Testing linearity and comparing linear response models for global surface temperatures

Hege-Beate Fredriksen[1], Kai-Uwe Eiselt[2], and Peter Good[3]

[1]Department of Physics and Technology, UiT The Arctic University of Norway
[2]The Arctic University of Norway
[3]Met Office

January 18, 2024

Abstract

Global temperature responses from different abrupt CO2 change experiments participating in Coupled Model Intercomparison Project Phase 6 (CMIP6) and LongRunMIP are systematically compared in order to study the linearity of the responses. For CMIP6 models, abrupt-4xCO2 experiments warm on average 2.2 times more than abrupt-2xCO2 experiments. A factor of about 2 can be attributed to the differences in forcing, and the rest is likely due to nonlinear responses. Abrupt-0p5xCO2 responses are weaker than abrupt-2xCO2, mostly because of weaker forcing. CMIP6 abrupt CO2 change experiments respond linearly enough to well reconstruct responses to other experiments, such as 1pctCO2, but uncertainties in the forcing can give uncertain responses. We derive also a generalised energy balance box model that includes the possibility of having oscillations in the global temperature responses. Oscillations are found in some models, and are connected to changes in ocean circulation and sea ice. Oscillating components connected to a cooling in the North Atlantic can counteract the long-term warming for decades or centuries and cause pauses in global temperature increase.

Testing linearity and comparing linear response models for global surface temperatures

Hege-Beate Fredriksen[1,2], Kai-Uwe Eiselt[1] and Peter Good[3]

[1]UiT the Arctic University of Norway, Tromsø, Norway
[2]Norwegian Polar Institute, Tromsø, Norway
[3]Met Office Hadley Centre, Exeter, United Kingdom

Key Points:

- We systematically compare different abrupt CO_2 change experiments from the Coupled Model Intercomparison Project 6 and LongRunMIP archives
- Linear response is overall a good assumption, but there is some uncertainty in how forcing varies with CO_2
- We derive a linear response model that can reproduce oscillations found in some models, linked to ocean circulation and sea ice changes

Corresponding author: Hege-Beate Fredriksen, `hege.fredriksen@npolar.no`

Abstract

Global temperature responses from different abrupt CO_2 change experiments participating in Coupled Model Intercomparison Project Phase 6 (CMIP6) and LongRunMIP are systematically compared in order to study the linearity of the responses. For CMIP6 models, abrupt-4xCO2 experiments warm on average 2.2 times more than abrupt-2xCO2 experiments. A factor of about 2 can be attributed to the differences in forcing, and the rest is likely due to nonlinear responses. Abrupt-0p5xCO2 responses are weaker than abrupt-2xCO2, mostly because of weaker forcing. CMIP6 abrupt CO_2 change experiments respond linearly enough to well reconstruct responses to other experiments, such as 1pctCO2, but uncertainties in the forcing can give uncertain responses. We derive also a generalised energy balance box model that includes the possibility of having oscillations in the global temperature responses. Oscillations are found in some models, and are connected to changes in ocean circulation and sea ice. Oscillating components connected to a cooling in the North Atlantic can counteract the long-term warming for decades or centuries and cause pauses in global temperature increase.

Plain Language Summary

We compare the global surface temperature responses in climate model experiments where the CO_2 concentration is abruptly changed from preindustrial levels and thereafter held constant. A quadrupling of CO_2 is expected to result in approximately twice the response to a doubling of CO_2. The ratio varies with time, but is on average 2.2 over the first 150 years. A factor 2 can be attributed to the radiative forcing, that is, how much the energy budget changes due to the change in CO_2. The remaining increase is likely due to stronger feedbacks. Experiments with half the CO_2 level are expected to have approximately the opposite response of a doubling, but we find their responses to be weaker. The reason appears to be a weaker radiative forcing. The evolution of the global temperature with time is also affected by changes in ocean heat uptake, ocean circulation, sea ice, cloud changes, etc., and these effects may be different with a stronger warming. Changes in the ocean circulation can also lead to oscillations appearing in addition to the warming. In some models, this effect may be strong enough to pause the long-term warming for decades or centuries, before it catches up again.

1 Introduction

Linear response is assumed for global surface temperature in many papers, resulting from e.g. box models (Geoffroy, Saint-Martin, Olivié, et al., 2013; Fredriksen & Rypdal, 2017; Caldeira & Myhrvold, 2013), and used in emulators like FaIR (Millar et al., 2017; Smith et al., 2018; Leach et al., 2021). It is based on the assumption that the global temperature response is independent of the climate state, and we can think of it as a powerful first-order approximation of the temperature response to a perturbation of the top-of-atmosphere (TOA) energy budget. For strong enough responses, state-dependent mechanisms like the albedo feedback will become important, so the question is: In what range of climate states can a linear response be considered a good assumption?

With a linear/impulse response model we can emulate the response to any known forcing within a few seconds, given knowledge about how the global temperature responds to an impulse. Alternatively, we can also gain this knowledge from step responses, since these are the integral of the impulse responses. The step-responses from experiments with abrupt quadrupling of the CO_2 concentration are typically used. This experiment is one of the DECK experiments required to participate in the Coupled Model Intercomparison Project (CMIP), and is therefore widely available.

Until recently, step-experiments with other CO_2 levels have only been available for a few models. Following the requests of nonlinMIP (Good et al., 2016), several CMIP6 models now make abrupt-2xCO2 and abrupt-0p5xCO2 experiments available. In addition, various abrupt CO_2 experiments are published through LongRunMIP (Rugenstein et al., 2019). The main motivation of this paper is to investigate the linearity of the temperature response by systematically comparing these different step experiments. That is, we want to test if the impulse response function derived from abrupt doubling of CO_2 experiments is equal (within expected uncertainties) to that derived from e.g. quadrupling of CO_2. This has implications for the concept of climate sensitivity – will the response to another doubling of CO_2 be similar to the first doubling?

In addition, we will discuss commonly used linear response models, derive the solution to a generalised box model, and study how well we can reconstruct the results of experiments that gradually increase the CO_2 concentration. With the generalised box model we demonstrate also how oscillations can appear in linear response models. The negative phase of oscillatory solutions may counteract the long-term warming for several decades, and these solutions can therefore be useful tools in understanding how plateaus or oscillations can appear in the global temperature responses to a step forcing, and how it is linked to changes in the ocean circulation and sea ice.

The generalised box model is described in Section 2. In Section 3 we discuss separation of forcing and response, and the linearity of global surface temperature response in the context of modifying the forcing-feedback framework to account for the non-constancy (or implicit time-dependence (Rohrschneider et al., 2019)) of global feedbacks. A non-constant feedback parameter just due to the pattern effect (a modulation of the global feedback from different paces of warming in different regions (Armour et al., 2013; Stevens et al., 2016; Andrews et al., 2015)) can be consistent with a linear response model, while state-dependent feedbacks imply a nonlinear response model. Section 4 describes the data included in this study and Section 5 describes estimation methods. Results are presented in sections 6 and 7, followed by a discussion in Section 8 and conclusions in Section 9.

2 Different linear response models, and their physical motivation

Generally, a linear response model for a climate state variable $\Phi(t)$ responding to a forcing $F(t)$ takes the form

$$\Phi(t) = G(t) * F(t) = \int_0^t G(t-s)F(s)ds, \tag{1}$$

assuming $F(t) = 0$ for $t \leq 0$ (Hasselmann et al., 1993). $G(t)$ is the Green's function, and $*$ denotes a convolution.

For global surface temperature, this integral can be interpreted as a part of the solution of a multibox energy balance model (see Fredriksen et al. (2021) and Appendix A),

$$\mathbf{C}\frac{d\mathbf{T}(t)}{dt} = \mathbf{K}\mathbf{T}(t) + \mathbf{F}(t) \tag{2}$$

where $\mathbf{C}$ is a diagonal matrix of heat capacities of different components of the climate system, $\mathbf{K}$ is a matrix of heat exchange coefficients, $\mathbf{T}$ is a vector of temperature responses, and $\mathbf{F}$ is a forcing vector. The two-box model (e.g. Geoffroy, Saint-Martin, Olivié, et al., 2013; Geoffroy, Saint-Martin, Bellon, et al., 2013; Held et al., 2010) is a widely used example. In appendix A we derive a general solution that can be applied to any linear K-box model, and find that in the case of only negative eigenvalues γ_n in the matrix $\mathbf{C}^{-1}\mathbf{K}$,

$$G(t) = \sum_{n=1}^{K} k_n e^{\gamma_n t}. \tag{3}$$

Hasselmann et al. (1993) notes that eigenvalues can also appear in complex pairs, where k_n and γ_n from one term of the pair are complex conjugates of the other term. To our knowledge, complex eigenvalues have never been used for estimating response functions in this field before. If pairs of complex eigenvalues are present, pairs from the sum above can be replaced by damped oscillatory responses on the form $k_1 e^{pt} \cos qt + k_2 e^{pt} \sin qt$ (see Appendix A). For these solutions to be stable, the real part of the eigenvalues (p) should be negative.

The step-forcing responses for negative eigenvalue solutions take the form:

$$T(t) = \sum_{n=1}^{K} S_n (1 - e^{\gamma_n t}) \tag{4}$$

and for complex eigenvalues, pairs from this sum are replaced by pairs on the form:

$$S_{osc1} \left[1 - e^{pt} \left(\cos qt - \frac{q}{p} \sin qt \right) \right] + S_{osc2} \left[1 - e^{pt} \left(\cos qt + \frac{p}{q} \sin qt \right) \right] \tag{5}$$

In these terms, the exponentially relaxing responses are modulated by sines and cosines.

So why do we want to expand the method to allow oscillatory responses for some models? It is not given that all eigenvalues of the linear model have to be negative if we allow the matrix $\mathbf{K}$ to have asymmetric terms. Asymmetric terms could for instance explain anomalies in energy fluxes following the ocean circulation, going only in one direction between two boxes. So if for instance the Atlantic Meridional Overturning Circulation (AMOC) has a strong response, this might require complex eigenvalues in a linear model for the surface temperature. And as we show in this paper, there are indeed models showing oscillations that can be described with such an oscillatory response function.

Since there could be many configurations of the box model (with different physical interpretations) leading to the same solution, from now on we will just work with the parameters in Eqs. (3, 4, 5) and not convert these to the parameters in the original box model in Eq. (2). When doing this we only have to specify the number of boxes used, and not worry about what is the best configuration of the boxes.

