warming) plotted as a function of the measured values of T140, EffCS and A140 respectively.

Predictor	T140	EffCS	A140	RCP8.5 2000-2100	RCP2.6 2000-2100
T140 (true)	1.00	0.78	0.77	0.99	0.65
EffCS (true)	0.78	1.00	0.70	0.77	0.62
A140 (true)	0.77	0.70	1.00	0.76	0.91
T140 (drift)	0.74	0.58	0.59	0.73	0.50
EffCS (drift)	0.73	0.94	0.67	0.73	0.59
A140 (drift)	0.74	0.67	0.95	0.73	0.86
T140 (noise)	0.99	0.77	0.76	0.98	0.65
EffCS (noise)	0.78	1.00	0.69	0.77	0.61
A140 (noise)	0.78	0.70	1.00	0.77	0.91
T140 (drift+noise)	0.70	0.55	0.55	0.69	0.47
EffCS (drift+noise)	0.72	0.93	0.65	0.71	0.58
A140 (drift+noise)	0.73	0.66	0.94	0.72	0.85

Table 2. A table showing R^2 regression statistics relating a set of predictors to a set of unbiased model properties. Predictors are Transient Climate Sensitivity at quadrupling of CO2 (T140), Effective Climate Sensitivity (EffCS) and warming 140 years after a quadrupling of CO2 (A140), additional rows show these values measured experiments conducted with unequilibrated base climates (drift), additive autoregressive noise (noise) and a combination of both factors (drift+noise). 'True' output model properties (T140, EffCS, A140, RCP8.5 and RCP2.6 warming from 2000 to 2100) are derived from the equilibrated model without noise.

Figure 5. Plots showing the correlation between TCR (a), EffCS (b) and A140 (c) with 21st century warming, here represented by the difference between 2001-2020 and 2081-2100 global mean temperatures in the 1st ensemble member for each model in the CMIP5 archive for the RCP8.5 scenario. Each plot shows the 'original' calculation, where the baseline temperatures (and TOA fluxes for EffCS) are taken as the PIControl mean. In the 'corrected' calculation, a correction term for the baseline temperature and control drift is applied. Correlation coefficients are shown for the original and corrected cases.



RCP8.5 from 0.86 to 0.89. It also improves the correlation between EffCS and 21st century warming slightly from 0.94 to 0.95 (and A140 from 0.89 to 0.91).

These 'corrected' values (listed in Table 3) are estimates only, given we would expect the regression estimate based on a short 20 year period to be itself subject to internal variability noise, and we are assuming that the abrupt4xCO2 simulation and 1pctCO2 simulation have the same baselines. However, the improvement in correlation with future warming seen over the case with the pre-industrial average baseline supports the hypothesis that control drift adds uncertainty to the estimation of all quantities (and particularly TCR). However, it is not a complete explanation - and even after this adjustment, EffCS remains better correlated to RCP8.5 transient warming than TCR in the multi-model ensemble.

3 Conclusions

The question of which metric of climate sensitivity is most useful for summarizing uncertainty in future projections is conditional on a number of factors. ~~Clearly, any~~ Any single metric of sensitivity, even if known perfectly, ~~will not~~ cannot constrain Earth System response on all timescales and scenarios. We have shown here that one can produce a number of model variants which can exhibit the same value of EffCS or TCR, but with a range of responses, especially in a mitigation scenario such as RCP2.6.

In an idealized environment where models can be brought to a complete equilibrium control state, and ensemble sizes for ~~'~~'1pctCO2' simulations are large enough to avoid the effects of internal variability, the T140 metric would be the best idealized warming measure for century-scale warming under a high emissions scenario. However, the presence of even moderate control drift can act as a significant source of error in the measurement of T140, and so here we find that EffCS is likely to be a ~~better predictor of high emission warming in real-world applications~~ more accurate practical sensitivity metric in Earth System Model applications where full equilibration is difficult to achieve.

EffCS itself has limitations, it is relatively insensitive to slow timescale feedbacks, which means that it poorly correlated with century-scale warming under RCP2.6 (where a large fraction of warming occurs due to slow feedback response to historical emissions), and for warming on multi-century timescales under a high emissions scenario (where concentrations stabilize post-2100). We find ~~an~~ that a simple, but useful alternative is to simply use the mean warming from ~~the end~~ years 131-150 of the abrupt-4xCO2 simulation - which is comparably skilled to EffCS in predicting RCP8.5 warming in 2100, but more sensitive to century timescale feedbacks than EffCS - so therefore it is better correlated with RCP2.6 end of century warming~~(though it is subject to greater fractional error due to control model drift than EffCS, but less so than T140)~~.