3 Distinguishing between forcing and response

The temperature response $T(t) = G(t) * F(t)$ cannot alone tell us how to distinguish between what is forcing and what is response to the forcing, since we can just move a factor between G and F without changing T. This separation is often done using the linear forcing - feedback framework, expressing the global top-of-the-atmosphere radiation imbalance (N) as

$$N = F + \lambda T \tag{6}$$

where $\lambda < 0$ is the feedback parameter, T is the global temperature response and F is the radiative forcing. This tells us how we can use the additional knowledge about the time series N to distinguish between F and T. However, it is now well known that the feedback parameter is not well approximated by a constant, so several modifications to this framework have been proposed to account for this. Note that how N relates to T does not impact the mathematical structure of the temperature response (as long as it is a linear relation), only how the forcing and feedbacks should be defined.

We can distinguish between three main classes of modifications:
(1) Assuming that N is a nonlinear function of T, e.g:

$$N = F + c_1 T + c_2 T^2 \tag{7}$$

This describes how λ could change with state (temperature) (Bloch-Johnson et al., 2015, 2021). Some examples of feedbacks that are well known to depend on temperature are the ice-albedo feedback and the water vapour feedback.

(2) Decomposing the surface temperature as

$$T = \sum_{n=1}^{K} T_n \tag{8}$$

and associate a feedback parameter λ_n with each component T_n, such that:

$$N = F + \sum_{n=1}^{K} \lambda_n T_n. \tag{9}$$

This can describe the pattern effect, if assuming different regions have different feedbacks and different amplitudes of the temperature response, which modulates the global value of λ with time (Armour et al., 2013). Proistosescu and Huybers (2017); Fredriksen et al. (2021, 2023) use such a decomposition of the temperature into linear responses with different time-scales.

Extending the decomposition of N in Eq. (9) to include oscillatory components may not be straight-forward if oscillations are in fact connected to the North Atlantic temperatures and changes in AMOC. The troposphere is very stable in this region and surface temperature changes are therefore confined in the lower troposphere, and not necessarily causing much change in the TOA radiation (Eiselt & Graversen, 2023; Jiang et al., 2023). Increasing surface temperatures in such stable regions lead to increased estimates of the climate sensitivity, interpreted as a positive lapse rate feedback (Lin et al., 2019). In the framework of Eq. (9) a possibility is to ignore or put less weight on the North Atlantic temperature component, due to the weaker connection between T and N here, but this needs to be further investigated in a future paper. Related effects can also play a role, for instance can AMOC changes lead to TOA radiation changes in surrounding areas, such as through low cloud changes in the tropics (Jiang et al., 2023). Such effects are likely model dependent.

(3) Descriptions using a heat-uptake efficacy factor ε, that describe how N depends on the heat uptake in the deeper ocean exist as well. This is mathematically equivalent to the second class for global quantities (Rohrschneider et al., 2019). In this description, the sum $T = \sum_{n=1}^{K} T_n$ is not necessarily considered a decomposition of the surface temperature, but includes also components describing temperature anomalies in the deeper ocean. If these temperatures are part of a linear model, typically a two- or three- box model, N can still be expressed as in Eq. (9). As these temperature components are just linear combinations of the components in Fredriksen et al. (2021); Proistosescu and Huybers (2017), it is only a matter of choice if expressing N using the temperatures in each box, or using the components of the diagonalized system, associated with different time scales of the system.

Descriptions with heat-uptake efficacy take slightly different forms in different papers. Winton et al. (2010) describes efficacy without specifying a model for the ocean heat uptake, while Held et al. (2010); Geoffroy, Saint-Martin, Bellon, et al. (2013) include it in the two-box model:

$$c_F \frac{dT}{dt} = -\beta T - \varepsilon H + F \tag{10}$$

$$c_D \frac{dT_D}{dt} = H \tag{11}$$

where T and T_D are the temperature anomalies of the surface and deep ocean boxes, respectively, and $H = \gamma(T - T_D)$ is the heat uptake of the deep ocean. The sum of the heat uptake in both layers equals N, leading to:

$$N = F - \beta T - (\varepsilon - 1)\gamma(T - T_D) \tag{12}$$

The concept of efficacy can be considered a way of retaining a "pattern effect" in box models with only one box connected to the surface, by relating the evolving spatial pattern of surface temperature change to the oceanic heat uptake (Held et al., 2010; Geoffroy & Saint-Martin, 2020). Similarly, efficacy of forcing (Hansen et al., 2005) has also been shown to be related to a "pattern effect" (Zhou et al., 2023), since forcing in different regions can trigger different atmospheric feedbacks.

Cummins et al. (2020); Leach et al. (2021) have modified this description to use it with a 3-box model, and use the heat uptake from the middle box to the deep ocean box to modify the radiative response

$$N(t) = F(t) - \lambda T_1(t) + (1-\varepsilon)\kappa_3[T_2(t) - T_3(t)] \tag{13}$$

If writing this equation in the form of Eq. (9), we find that the feedback parameters associated with $T_2(t)$ and $T_3(t)$ have equal magnitudes and opposite signs. This could put unfortunate constraints on parameters in this system, like net positive regional feedbacks, if interpreted as a pattern effect. We suggest avoiding this indirect description of the pattern effect with an efficacy parameter when using more than two boxes, and instead use a more direct interpretation of the parameters as describing a spatial pattern, such as Eq. (9).

3.1 Forcing defined using fixed-SST experiments

An alternative, that is not based on assumptions about the evolution of the feedbacks, is to run additional model experiments where sea-surface temperatures are kept fixed (Hansen et al., 2005; Pincus et al., 2016). These experiments aim to simulate close to 0 surface temperature change, such that $N \approx F$. Forcing estimated from these experiments have less uncertainty than regression methods based on the above-mentioned relationships between N, T and F (P. M. Forster et al., 2016), but are contaminated by land temperature responses. A forcing definition that includes all adjustements in N due to the forcing, but no adjustments due to surface temperature responses is the effective radiative forcing (ERF). This is considered the best predictor of surface temperatures, since it has forcing efficacy factors closest to 1 (Richardson et al., 2019). Ideally ERF should be estimated in models by fixing all surface temperatures, but this is technically challenging (Andrews et al., 2021). Instead, it is more common to correct the fixed-SST estimates for the land response (Richardson et al., 2019; Tang et al., 2019; Smith et al., 2020). We have not used these estimates in this paper, since they are not available for many models.

4 Choice of data

We compare abrupt-4xCO2 global temperature responses to all other abrupt CO2 experiments we can find. In the CMIP6 archive we have 12 models with abrupt-2xCO2 and 9 models with abrupt-0p5xCO2. In LongRunMIP we find 6 models with at least two different abrupt CO2 experiments, and we use the notation abruptNx to describe these, where N could be 2, 4, 6, 8 or 16. The advantage of models in LongRunMIP is that we can study responses also on millennial time scales, while for CMIP6 models the experiments are typically 150 years long.

There exist also similar comparisons of abrupt CO_2 experiments for a few other models outside of these larger data archives (e.g., Mitevski et al., 2021, 2022; Meraner et al., 2013; Rohrschneider et al., 2019). These data are not analysed in this study, but will be included in our discussion.

CMIP6 abrupt CO_2 experiments are used to reconstruct 1pctCO2 experiments, and the reconstructions are compared to the coupled model output of CMIP6 models. The reason for choosing this experiment is that the forcing is relatively well known. If assuming the forcing scales like the superlogarithimic formula of Etminan et al. (2016), it should increase slightly more than linearly until CO_2 is quadrupled, and end up at the same forcing level as the abrupt-4xCO2 experiments. The Etminan et al. (2016) forcing includes stratospheric adjustments, but not tropospheric and cloud adjustments like the ERF. However, we don't use the absolute values of this forcing, only the forcing ratios. We may also take these ratios as approximate ERF ratios if assuming the Etminan et al. (2016) forcing can be converted to ERF with a constant factor.

For other experiments, the uncertainty in forcing estimates is an even more important contribution to uncertainties in the responses. Jackson et al. (2022) test emulator responses to the Radiative Forcing Model Intercomparison Project (RFMIP) forcing for 8 models, and find large model differences in emulator performance. Using a different forcing estimation method (Fredriksen et al., 2021) for the CMIP6 models, Fredriksen et al. (2023) find a generally good emulator performance for historical and SSP experiments. An important difference between the forcing estimates is that the RFMIP forcing used by Jackson et al. (2022) is not corrected for land temperature responses, while the regression-based forcing in Fredriksen et al. (2023) is defined for no surface temperature response. The method described in Fredriksen et al. (2021, 2023) is actually designed to make forcing estimates compatible with a linear temperature response, and we therefore refer to these results for performance of linear response models for historical and future scenario forcing. However, if the linear response assumption is poor for the temperatures, this influences performance of the forcing estimation method as well. For this reason it is important to test the linear response hypothesis with idealized experiments, which is the focus of this paper.

4.1 AMOC and sea ice

In our discussion of oscillatory responses and plataeus in global temperature, we consider also AMOC and sea ice changes in the models. The AMOC index is calculated as the maximum of the meridional overturning stream function (*mstfmz* or *mstfyz* in CMIP6 and *moc* in LongRunMIP) north of 30°N in the Atlantic basin below 500 m depth.

The sea-ice area is calculated by multiplying the sea-ice concentration (*siconc* or *siconca* in CMIP6 and *sic* in LongRunMIP) with the cell area (*areacello* or *areacella*) and then summing separately over the northern and southern hemispheres.

5 Estimation

5.1 Forcing ratios for step experiments

A linear temperature response assumption predicts the response in any abrupt CO_2 experiment to be a scaled version of that of the abrupt-2xCO2 experiment, since only the forcing is different in these experiments. So when comparing abrupt CO2 experiments, they are all scaled to correspond to the abrupt-2xCO2 experiment. However, choosing the best scaling factor is challenging, since the forcing is uncertain, and it is not easy to distinguish between differences due to forcing and possible nonlinear temperature responses. Therefore, we have used three different types of scaling factors in our analysis:

1) Use the same scaling factor for all models, and assume a forcing scaling like the superlogarithmic radiative forcing (RF) formula in Etminan et al. (2016) in the CO_2 range where this formula is valid, and logarithmic forcing outside this range (just to have something in lack of a valid non-logarithmic description). The factors used are 0.478 for abrupt-4xCO2 and 0.363 for abrupt-6xCO2. A logarithmic dependence on the CO_2 concentrations results in the factors -1, 1/4 and 1/8 for the abrupt- 0p5xCO2, 8xCO2 and 16xCO2 experiments.

2) Estimate ratios by performing Gregory regressions (Gregory et al., 2004) of the first 5, 10, 20 and 30 years of the experiments.

3) Use the mean temperature ratio to the abrupt-2xCO2 experiment over the first 150 years as the scaling factor. This is not meant to be an unbiased estimate of the forcing ratio, but investigates the forcing ratios in the hypothetical case of per-

fectly linear responses. However, some degree of nonlinear response is expected e.g. from differences in feedbacks (Bloch-Johnson et al., 2021). After scaling temperature responses with this factor, it is easier to visualise how nonlinear responses affect different time scales of the response.

5.2 Reconstructing 1pctCO2 experiments

Performing an integration by parts of Eq. (1) leads to

$$T(t) = \int_0^t \frac{dF}{ds} R(t-s)ds, \tag{14}$$

where $R(t) = \int_0^t G(t-s)ds$ is the response to a unit-step forcing. Discretising this equation leads to the expression used to compute impulse responses in Good et al. (2011, 2013, 2016); Larson and Portmann (2016):

$$T_i = \sum_{j=0}^{i} \frac{\Delta F_j R_{i-j}}{\Delta F_s} \tag{15}$$

where ΔF_j are annual forcing increments, and the discretised step response R_{i-j} is a response to a general step forcing ΔF_s, and must therefore be normalised with this forcing. Further details of the derivation are provided in Fredriksen et al. (2021) Supplementary Text S2.

With Eq. (15) we can use datapoints from abrupt CO_2 experiments and knowledge of forcing to directly compute the responses to other experiments. Then we can avoid the additional uncertainty related to what model to fit and its parameter uncertainties. Fitting a box model first would smooth out internal variability from the step response function, which could be an advantage when studying responses to experiments with more variable forcing. Another advantage of box models is that the response function can be extrapolated into the future, while with Eq. (15) the length of the reconstruction is restricted by the length of the step experiment. Here we will use 140 years of data for the reconstruction of 1pctCO2 experiments, and as we will see, the reconstructed responses to 1pctCO2 experiments are already very smooth, so smoothing the response function with exponential responses should not change the results significantly, as long as the smoothed model provides a good fit to the datapoints.

To test this reconstruction, we will use CMIP6 annual anomalies from the experiments abrupt-4xCO2, abrupt-2xCO2 and abrupt-0p5xCO2. The input forcing ratio starts at 0, and increases either linearly, consistent with a logarithmic dependence on CO_2 concentration, or as a ratio scaling like the superlogarithmic formula (Etminan et al., 2016). For abrupt-4xCO2, we assume the ratio becomes 1 in year 140, the time of quadrupling, and for abrupt-2xCO2, we assume the ratio is 1 in year 70, the time of doubling. The positive 1pctCO2 forcing does not equal the negative abrupt-0p5xCO2 forcing at any time point, so we just assume the abrupt-0p5xCO2 forcing to be the negative of the abrupt-2xCO2 forcing.

5.3 Fitting response functions

We will compare estimated response models from a two-box model, three-box model, and a four-box model with one pair of complex eigenvalues. These response models consist of two or three exponential responses, or two exponential plus two damped oscillatory responses. Decomposing the response using box models may also help us gain insight into the physical reasons why a linear response model works or not.

We apply the python package lmfit to estimate the parameters of the response models. It takes in an initial parameter guess, and then searches for a solution that minimizes the least-squared errors. The final parameter estimates can be sensitive to the initial guesses,

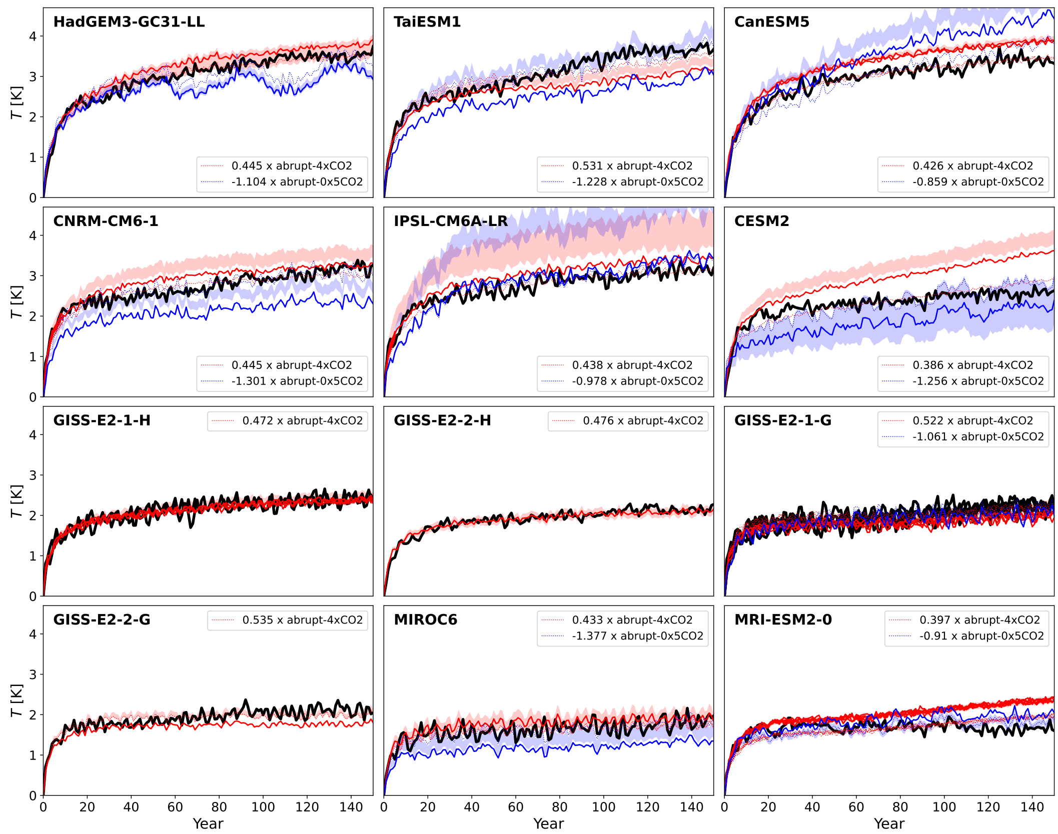

Figure 1. Comparing abrupt CO_2 experiments for CMIP6 models, where the abrupt-4xCO2 and abrupt-0p5xCO2 experiments are scaled in three different ways to correspond to the abrupt-2xCO2 experiment. Models are sorted by their abrupt-2xCO2 response in year 150. The black curves are abrupt-2xCO2 experiments, the red are scaled abrupt-4xCO2 and the blue are scaled abrupt-0p5xCO2 experiments. Solid curves use the same scaling factor for all models: 0.478 for abrupt-4xCO2 and -1 for abrupt-0p5xCO2. Thin dotted curves use the mean temperature ratio as the scaling factor (shown in legends and supplementary figure S1), and shading shows the range of the ratios of the Gregory regressions given in Supporting Tables S1 and S2.

since the optimization algorithm may just have found a local minimum. The more parameters we have in the model, the less we can trust the estimates. We see this in particular when including oscillatory responses; then we need to estimate 8 parameters, and are at risk of overfitting for the typical 150 year long experiments. As we will see, there could be different solutions containing oscillations that all provide good fits to the data. Longer time series (or some physical reasoning) would be needed in order to select the optimal fit for these records. For longer time series such as those from LongRunMIP we obtain more useful estimates.

6 Linear response results

6.1 Comparing abrupt CO_2 experiments

The curves in Figures 1 and 2 are all scaled to correspond to the abrupt-2xCO2 experiment, where the different scaling factors used illustrate the problem with the forcing uncertainty. The thick solid curves use the same scaling factor for all models (method 1), while the factors from the second and third method are model specific. The shading shows the range using the four different forcing ratios computed with Gregory regressions (method

2), that is, the minimum and maximum values from Tables S1 - S3. The thin dashed curves use the mean temperature ratios (method 3). These values are given in the subfigure legends, and shown in supporting figures S1 - S2. By definition, the black curves and the dotted red and blue curves all have the same time mean. Model specific factors can be explained by their different fast adjustments to the instantaneous radiative forcing. In addition, models can have different instantaneous forcing values, as this is shown to depend on the climatological base state (He et al., 2023). From the mean temperature ratios of the first 150 years of CMIP6 we find also that abrupt-4xCO2 warms on average 2.2 times more than abrupt-2xCO2, and abrupt-0p5xCO2 cools on average 9 % less than abrupt-2xCO2 warms (see Table S4). For LongRunMIP, abrupt4x warms 2.13 times abrupt2x when averaging all available years, or 2.18 times if averaging just the first 150 years (see Table S5, and both estimates exclude FAMOUS).

Significant differences between the curves in Figures 1 and 2 that cannot be explained by their different forcing must be explained by a nonlinear/state-dependent response. A first order assumption could be that models that warm more should tend to be more nonlinear. To investigate this we have ordered the models by their abrupt-2xCO2 response in year 150 in Figure 1 and year 500 for the longer experiments in Figure 2. We find that there are some clear differences for the warmest CMIP6 models, but also for the coldest (MRI-ESM2-0). The four different GISS models appear to be very linear.

For the two LongRunMIP models with the strongest 2xCO2 warming (CNRM-CM6-1 and FAMOUS) there are some clear differences between the curves (see Figure 2). The initial warming for CNRM-CM6-1 is halted in the 2xCO2 compared to the 4xCO2 experiment. For FAMOUS the scaling factor is particularly uncertain, and after a few centuries the pace of warming is slower in the scaled abrupt-4xCO2 experiment than in the abrupt-2xCO2 experiment. We observe only minor differences for MPI-ESM1-2, HadCM3L and CCSM3 when scaling with the mean temperature ratios. For CESM104 we observe that the abrupt2x experiment has some oscillations that are not seen in the other experiments, in addition to an abrupt change in the abrupt8x experiment.