~~Particularly concerning is that the two~~ It is notable that the most common metrics of sensitivity, EffCS, T140 and TCR, provide very little guidance on peak warming expected under climate mitigation. The focus on these metrics has also given rise to the issue that slow feedbacks in Earth System Models are not well constrained by the set of experiments currently conducted by default in CMIP. The standard 150 year simulation used to calculate Effective Climate Sensitivity does not constrain true Equilibrium Climate Sensitivity~~(see Additional Material)~~, and only a limited set of CMIP-class models have run models for long enough to be informative about equilibrium response (Rugenstein et al., 2019).

Model	EffCS (org)	EffCS (corr)	A140 (org)	A140 (corr)	T140 (org)	T140 (corr)	RCP8.5 2000-2100	RCP2.6 2000-2100
ACCESS1_0	3.48	3.53	5.48	5.60	4.45	4.57	3.72	-
ACCESS1_3	3.30	3.38	4.84	5.02	3.93	4.11	3.59	-
BNU_ESM	3.86	3.80	6.17	6.05	4.98	4.86	3.88	0.63
CCSM4	2.84	2.87	4.80	4.86	4.02	4.08	3.20	0.44
CESM1_CAM5_1_FV2	3.31	2.89	5.29	4.44	-	-	-	-
CNRM_CM5	3.22	3.28	5.17	5.30	4.42	4.54	3.06	0.68
CNRM_CM5_2	3.37	3.37	5.11	5.12	4.29	4.29	-	-
CSIRO_Mk3_6_0	3.53	3.63	5.66	5.86	4.25	4.45	3.67	1.09
CanESM2	3.61	3.59	5.92	5.89	5.08	5.05	3.91	0.92
FGOALS_s2	3.85	3.78	5.90	5.76	4.76	4.62	-	-
GFDL_CM3	3.69	3.87	5.66	6.02	4.55	4.90	4.08	1.25
GFDL_ESM2G	2.37	2.34	3.86	3.80	-	-	2.49	-0.08
GFDL_ESM2M	2.52	2.60	3.78	3.93	-	-	2.39	0.32
GISS_E2_H	2.20	2.42	3.79	4.23	3.35	3.79	2.53	0.36
GISS_E2_R	2.03	2.01	3.37	3.34	2.98	2.94	2.18	0.09
HadGEM2_ES	4.25	4.34	6.27	6.45	5.13	5.30	4.10	0.87
IPSL_CM5A_LR	3.90	3.92	5.78	5.78	4.81	4.81	4.03	0.80
IPSL_CM5A_MR	3.96	4.01	5.84	5.93	4.84	4.93	3.91	0.59
IPSL_CM5B_LR	2.43	2.54	4.20	4.43	3.45	3.67	3.07	-
MIROC_ESM	4.45	4.51	6.35	6.56	5.10	5.30	4.34	1.26
MIROC5	2.60	2.62	4.20	4.27	3.61	3.68	2.93	0.62
MPI_ESM_LR	3.50	3.45	5.91	5.82	4.82	4.74	3.20	0.43
MPI_ESM_MR	3.35	3.42	5.71	5.84	4.66	4.80	3.15	0.36
MPI_ESM_P	3.34	3.31	5.71	5.64	4.57	4.49	-	-
NorESM1_M	2.63	2.68	4.19	4.29	3.39	3.49	2.89	0.55
bcc_csm1_1	2.77	2.77	4.85	4.87	4.00	4.02	3.01	0.52

~~Summary metrics may have value if the context of those metrics , and their range of applicability in relation to real-world futures is well understood, but their limitations should be kept in mind. Although it has been convincingly demonstrated that the~~ It should be noted that these conclusions are derived from the consideration of a relatively simple two-timescale pulse response model. In this model, we can show that certain sensitivity metrics are insufficient to constrain future projections, and that non-equilibration may confound measurement. However, the constrained distributions for the metrics are subject to the structural assumptions of the model. The real world may have more than two response timescales Aengenheyster et al. (2018) , or may be better described as a continuous sum Ragone et al. (2016); Lembo et al. (2019). Further work should identify how such complexity impacts uncertainty in relevant climate metrics.

The diversity of simulated global mean dynamical response to greenhouse gas forcing over the coming centuries can be represented in simple models with a relatively small number of parameters (Smith et al., 2018; Meinshausen et al., 2011), ~~this number is greater than one.~~ but we cannot reduce uncertainty in climate projections on all timescales to a single degree of freedom. Summary metrics of climate response have value if the context of those metrics (and their range of applicability in relation to projection uncertainty) is well understood, but their limitations should be kept in mind.

Data availability. CMIP5 and CMIP6 data are available through a distributed data archive developed and operated by the Earth System Grid Federation (ESGF).

Code and data availability. Code for this study is available on Github at https://github.com/benmsanderson/matlab_pulse

4 ~~Methods~~

3.1 ~~Equilibrium sensitivity calculation~~

~~The two-timescale impulse response model follows the thermal feedback-timescale implementation from the FAIR simple climate model(Smith et al., 2018; Millar et al., 2017),resulting in a simple model for temperature and radiation response to a step change in forcing:~~

$$P(t) = F_{4xCO2} \sum_{n=1}^{2} q_n (1 - exp(-t/d_n))$$

$$R(t) = F_{4xCO2} \sum_{n=1}^{2} r_n (exp(-t/d_n)),$$

~~where $P(t)$ is the annual global mean temperature and $R(t)$ is the net top-of atmosphere radiative imbalance, and F_{4xCO2} is the instantaneous global mean radiative forcing associated with a quadrupling of CO_2, taken here to be $3.7 W m^{-2}$(Myhre et al., 2013) .~~

Constraining thermal parameters from historical temperatures and concentrations requires a consideration of other climate forcers. MCMC optimization of even a simple model of this form requires 10^7 or more calculations, so a very rapid model is required for computational tractability.

This study employs a fast pulse-response model to represent the response of surface global mean surface temperatures to forcing changes, where the model is implemented as a digital filter in MATLAB (see attached code) - allowing efficient computation and enabling Markov-Chain Monte Carlo parameter estimation for the physical parameters.

The thermal response is calculated by expressing the derivative of the forcing timeseries $F(t)$ as a series of step functions and using the CO2 quadrupling response T_p from equation 12 to calculate the integrated thermal response.

$$T(t) = \int_0^t \frac{\frac{dF}{dt}(t')}{F_{4xCO2}} T_p(t-t')dt',$$

Heat fluxes into the deep ($D(t)$) and shallow ($H(t)$) ocean components are estimated by the slow (n=1) and fast (n=2) components of $R(t)$.

3.0.1 Model Optimization

The model input time-series for calibration are observed CO_2 concentrations, along with radiative estimates from (Meinshausen et al., 2011) of non-CO2 forcing agents. We optimize the thermal model parameters for 2 timescales $[\mathbf{q}, \mathbf{d}, \mathbf{r}]$ and the non-CO2 forcing factor f_r. Optimization is conducted with the (Goodman and Weare, 2010) MCMC implementation, using flat initial parameter distributions as shown in Table 1, 200 walkers and 50,000 iterations for each optimization. Cost functions are computed for global mean temperature and global CO_2 concentrations.

$$E_T = \sum_t \left(\frac{(T(t) - T_{GCM}(t))}{\sqrt{2}\sigma_T} \right)^2$$

$$E_H = \sum_t \left(\frac{(H(t) - H_{GCM}(t))}{\sqrt{2}\sigma_H} \right)^2,$$

$$E_D = \sum_t \left(\frac{(D(t) - D_{GCM}(t))}{\sqrt{2}\sigma_D} \right)^2,$$

where σ_T is defined as for the abrupt-CO2 case as the standard deviation of HadCRUT 1850-1950 values. Shallow and Deep Ocean heat fluxes are taken as the 0-300m and 300m+ heat content derivatives respectively in (Zanna et al., 2019), with σ_H and σ_D taken as 1850-1950 standard deviations from the same dataset.

Flat priors are used for all parameters, with an additional prior on true equilibrium climate sensitivity using the likely value and upper bound on Equilibrium Climate Sensitivity from (Goodman and Weare, 2010) fit the median and 90th percentile of a gamma distribution for equilibrium (i.e. warming as $t \to \infty$).

Long name Symbol Min Max Thermal equilibration of deep ocean Sensitivity (KWm^{-2}) q_1 0 10* Thermal adjustment of upper ocean Sensitivity (KWm^{-2}) q_2 0 10 Thermal equilibration of deep ocean timescale $(years)$ d_1 100 4000 Thermal adjustment of upper ocean timescale $(years)$ d_2 10 100 Fraction of forcing in deep ocean response f_r 0. 1 Non-CO2 Forcing ratio f_r .7 1.3 A table showing model parameter values and minimum and maximum values allowed in model optimization.

3.0.1 Idealized Simulations

The posterior distribution of model configurations is then used to simulate a range of self-consistent values for various climate sensitivity metrics . Effective Climate Sensitivity is measured by implementing a step-change abrupt CO_2 quadrupling, and following (Gregory et al., 2004) to assess the linear extrapolation of warming at the point of net top of atmosphere energetic balance. A140 is calculated as the average of year 131-150 of the abrupt4xCO2 simulation. TCR and T140 are calculated as the average of years 61-80 and 131-150 respectively of the 1pctCO2 simulation, where CO_2 concentrations are increased annually by 1pct resulting in a linear increase in climate forcing. RCP scenario temperature trajectories are calculated for each parameter set using concentration and forcing timeseries from (Meinshausen et al., 2011) from 1850 until 2300.