If more warming increases the likelihood of finding nonlinear responses, we should also expect nonlinear responses to become more apparent towards the end of the simulations. We can then hypothesize that differences in forcing should explain initial differences (maybe up to a decade), and nonlinear responses explain differences at later stages. Following this, we should put more trust in the forcing scaling factors that make the initial temperature increase most similar to the abrupt-2xCO2 experiment. Which factor this is differs between models. In general, method 2 should put more emphasis on describing the first years correctly, while method 3 emphasises a good fit on all scales.

Although the individual forcing estimates are uncertain, it is a noteworthy result that the abrupt-2xCO2 regression forcing (method 2) is on average half of the abrupt-4xCO2 forcing (see Tables S1 and S3). The uncertainty of this mean is however too large to rule out that the forcing for a second CO_2 doubling is in fact larger than the first doubling, according to the findings of Etminan et al. (2016); He et al. (2023). And consistent with these expectations, for CMIP6 abrupt-0p5xCO2 we find a weaker negative forcing than logarithmic (Table S2). Our forcing ratios based on the LongRunMIP simulations for abrupt 6x, 8x and 16x CO_2 indicate that the forcing is weaker than logarithmic for higher CO_2 concentrations. Although based on very few simulations, this result is the opposite of the expectation that each CO_2 doubling produces stronger forcing (He et al., 2023).

An average forcing factor of 2 means the forcing alone is unlikely to explain the 2.2 factor difference in warming between CMIP6 abrupt-2xCO2 and abrupt-4xCO2. This conclusion is also supported by the differences in the pace of warming between abrupt-2xCO2 and abrupt-4xCO2 for several models (best visualised with the dotted curves from method 3 in Figure 1). The abrupt-4xCO2 temperatures scaled using method 2 in Figure 1 are

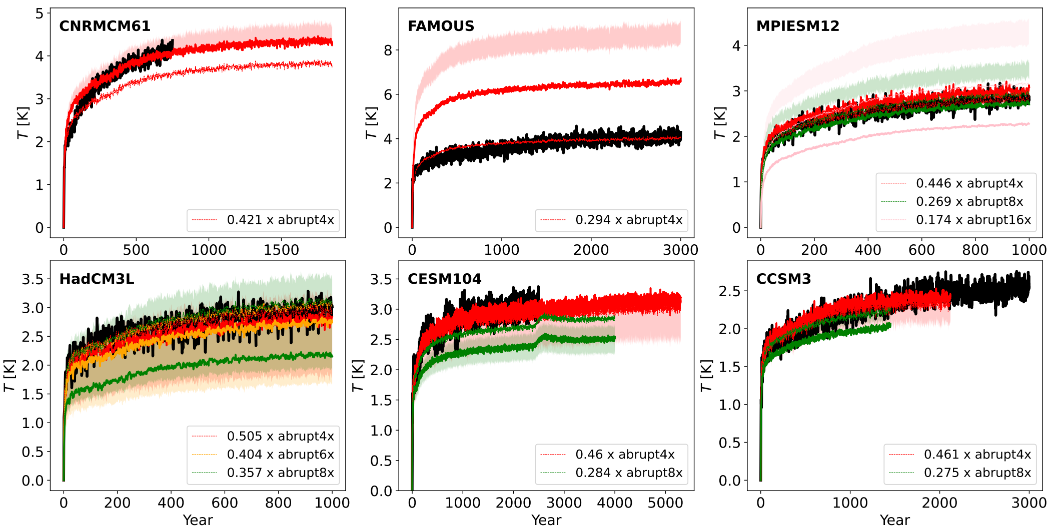

Figure 2. Comparing abrupt CO_2 experiments for LongRunMIP models. The scaling factors for the thick curves are 0.478 for 4x, 0.363 for 6x, 1/4 for 8x, 1/8 for 16x. For the thin dashed curves, the factors are computed from the mean T ratios to the first 150 years of abrupt2x, shown in Supporting figure S2, and shown in the legends here. The models are sorted by their abrupt2x temperature response in year 500. Note their different lengths and temperature scales.

on average 10 % stronger than the abrupt-2xCO2 experiments (computed from the ratio 2.2/2). The scaled abrupt-0p5xCO2 temperatures are on average 2 % stronger than the abrupt-2xCO2 temperatures (see Table S4), suggesting that the weak forcing can explain much of the weak response for abrupt-0p5xCO2. For LongRunMIP models, the average forcing ratio between 2x and 4x CO_2 reduces to 0.46 when excluding FAMOUS, making differences in the scaled temperatures over the first 150 years vanish (computed with method 2, see Table S5). For some models (CESM104 and CCSM3) the scaled temperatures deviate more from abrupt2x on millennial time scales.

Bloch-Johnson et al. (2021) suggests that feedback temperature dependence is the main reason why abrupt-4xCO2 warms more than twice the abrupt-2xCO2. This is consistent with the nonlinear responses we observe for several models. If the mean temperature ratio was a valid estimate of the forcing ratio, then in a linear framework, the same factors we found for the temperature ratios should be able to explain the ratios in top-of-atmosphere radiative imbalance. For some models this is not a good approximation (see supporting figures S1 and S2), consistent with the findings of Bloch-Johnson et al. (2021). FAMOUS has a particularly large difference in T and N ratios. Its abrupt4x warming is also so extreme that the quadratic model in Bloch-Johnson et al. (2021) suggests a runaway greenhouse effect.

6.2 Reconstructing 1pctCO2 experiments

In general, we find that both abrupt-4xCO2 experiments (see Figure 3) and abrupt-2xCO2 experiments (see Figure 4) can reconstruct the 1pctCO2 experiment very well. The largest deviation we find for the model KIOST-ESM, but we suspect the 1pctCO2 experiment from this model may have errors in the branch time information or the model setup. For many models the abrupt-0p5xCO2 experiment can also be used to make a good reconstruction, but not all (see Figure 4). For several models where abrupt-0p5xCO2 makes a poor reconstruction (TaiESM1, CNRM-CM6-1, CESM2, MIROC6), our assumptions about the forcing seems to be the limiting factor. If upscaling the negative of the abrupt-0p5xCO2 response for these models with a different factor than -1 to correspond better with the abrupt-2xCO2 experiment, we would have obtained a better reconstruction of 1pctCO2.

For many models we find that reconstructions with abrupt-4xCO2 slightly overestimates the 1pctCO2 response in the middle parts of the experiment, similar to earlier findings by Good et al. (2013); Gregory et al. (2015). In Figure 3 we compare reconstructions with a linear forcing (from logarithmic dependence on CO_2) and a forcing scaling like the superlogarithimic formula (Etminan et al., 2016). We find that reconstructions using the superlogarithmic forcing (shown in brown) explains the middle part of the 1pctCO2 experiment a little better than the logarithmic forcing (shown in red), since this forcing is slightly weaker in the middle. Even with the superlogarithmic forcing ratio, the model average reconstruction with abrupt-4xCO2 is a little overestimated in the middle part of the experiment (Figure 5). The average reconstruction with abrupt-2xCO2 explains the middle part of the experiment well, but slightly underestimates the latter part.

Which of abrupt-2xCO2 or abrupt-4xCO2 make the best reconstruction is model dependent. The 1pctCO2 experiment goes gradually to $4xCO_2$, and if there is a state-dependence involved in the response, we might expect something in between abrupt-2xCO2 and abrupt-4xCO2 responses to make the best prediction. MRI-ESM2-0 is a good example where this might be the case. For this model we observe a small underestimation with abrupt-2xCO2 and a small overestimation with abrupt-4xCO2. The reconstruction is very good with abrupt-0p5xCO2, which has an absolute response looking like an average of abrupt-2xCO2 and abrupt-4xCO2 (see Figure 1). CESM2 is also a good example where state-dependent effects are visible, since the abrupt-2xCO2 underestimates and abrupt-4xCO2 overestimates the response in the latest decades of the 1pctCO2 experiment.

For TaiESM1 and CNRM-CM6-1 the paces of warming differ a little for abrupt-2xCO2 and abrupt-4xCO2 during the middle/late stages of the experiments. Although the differences are not very significant, this is an indication of a nonlinear response. For some models (CanESM5, CNRM-CM6-1, HadGEM3-GC31-LL, IPSL-CM6A-LR, MIROC6) it is unclear if the small errors in the reconstructions are due to incorrect scaling of the forcing or nonlinear responses. The four GISS models are the most linear models, where we make good and very similar reconstructions with both abrupt-4xCO2 and abrupt-2xCO2. We observe just a small underestimation in the end of the experiment for GISS-E2-2-G abrupt-4xCO2.