We consider the range of control drifts observed in the CMIP5 and CMIP6 ensembles (illustrated in Figure 3(L)) which range from -.3 to +.6K /century in the CMIP5 and CMIP6 models considered in this study. An idealized distribution of drift in the simple model ensemble is created by initializing the model 500 years before the abrupt4xCO2 or 1pctCO2 simulation with a non-zero, constant forcing drawn from a flat distribution ranging from -1 to $+1 Wm^{-2}$, which results in a distribution of control drift of -.4K to +.4K per century (i.e. broadly comparable to the CMIP case). For each simulation we consider a baseline for temperature to be defined by the average global mean temperature in years 400-500.

To represent the first order effect of climate noise, we fit a 2nd order autoregressive model to the detrended global mean temperature timeseries in each available model in the CMIP5/6 ensemble. Taking CMIP mean parameters for the variance and autoregressive parameters, we generate noise for each realization of the simple model (though we note, in practise that the noise characteristics vary by model).

Author contributions. The author performed all analysis and writing for this project

Competing interests. The author declares no competing interests

Acknowledgements. This work is funded by the French National Research Agency, project number ANR-17-MPGA-0016. Benjamin Sanderson is an affiliate scientist with the National Center for Atmospheric Research, sponsored by the National Science Foundation.

References

Aengenheyster, M., Feng, Q. Y., Van Der Ploeg, F., and Dijkstra, H. A.: The point of no return for climate action, Earth System Dynamics, 9, 2018.

Andrews, T., Gregory, J. M., and Webb, M. J.: The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models, Journal of Climate, 28, 1630–1648, 2015.

Andrews, T., Gregory, J. M., Paynter, D., Silvers, L. G., Zhou, C., Mauritsen, T., Webb, M. J., Armour, K. C., Forster, P. M., and Titchner, H.: Accounting for changing temperature patterns increases historical estimates of climate sensitivity, Geophysical Research Letters, 45, 8490–8499, 2018.

Armour, K. C., Bitz, C. M., and Roe, G. H.: Time-varying climate sensitivity from regional feedbacks, Journal of Climate, 26, 4518–4534, 2013.

Charney, J., Arakawa, A., Baker, D., Bolin, B., Dickinson, R., Goody, R., Leith, C., Stommel, H., and Wunsch, C.: Carbon Dioxide and Climate: A Scientific Assessment: Report of an Ad Hoc Study Group on Carbon Dioxide and Climate, Woods Hole, Massachusetts, July 23-27, 1979 to the Climate Research Board, Assembly of Mathematical and Physical Sciences, National Research Council, National Academies, 1979.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J.: emcee: the MCMC hammer, Publications of the Astronomical Society of the Pacific, 125, 306, 2013.

Geoffroy, O., Saint-Martin, D., Bellon, G., Voldoire, A., Olivié, D., and Tytéca, S.: Transient climate response in a two-layer energy-balance model. Part II: Representation of the efficacy of deep-ocean heat uptake and validation for CMIP5 AOGCMs, Journal of Climate, 26, 1859–1876, 2013.

Goodman, J. and Weare, J.: Ensemble samplers with affine invariance, Communications in applied mathematics and computational science, 5, 65–80, 2010.

Gregory, J., Ingram, W., Palmer, M., Jones, G., Stott, P., Thorpe, R., Lowe, J., Johns, T., and Williams, K.: A new method for diagnosing radiative forcing and climate sensitivity, Geophysical Research Letters, 31, 2004.

Grose, M. R., Gregory, J., Colman, R., and Andrews, T.: What Climate Sensitivity Index Is Most Useful for Projections?, Geophysical Research Letters, 45, 1559–1566, 2018.

Hansen, J., Lacis, A., Rind, D., Russell, G., Stone, P., Fung, I., Ruedy, R., and Lerner, J.: Climate sensitivity: Analysis of feedback mechanisms, Climate processes and climate sensitivity, 29, 130–163, 1984.

Hasselmann, K., Sausen, R., Maier-Reimer, E., and Voss, R.: On the cold start problem in transient simulations with coupled atmosphere-ocean models, Climate Dynamics, 9, 53–61, 1993.

Held, I. M., Winton, M., Takahashi, K., Delworth, T., Zeng, F., and Vallis, G. K.: Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing, Journal of Climate, 23, 2418–2427, 2010.

Knutti, R. and Rugenstein, M. A.: Feedbacks, climate sensitivity and the limits of linear models, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373, 20150 146, 2015.

Knutti, R., Rugenstein, M. A., and Hegerl, G. C.: Beyond equilibrium climate sensitivity, Nature Geoscience, 10, 727, 2017.

Lembo, V., Lunkeit, F., and Lucarini, V.: TheDiaTo (v1. 0)–a new diagnostic tool for water, energy and entropy budgets in climate models, Geoscientific Model Development, 12, 3805–3834, 2019.

Lewis, N.: An objective Bayesian improved approach for applying optimal fingerprint techniques to estimate climate sensitivity, Journal of Climate, 26, 7414–7429, 2013.

Lucarini, V., Ragone, F., and Lunkeit, F.: Predicting climate change using response theory: Global averages and spatial patterns, Journal of Statistical Physics, 166, 1036–1064, 2017.

MacKay, D. J. and Mac Kay, D. J.: Information theory, inference and learning algorithms, Cambridge university press, 2003.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., et al.: Tuning the climate of a global model, Journal of advances in modeling Earth systems, 4, 2012.

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M., Lamarque, J.-F., Matsumoto, K., Montzka, S., Raper, S., Riahi, K., et al.: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300, Climatic change, 109, 213, 2011.

Millar, R. J., Nicholls, Z. R., Friedlingstein, P., and Allen, M. R.: A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions, Atmospheric Chemistry and Physics, 17, 7213–7228, 2017.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, Journal of Geophysical Research: Atmospheres, 117, 2012.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., and Zhang, H.: Anthropogenic and natural radiative forcing, pp. 659–740, Cambridge University Press, Cambridge, UK, https://doi.org/10.1017/CBO9781107415324.018, 2013.

Otto, A., Otto, F. E., Boucher, O., Church, J., Hegerl, G., Forster, P. M., Gillett, N. P., Gregory, J., Johnson, G. C., Knutti, R., et al.: Energy budget constraints on climate response, Nature Geoscience, 6, 415, 2013.

Proistosescu, C. and Huybers, P. J.: Slow climate mode reconciles historical and model-based estimates of climate sensitivity, Science Advances, 3, e1602 821, 2017.

Ragone, F., Lucarini, V., and Lunkeit, F.: A new framework for climate sensitivity and prediction: a modelling perspective, Climate Dynamics, 46, 1459–1471, 2016.

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5—A scenario of comparatively high greenhouse gas emissions, Climatic Change, 109, 33, 2011.

Ruelle, D.: General linear response formula in statistical mechanics, and the fluctuation-dissipation theorem far from equilibrium, Physics Letters A, 245, 220–224, 1998.

Rugenstein, M., Bloch-Johnson, J., Gregory, J., Andrews, T., Mauritsen, T., Li, C., Frölicher, T., Paynter, D., Danabasoglu, G., Yang, S., Dufresne, J.-L., Cao, L., SchmidtG.A., Abe-Ouchi, A., Geoffroy, O., and Knutti, R.: Equilibrium climate sensitivity estimated by equilibrating climate models, Geophysical Research Letters, in press, 2019.

Rugenstein, M. A., Caldeira, K., and Knutti, R.: Dependence of global radiative feedbacks on evolving patterns of surface heat fluxes, Geophysical Research Letters, 43, 9877–9885, 2016.

Sanderson, B. M. and Knutti, R.: On the interpretation of constrained climate model ensembles, Geophysical Research Letters, 39, 2012.

Senior, C. A. and Mitchell, J. F.: The time-dependence of climate sensitivity, Geophysical Research Letters, 27, 2685–2688, 2000.

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., and Regayre, L. A.: FAIR v1. 3: A simple emissions-based impulse response and carbon cycle model, Geoscientific Model Development, 11, 2273–2297, 2018.

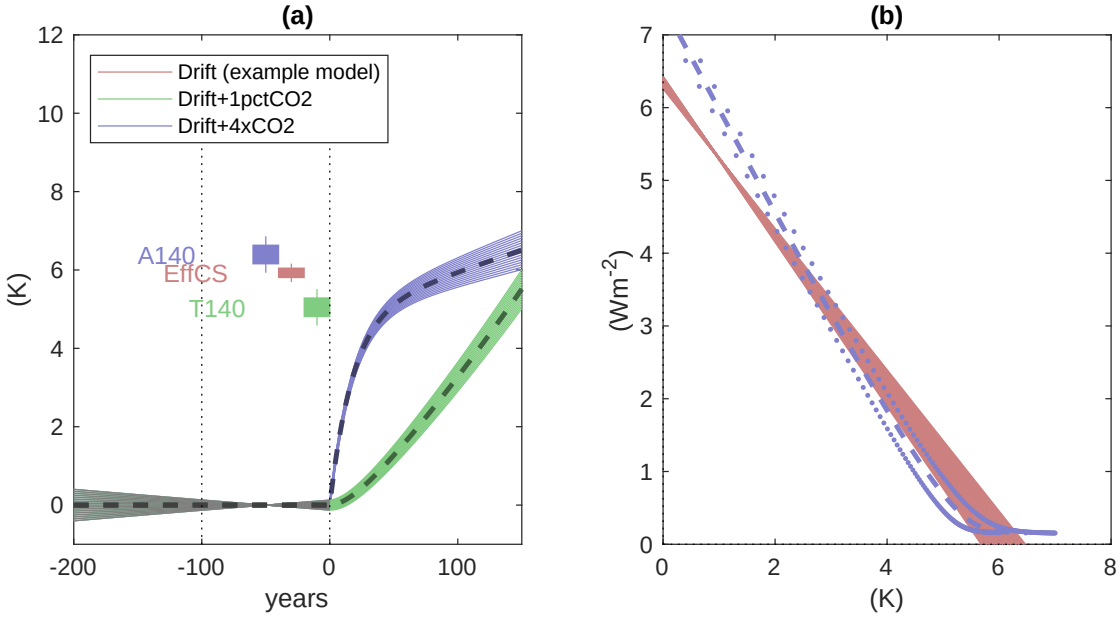
Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M. M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M.: Climate Change 2013 The Physical Science Basis, IPCC, 2013.