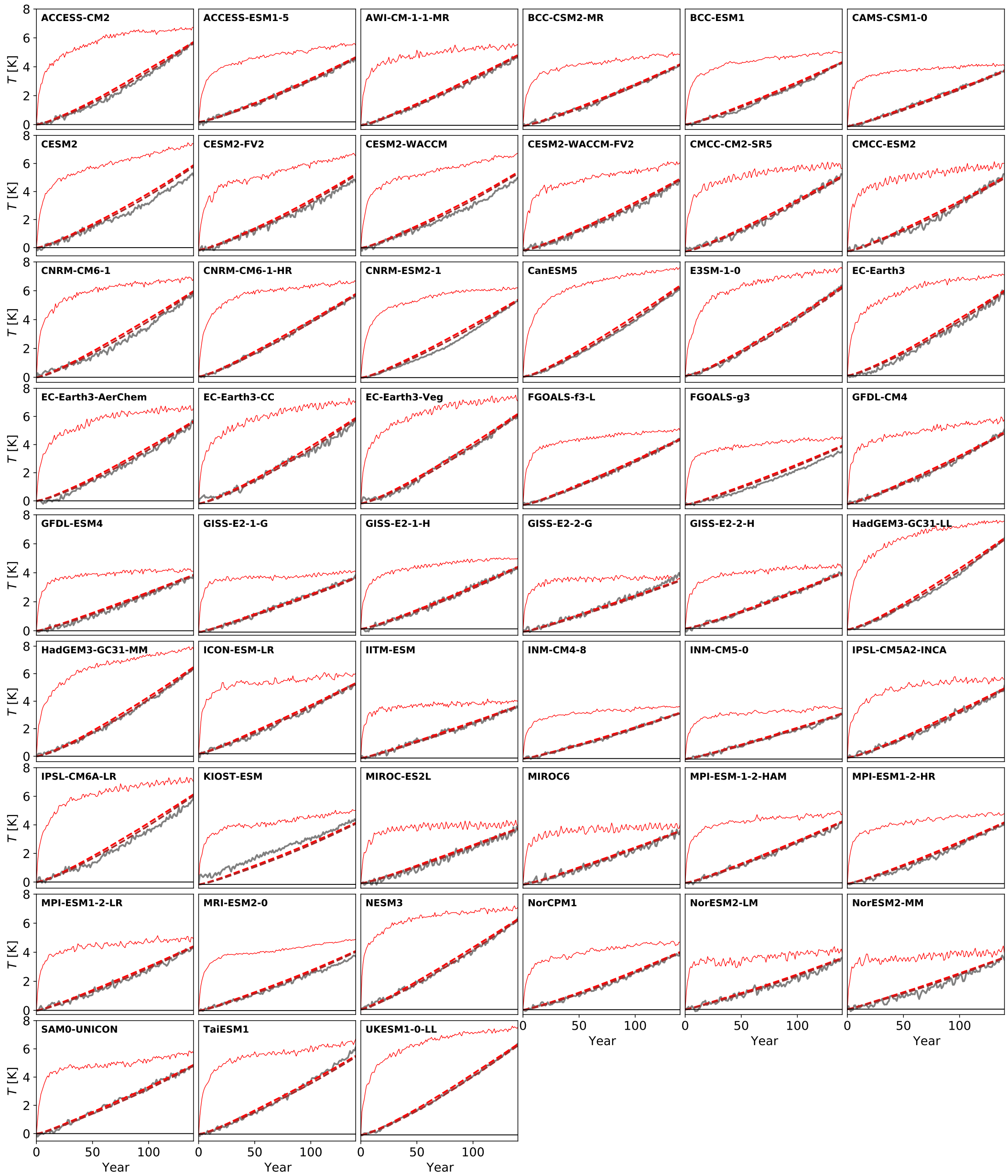

Figure 3. Red/brown dashed curves show reconstructions of the 1pctCO2 experiment (gray) using the data from the abrupt-4xCO2 experiment (red). The dashed red curve is a reconstruction based on a linearly increasing forcing, and the dashed brown curve is a reconstruction based on a forcing scaling like the superlogarithmic (Etminan et al., 2016) formula. For the experiments where several members exist, we have plotted the ensemble mean.

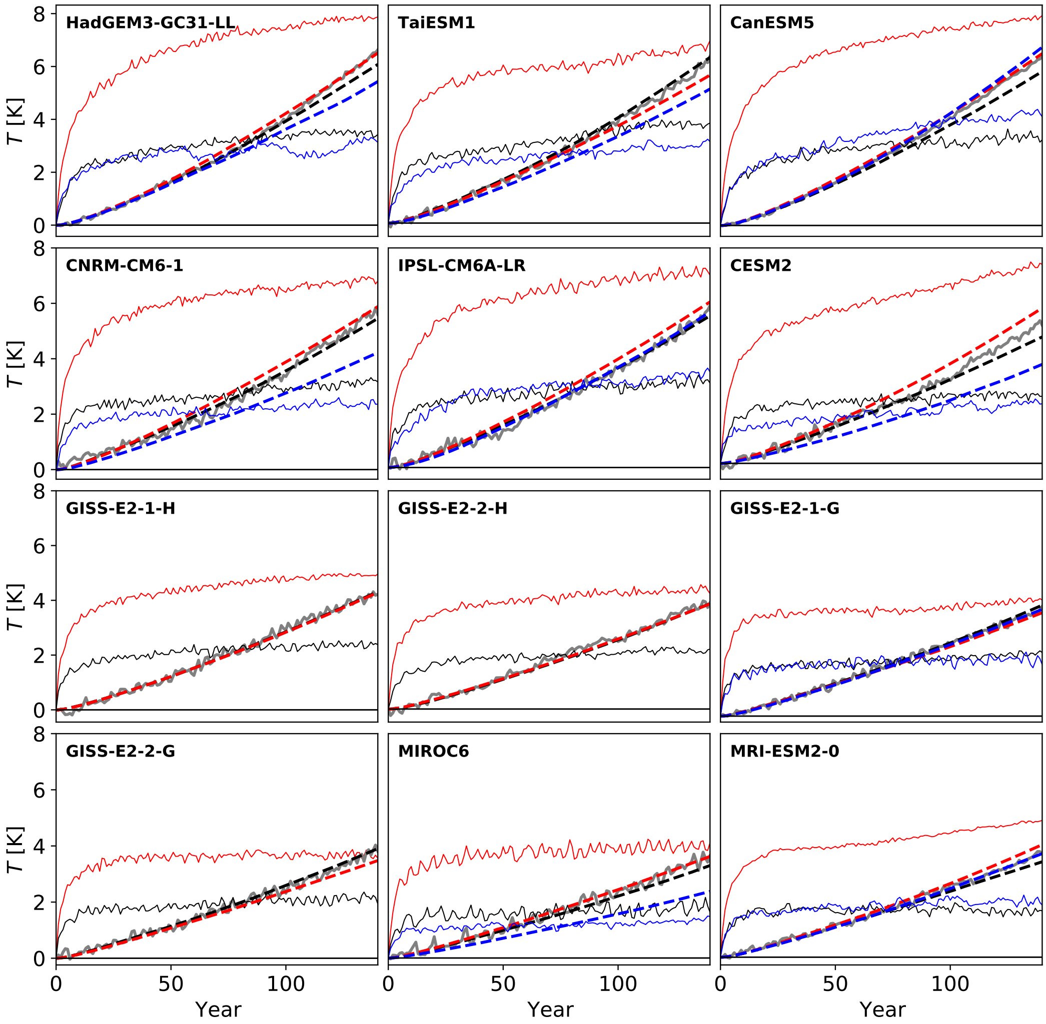

Figure 4. The dashed curves are reconstructions of the 1pctCO2 experiment (gray) using data from the abrupt-4xCO2 (red), abrupt-2xCO2 (black) and abrupt-0p5xCO2 (blue) experiments (solid curves). The forcing is assumed to scale like the superlogarithmic forcing in the reconstruction. The sign is flipped when plotting data from the abrupt-0p5xCO2 experiment. For the experiments where several members exist, we have plotted the ensemble mean.

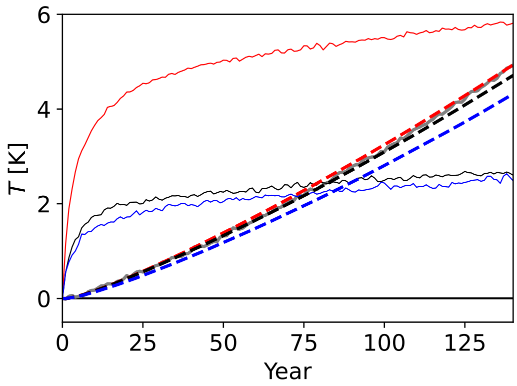

Figure 5. The model means of all curves in Figure 4.

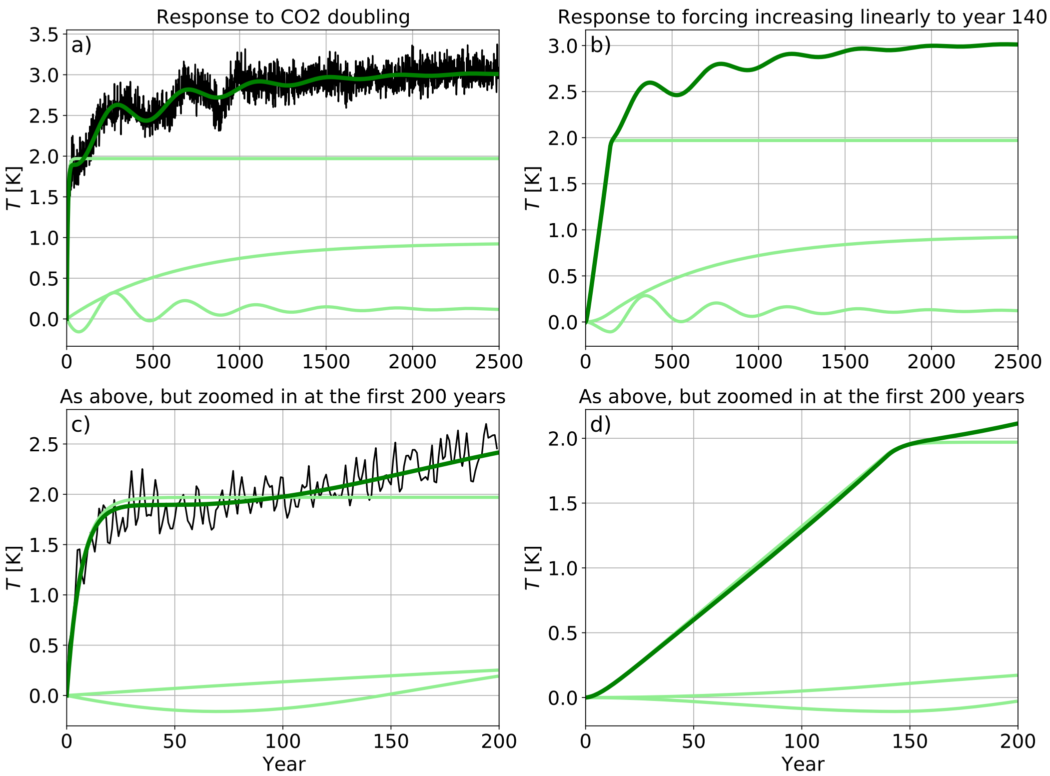

Figure 6. a) Result of fitting a two-exp and a pair of oscillatory responses to CESM104 abrupt2x. The dark green curves are the total responses to either an abrupt doubling of CO_2 (left) or a forcing increasing linearly to doubling of CO_2 in year 140, and is thereafter kept constant (right). The light green curves are components of the total response: Two exponential responses with time scales of approximately 7 and 639 years, and one oscillatory response with a period of approximately 410 years and damping time scale of 619 years.

6.3 Comparing different response functions

We fit two-exp, three-exp and two-exp + oscillatory response for all CMIP6 models. The resulting root mean squared error (RMSE) of these fits are summarised in Tables S6 and S7 for abrupt-4xCO2, Table S8 for abrupt-2xCO2 and Table S9 for abrupt-0p5xCO2. The results for LongRunMIP experiments are listed in Table S10. As expected, RMSE is always smaller or unchanged for the three-exp model compared to the two-exp model. With an ideal estimation method, the two-exp + osc. should be reduced to a three-exp (by setting $q = 0$ and $S_2 = 0$) if the oscillatory solution is not a better description than the three-exp. Hence all results here with increased RMSE are just the results of not finding the optimal parameters. However, for the models where we estimate higher RMSE values for the two-exp + osc, this model is very unlikely to be a good description. Going further, we will therefore just focus on the models where adding oscillations provides a better description.

Including oscillations provides a smaller RMSE compared to the three-exp model for 11/22 LongRunMIP abrupt experiments. For most of these experiments, the improvement is very minor, and probably not worth the additional parameters. However, for one of these simulations an oscillatory response provides a visually significant better description: the CESM104 abrupt2x, shown in Figure 6 a). This experiment is also studied in further detail in Section 7.1 and Figure 8.

In Figure 6 b) and d) we estimate the temperature response to a forcing that increases linearly until doubling (in year 140), and is then kept constant thereafter. This will be approximately half the output of 1pctCO2 experiments, and demonstrates that with this linear oscillatory model, the oscillations cannot be seen during the 140 years with linear forcing. The negative response of the oscillatory part is to a large degree cancelled out by the slow exponential part, and the majority of the temperature response is described by the fastest exponential response.

42/71 runs for CMIP6 abrupt-4xCO2 have smaller RMSE if including oscillations (note that we count different members from the same model). Also for these models, most improvements are so minor that we cannot really argue that the extra parameters are needed. Despite large estimation uncertainties for these shorter runs, we find indications that there may be oscillations in many models. In the following, we highlight results for members from the 8 models where we have the largest improvements in RMSE for abrupt-4xCO2: ACCESS-CM2, GISS-E2-1-G, ICON-ESM-LR, KIOST-ESM, MRI-ESM2-0, NorESM2-LM, SAM0-UNICON, TaiESM1. We note the generally close resemblance between these runs (see Figure 7) and the first 150 years of the CESM104 abrupt2x run in Figure 6 c).

The two-exp and oscillatory fits in Figure 7 show that the oscillatory component can take various shapes. For most members (e.g. TaiESM1 r1i1p1f1), the best fit includes an oscillatory component that resembles the purely exponential components, but where the initial warming overshoots before stabilizing at a lower equilibrium temperature. In these cases the estimated oscillations have a quick damping time scale (τ_p), typically 20-30 years. For MRI-ESM2-0 members r7 and r10 we have instead an oscillation starting with an initial cooling, which is part of a slow oscillation that could develop as in the CESM104 abrupt2x run. When including this slow oscillation, we find only shorter time scales (annual and decadal) for the two purely exponential parts. For the members where the oscillation has a shorter period, we have a centennial-scale purely exponential part to explain the slow variations in the temperature. Since we know from longer runs that a centennial-millennial scale exponential component is necessary to explain the full path to equilibrium, the fits for MRI-ESM2-0 members r7 and r10 are unlikely to explain the future of these experiments. This could in theory be resolved by combining the two short time-scale exponential parts to one, and allowing the second exponential part to take a long time scale instead. However, with only 150 years of data, a fit containing several components varying on centennial to millennial scales will be poorly constrained. The take-home message from this is that we cannot really tell from the global surface temperature of these short experiments if we deal with a short-period and quickly damped out oscillation or an oscillation lasting for centuries. Longer experiments are needed, but a closer look at the AMOC evolution and the spatial pattern of warming may also give some hints.

Of these 8 models, 3 models have also run abrupt-2xCO2 and abrupt-0p5xCO2 experiments. We see no clear signs of oscillations in these abrupt-0p5xCO2 runs. For GISS-E2-1-G abrupt-2xCO2 we observe a small flattening out of the temperature as for abrupt-4xCO2, for MRI-ESM2-0 abrupt-2xCO2 the temperature flattens out, and does not start to increase again. For TaiESM1 abrupt-2xCO2, the temperature behaves similarly as for abrupt-4xCO2 (although our estimated decomposition looks a bit different). Hence there are hints that the same phenomenon appears also for abrupt-2xCO2, but the responses may not be perfectly linear.

7 Oscillations and plateaus in global temperatures

7.1 Oscillation in CESM1 warming experiments

The CESM1 abrupt CO_2 responses are further investigated (Figure 8) by looking at the Northern Hemisphere (NH) and Southern Hemisphere (SH) temperatures separately (a), and by comparing with the AMOC index (b) and NH and SH sea ice areas (c). We find that the oscillations happen only in the NH, and that the abrupt2x (blue) NH temper-

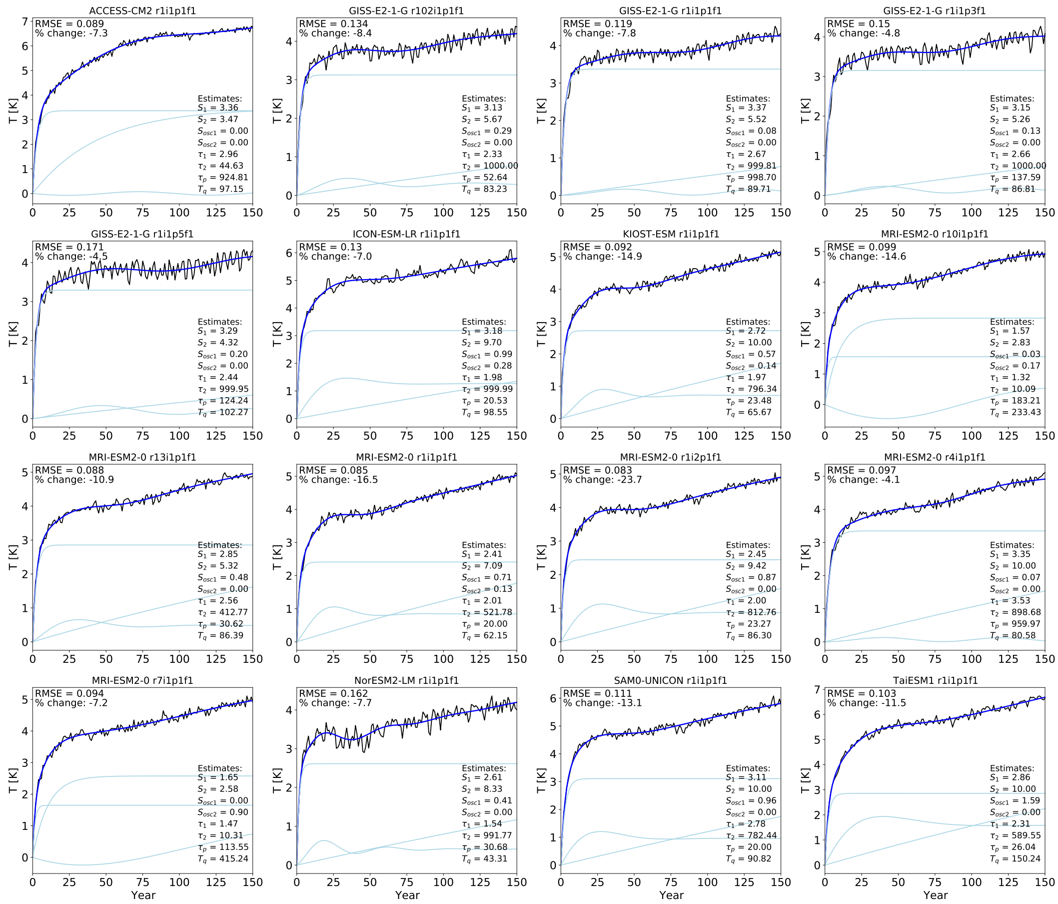

Figure 7. The two-exponential + oscillatory model fits (blue curves) for 16 different abrupt-4xCO2 runs (black curves). The light blue curves show the decomposition of the blue curve into two exponential components and one oscillatory component. The estimated parameters are listed in the figures, and the % change refers to the improvement in RMSE from three-exponential fit to the two-exponential + oscillatory model fit.