Van Vuuren, D. P., Stehfest, E., den Elzen, M. G., Kram, T., van Vliet, J., Deetman, S., Isaac, M., Goldewijk, K. K., Hof, A., Beltran, A. M., et al.: RCP2. 6: exploring the possibility to keep global mean temperature increase below 2 C, Climatic Change, 109, 95, 2011.

Winton, M., Takahashi, K., and Held, I. M.: Importance of ocean heat uptake efficacy to transient climate change, Journal of Climate, 23, 2333–2344, 2010.

Zanna, L., Khatiwala, S., Gregory, J. M., Ison, J., and Heimbach, P.: Global reconstruction of historical ocean heat storage and transport, Proceedings of the National Academy of Sciences, 116, 1126–1131, 2019.

Figure S1. Plots illustrating how different types of sensitivity metric are influenced by climatological drift. Each line describes the evolution of the model (with default parameters), where the control simulation is initialized 500 years in advance of the sensitivity experiment with a non-zero forcing ranging from -1 to 1 Wm^{-2}. (a) shows the global mean temperature time evolution of the abrupt4xCO2 simulations (blue) and the 1pctCO2 simulation (green), with box-whisker plots showing the range of biased values which are measured due to climate drift for A140, T140 and EffCS. (b) shows the trend lines used to compute the EffCS estimates from the simple model. Blue lines show an example model configuration response to an abrupt 4xCO2 perturbation in for the equilibrated case (dashed), and end members (+/- 1Wm^{-2} imbalance). The red shaded area shows the range of fitted trend lines consistent with (a).



Appendix : Supplementary Material

Figure S2. Scatterplots of 21st century warming (difference between 20 year means in 2081-2100 and 1981-2000) and a range of sensitivity metrics for CMIP5. TCR, T140 and EffCS are reported values from (Stocker et al., 2013), A140 is calculated as the year 131-150 average global mean temperature above the control level (taken as the last 100 years of the relevant control simulation). Columns represent different RCPs, rows represent different sensitivity metrics considered in the text. Each point represents a single model from the archive. Only results from the 1st initial condition ensemble member are considered for each model (thus the plots are subject to initial condition variability).