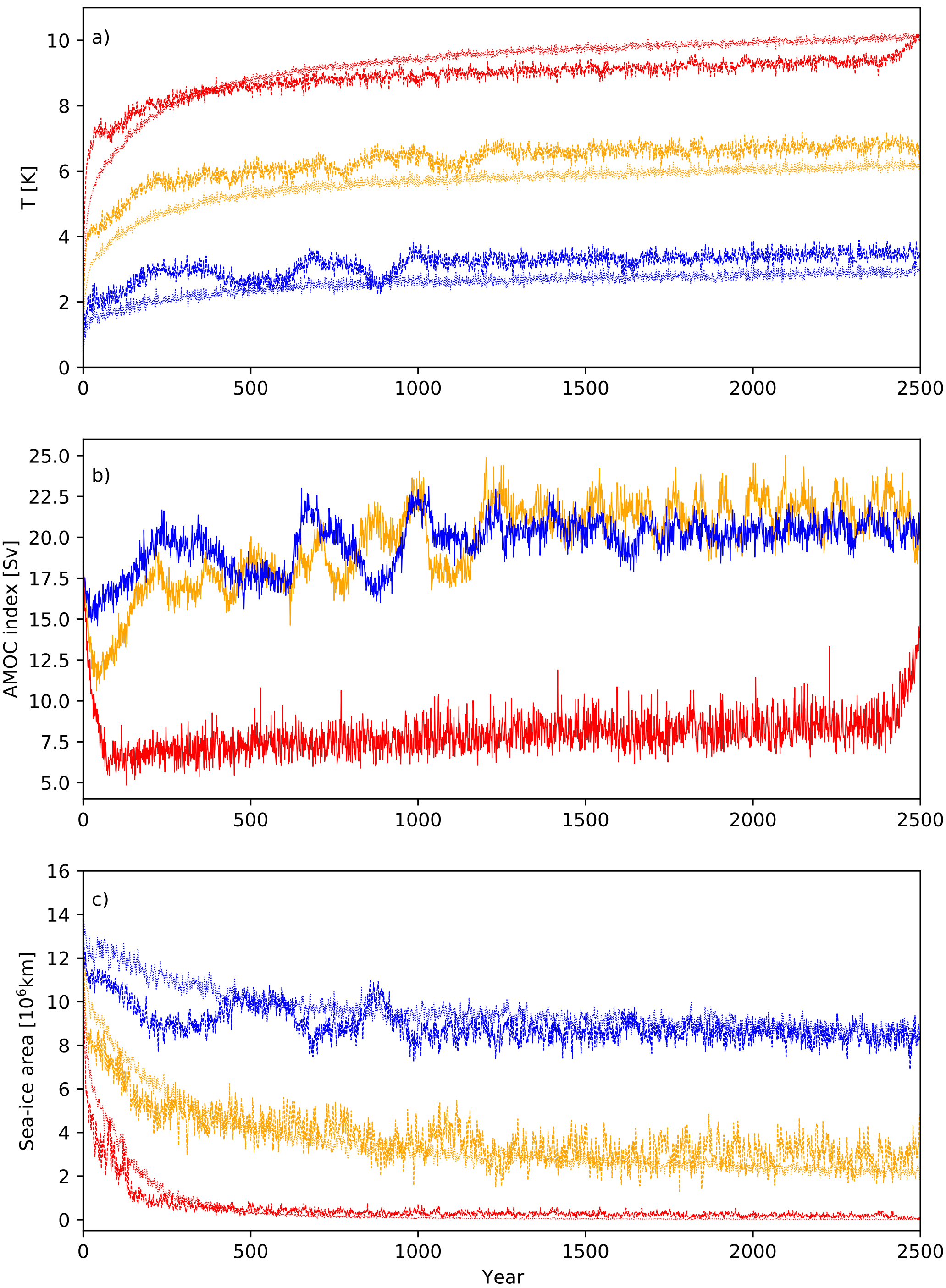

Figure 8. Mean surface temperature (a), AMOC index (b) and sea-ice area (c) for CESM104 abrupt2x (blue), abrupt4x (orange) and abrupt 8x (red). In a) and c), dashed curves are means over the Northern Hemisphere, and dotted (thinner) curves are means over the Southern Hemisphere.

ature is strongly correlated with the AMOC index (R = 0.796) and anticorrelated with the NH sea ice area (R = -0.919) if using all 2500 annual values for computation. If looking only at the first decades after the abrupt CO_2 doubling, we observe an anticorrelation between temperatures (which increase) and AMOC (which weakens). A plausible mechanism for this is that the strong initial warming inhibits the sinking of water in the North Atlantic by reducing its density. On longer time scales, AMOC changes also impact temperatures, by bringing more/less warm water northwards, which could explain the positive correlation.

The comparison with the abrupt 4x (orange) and 8x (red) simulations from the same model shows that all NH temperatures have a small plateau for some decades after the initial temperature increase, likely connected to their initial decrease in AMOC strength and sea-ice area. There are also some long-term variations later on in these experiments, but not following a similar oscillatory behaviour as the 2x experiment. We note for instance that the abrupt change around year 2500 in the abrupt8x experiment is strongly connected to an AMOC recovery. Hence, while linear response models estimated from the abrupt2x simulation may well describe the long-term responses to these other abrupt CO_2 experiments, the oscillatory behavior does not transfer to the same degree. In lack of more simulations with weaker forcing from this model, it is difficult to judge if the oscillatory phenomenon really is part of a linear model that can only be used for weaker forcings, or if it is a nonlinear effect or a random fluctuation.

7.2 Oscillation in cooling HadGEM experiment

Among models with abrupt-0p5xCO2 experiments, we find one (HadGEM-GC31-LL) with an interesting oscillation. This oscillation appears to have an increasing amplitude (see Figure 9 a)). To fit our model to these data, we need to allow the oscillatory part of the solution to have a positive real part eigenvalue, such that we get unstable/growing oscillations. This corresponds to a negative damping time scale τ_p. In b) we note that the oscillation appears mainly in the Southern Hemisphere, and is tightly connected to oscillations in the SH sea-ice extent. The Northern Hemisphere temperature is only slightly influenced by the oscillation, possibly through the atmosphere or because AMOC couples it to the SH. AMOC data are not provided for this experiment, but temperature changes in the North Atlantic (not shown) indicate that AMOC is changing. The estimated parameters are listed in the figure, and shows also that we have allowed negative values of S_{osc1} and S_{osc2}. The physical interpretation of this is that the SH sea ice actually decreases on average in extent, hence contributing to a warming on an otherwise cooling globe.

This oscillation seems to have a different physical origin than the oscillations/plateaus we observe in warming experiments. Similar changes in the SH were observed in the piControl experiment of this model (Ridley et al., 2022). In the piControl the deeper ocean has not yet reached an equilibrium state and the drifting temperatures eventually cause the water column in the Weddell and Ross seas to become unstable, and start to convect up warmer deeper ocean water that melts the sea ice. We suspect the oscillations in the abrupt-0p5xCO2 experiment is a similar phenomenon, except that in this run the cooling of the atmosphere and ocean surface layer brings the ocean column in the southern oceans faster into an unstable state. The more the surface is cooling, the larger the area can become where this instability and melting of sea ice happens, which can explain the growing oscillation and overall reduced sea ice cover.

7.3 Multidecadal pauses in global temperature increase

In Fig. 7 it can observed that the abrupt-4xCO2 simulations for several models (e.g., GISS-E2.1-G, MRI-ESM2.0, SAM0-UNICON) exhibit a plateau in their global mean surface temperature evolution after the initial fast-paced increase. This happens typically between years 30 and 70 and after year 70 the temperature starts increasing again. Averaging the temperature separately over northern and southern hemisphere (NH and SH,

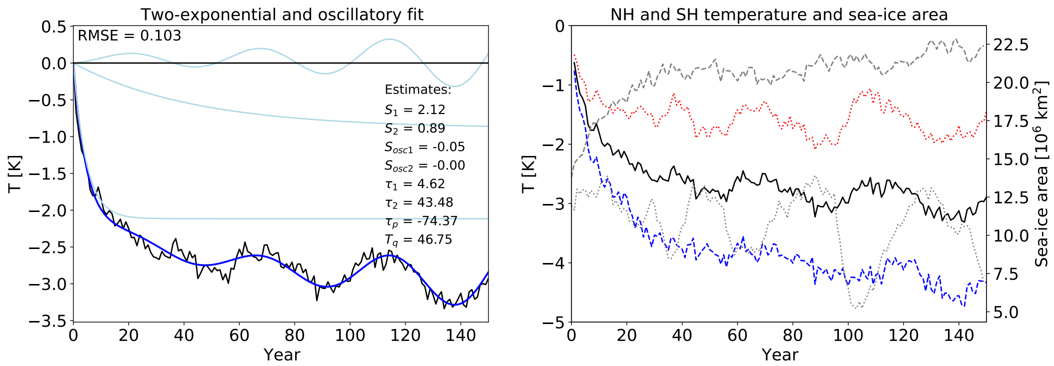

Figure 9. Results from HadGEM-GC31-LL abrupt-0p5xCO2 r1i1p1f3, where allowing an unstable (growing) oscillation makes a good fit. a) The black curve is the global surface air temperature change relative to piControl, the thick blue curve is the fitted model consisting of two exponential components (slowly varying light blue curves) and one oscillatory pair (plotted together as the oscillating light blue curve). Note that to make the fit the signs were flipped, such that the listed parameters $S_1, S_2, S_{osc1}, S_{osc2}$ are consistent with a positive response. b) The global temperature response (black) split up in Northern Hemisphere (NH, dashed blue) temperature and Southern Hemisphere (SH, dotted red) temperature. On the right axis we have the sea-ice area, which is plotted for the SH (dotted gray) and NH (dashed gray).

respectively; see Fig. 10 for the example of GISS-E2.1-G) reveals that the plateau of the global mean temperature results from a plateauing or even decrease of the NH temperature while the SH temperature increases monotonically. More specifically, maps of time slices of surface warming make clear that it is the North Atlantic that cools in response to the CO_2-forcing (Fig. 10, left column). Models that do not exhibit the plateauing global mean temperature typically exhibit neither the plateauing in the NH nor the cooling (or lack of warming) in the North Atlantic (E3SM-1.0 shown as an example in Fig. 10, right column). Though there may be models where the North Atlantic cools/warms less, but not enough to cause a significant slowdown of global temperature increase.

The difference in North Atlantic temperatures between models with and without plateau is found to be concomitant with a difference in the development of AMOC and the development of Arctic sea ice (see Figure 10), consistent with earlier studies (Bellomo et al., 2021; Mitevski et al., 2021). Models with plateauing global mean temperature tend to simulate a stronger AMOC decline in response to the CO_2-forcing (e.g. GISS-E2-1-G and SAM0-UNICON) than do the models without plateau. Notably, the pre-industrial AMOC also tends to be stronger in models with plateau than in those without plateau. Furthermore, models with plateau retain more of their Arctic sea ice than models without plateau. The connection between a plateauing global temperature, weakening AMOC, and enhanced NH sea ice cover was also noted by Held et al. (2010) for the GFDL Climate Model version 2.1.