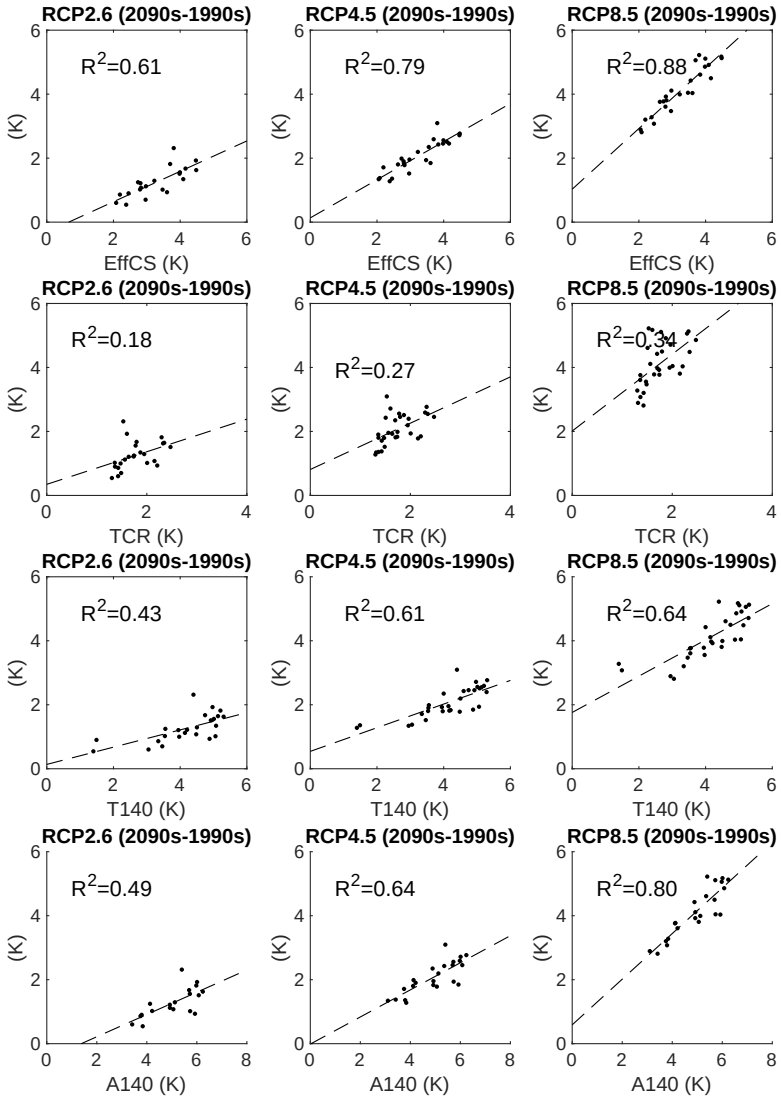


Figure S3. A 'corner-plot' showing the posterior parameter distribution attained by MCMC calibration of the simple climate model. Diagonal plots show posterior histograms for parameter values optimized in the calibration, while the horizontal range indicates the bounding values of the initial flat prior distribution. Off-diagonal plots show pairwise distributions of parameters in the posterior distribution.

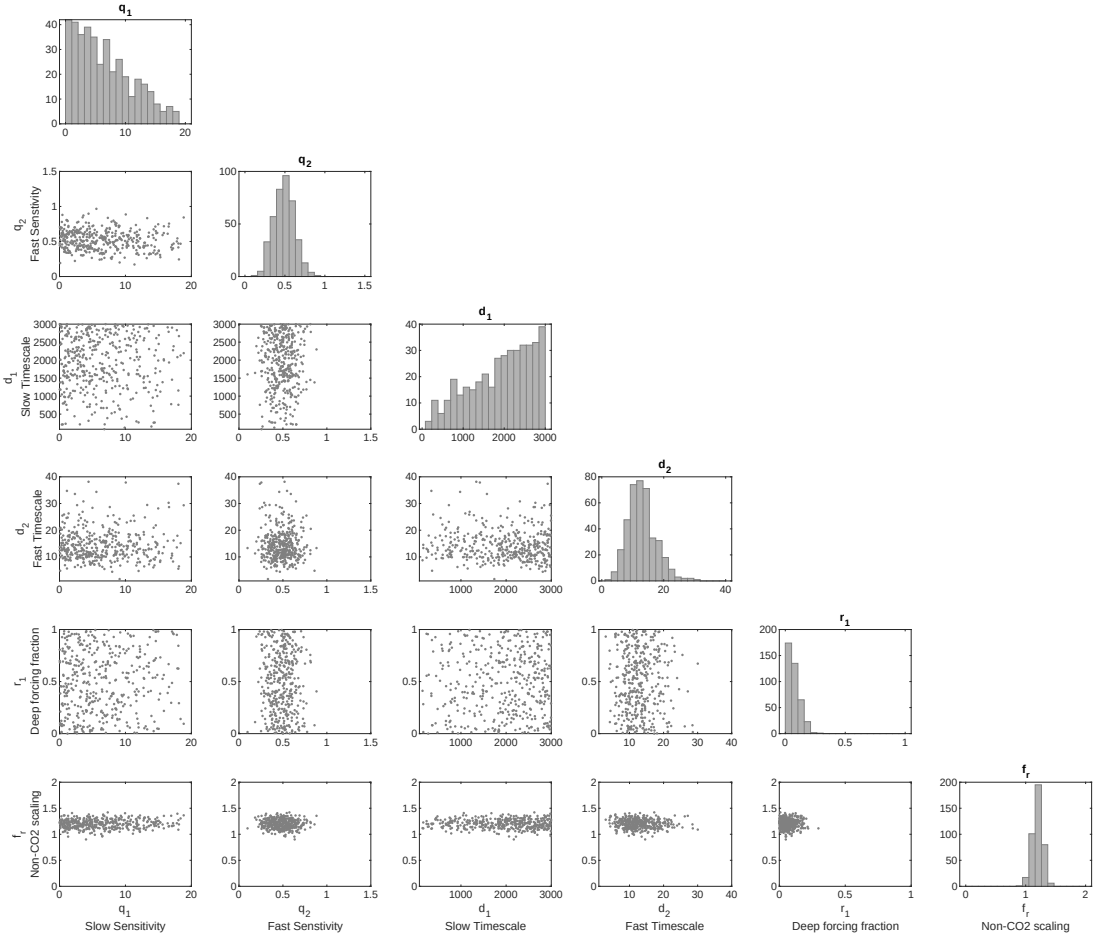


Figure S4. A demonstration of the simple model fitting strategy applied to historical simulations for a range of models in the CMIP5 archive. A pulse-response model is fitted treating each model's global mean temperature output in turn as truth for the period 1870-2019 (black line). 10th-90th percentiles of fitted temperature response for historical (grey area) and future projections are shown for RCP8.5 (pink area) and RCP2.6 (blue area) concentration pathways. Dotted lines show the median temperature in the ensemble projection, while solid colored lines show the evolution of the actual GCM for the corresponding scenario.

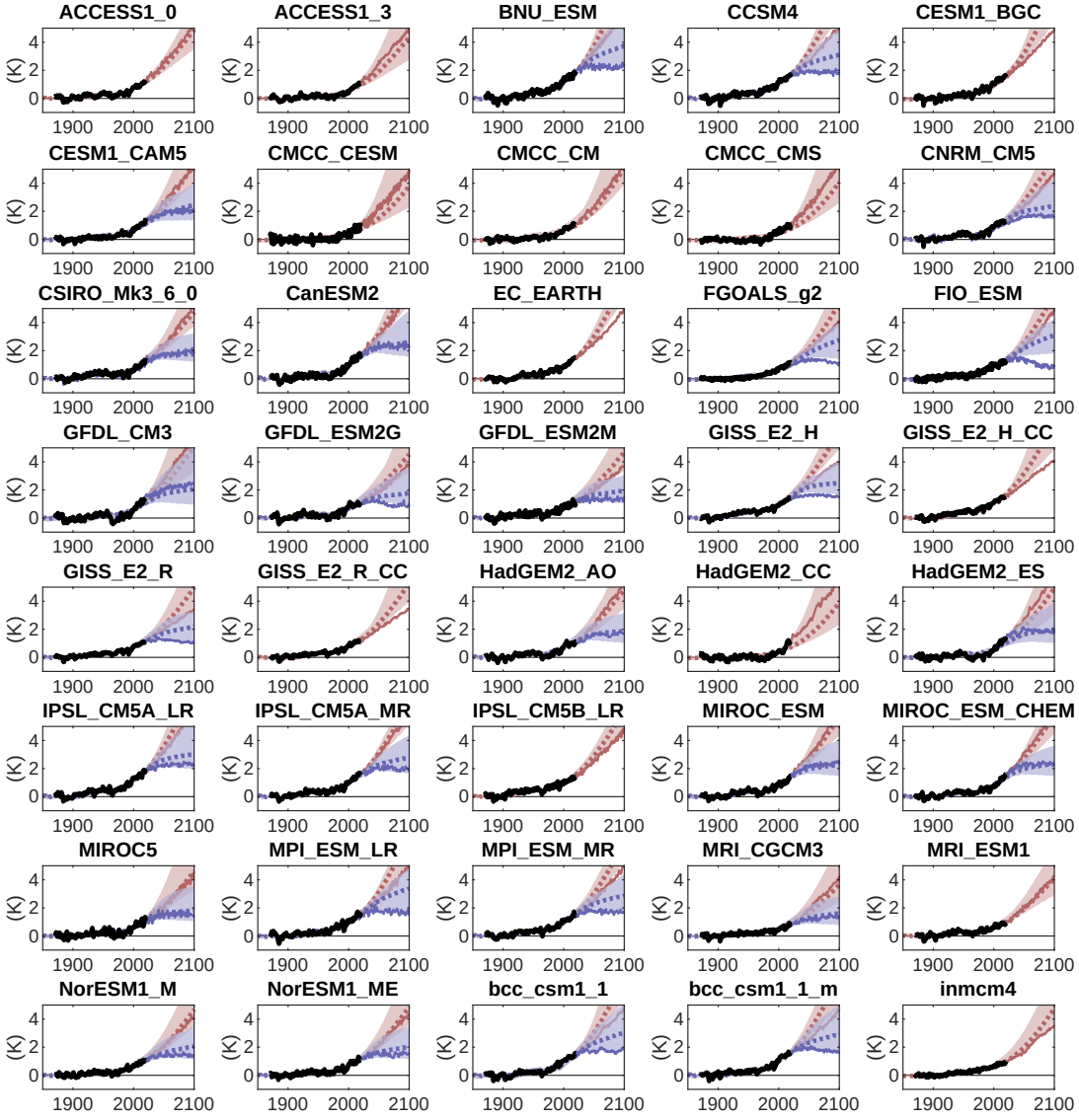


Figure S5. Figure illustrating the 'correction' employed for TCR and ECS in Figure 5. Corrected baseline temperatures are estimated by regression of the first 20 years of the control simulation, and branch-point from the control simulation is identifying by finding the year in which a linear fit to the control model evolution intersects the corrected baseline temperature. Branching in cases where there is no intersection are illustrated by the year in which the trendline is closest to the corrected baseline (either the first or last year).