A stronger decline in AMOC is consistent with lower North Atlantic temperatures (Bellomo et al., 2021) and less sea ice melt (Yeager et al., 2015; Liu et al., 2020; Eiselt & Graversen, 2023). The AMOC constitutes a part of the poleward energy transport in the climate system that is necessary to balance the differential energy input from solar radiation. The AMOC accomplishes northward energy transport by transporting warm water from the Tropics into the Arctic increasing the ocean heat release there and thus warming the North Atlantic. A decline of the AMOC will hence lead to a cooling or at least a hampering of the warming in response to a CO_2-forcing. Growing sea ice in response to a cooling

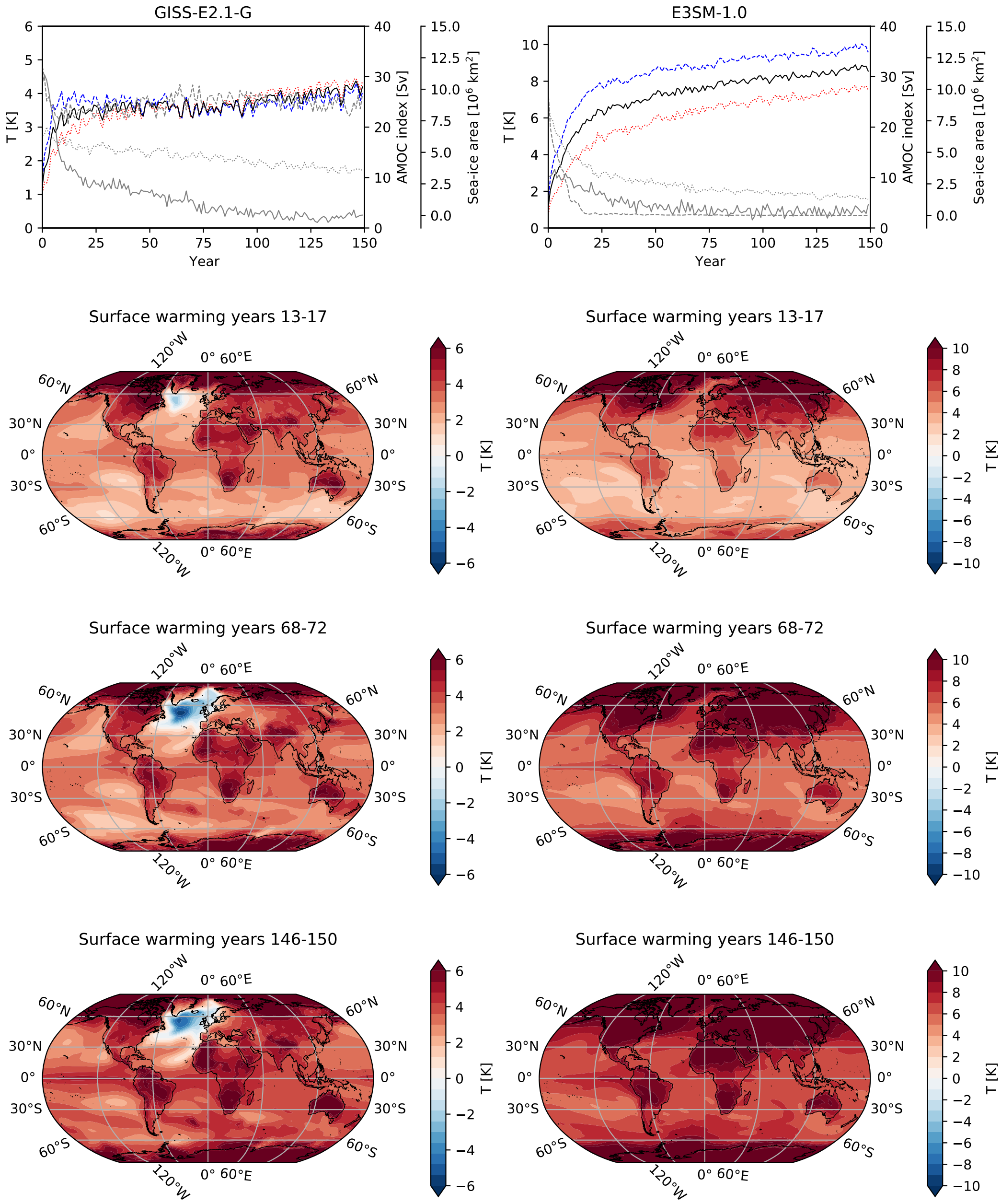

Figure 10. Example of models with and without plateaus in global temperature.

will contribute to keeping the temperature low for a while. Changes in sea ice has also been shown to affect AMOC (Sévellec et al., 2017; Liu et al., 2019; Madan et al., 2023). The growth of sea ice can therefore be an explanation for an eventual AMOC recovery, and finally lead to a decay of the oscillating component.

8 Discussion

Many earlier studies comparing different abrupt CO_2 experiments focus on experiments from single models, and are often mainly interested in the equilibrium response. Such studies find both decreasing and increasing climate sensitivities with stronger CO_2 forcing (see discussions in Meraner et al. (2013); Bloch-Johnson et al. (2021)), but the more comprehensive analysis by Bloch-Johnson et al. (2021) (including many of the same models as this paper) finds that climate sensitivity increases in most models.

Slab-ocean models are used in several studies (Colman & McAvaney, 2009; Meraner et al., 2013), and are useful tools for studying the temperature-dependence of atmospheric feedbacks. They are relatively cheap to run, and the pattern effect is somewhat suppressed in these models, partly because they go quicker to equilibrium and partly due to the lack of ocean dynamics that can change the pattern of the temperature response. This makes it easier to separate the nonlinear/temperature dependent feedbacks from the pattern effect, but ignores also possible permanent changes in feedbacks due to changes in the ocean circulation.

For a wide range of abrupt CO_2 increase experiments (1x to 8x), Mitevski et al. (2021) finds that the increase in effective climate sensitivity with increasing CO_2 is not monotonic in two fully coupled models (GISS-E2.1-G and CESM-LE), in contrast to the monotonic increase found in slab-ocean experiments (Meraner et al., 2013; Mitevski et al., 2021). The nonmonotonic increase is related to the decreasing temperatures in the North Atlantic and the weakening AMOC. For small enough abrupt CO_2 concentration increases (up to 2x and 3x CO_2 for GISS-E2.1-G and CESM-LE, respectively) the AMOC recovers after the initial decrease, while for higher concentrations it does not. For higher concentrations, the North Atlantic cools less however, because of the increased warming from CO_2.

Manabe and Stouffer (1993, 1994) also focused on studying the thermohaline circulation in the Atlantic Ocean in different abrupt CO_2 experiments. In their 2x and 4x experiments they observe a weakening of the thermohaline circulation. The circulation recovered again for $2xCO_2$, but remained weak for $4xCO_2$. For $0.5xCO_2$ Stouffer and Manabe (2003) finds a weak and shallow thermohaline circulation in the Atlantic.

The collapse of AMOC above a certain CO_2 level is an example of how a change in the ocean circulation can cause a nonlinear global temperature response. A change in circulation changes the surface temperature pattern, which further modulates which atmospheric feedbacks are triggered. In the case of a permanent collapse of AMOC, the new pattern and associated feedbacks are also permanently changed. In general, any change in effectiveness of deeper ocean heat uptake can depend on state, and therefore result in a nonlinear response. A warming of the surface can lead to a more stratified ocean with reduced vertical mixing. To some extent, however, the reduced heat uptake can still be approximated as a linear function of the surface temperature increase. We have also demonstrated the opposite effect here, that a cooling of the surface can lead to a linear oscillating response, as a result of ocean-sea ice dynamics in the Southern Ocean.

Linear response models can take many forms. Examples of physically motivated models are the upwelling-diffusion models (Hoffert et al., 1980) used in the First IPCC report, and the temperature component of the FaIR emulator (Millar et al., 2017; Smith et al., 2018; Leach et al., 2021) used in AR6 (P. Forster et al., 2021). They are powerful tools for e.g. the IPCC reports since they can be used to quickly explore a wider range of forcing scenarios than that simulated by coupled models. We suggest that a generalised box model is easier to interpret, test and generalise than box models using an efficacy factor, since temperature components and different feedback parameters are more directly associated with the pattern of surface temperature evolution, instead of being indirectly associated through an efficacy factor. We do not have to assume anything about the distribution of the boxes as long as we are interested in global quantities, but in order to better constrain the values of the different feedback parameters, the additional information about the pattern can be useful.

9 Conclusions

We find that linear response is overall a good assumption for global surface temperatures. However, good predictions with linear response models are crucially dependent on good forcing estimates. Distinguishing between forcing and response is a challenge, and the uncertainty of forcing estimates is the main limitation to determining if a model has a linear response or not.

Mitevski et al. (2022) and Geoffroy and Saint-Martin (2020) highlight the importance of taking into account the nonlogarithmic dependence of the forcing on the CO_2 concentration. This implies stronger forcing for each CO_2 doubling, also consistent with recent findings of (He et al., 2023). He et al. (2023) finds that the stratospheric temperature impacts CO_2 forcing, and that other forcing agents affecting the stratospheric temperature therefore can modulate the CO_2 forcing. Such nonlinear interaction between forcing agents should be studied in further detail, as this deviates from a linear framework. We hope also the effort initiated by RFMIP (Pincus et al., 2016) to better constrain forcing estimates will be continued for more models and experiments in the future.

For models with a plateau in the global temperature response to an abrupt increase in CO_2 stemming from a cooling of the North Atlantic, the cooling component (which can be modelled with an oscillatory part) can counteract the warming from the slow centennial-millennial scale component for a long time. For these models, a response model with a single exponential response can actually be sufficient for many short-term prediction purposes. In CESM104 abrupt2x a single exponential explains the majority of the first decades after abrupt doubling of CO_2, and for all 140 years with linearly increasing forcing.

Parameter estimation taking into account the possibility for centennial-scale oscillations is difficult for short time series, like the typical 150 year abrupt CO_2 experiments. We encourage more models to run longer abrupt CO_2 experiments, also for different levels of CO_2. Longer runs will help constrain linear response models better on the longer term, which can then further be used to quickly predict a wide range of other forcing scenarios. In particular, more and longer abrupt-2xCO2 would be useful, since these are very likely to be within the range where a linear response is a good approximation. Linear responses estimated from abrupt-4xCO2 are also quite good approximations, but there are some signs of nonlinear responses playing a role in these experiments (Fredriksen et al., 2023; Bloch-Johnson et al., 2021). CMIP6 abrupt-4xCO2 warms on average 2.2 times abrupt-2xCO2, and we estimate that about a factor 2 can be attributed to the forcing difference. The remaining 10% extra warming in abrupt-4xCO2 is likely attributed to nonlinear responses, such as feedback changes (Bloch-Johnson et al., 2021).

References

Andrews, T., Gregory, J. M., & Webb, M. J. (2015). The Dependence of Radiative Forcing and Feedback on Evolving Patterns of Surface Temperature Change in Climate Models. *Journal of Climate*, *28*(4), 1630–1648. doi: 10.1175/JCLI-D-14-00545.1

Andrews, T., Smith, C. J., Myhre, G., Forster, P. M., Chadwick, R., & Ackerley, D. (2021). Effective Radiative Forcing in a GCM With Fixed Surface Temperatures. *Journal of Geophysical Research: Atmospheres*, *126*, e2020JD033880. doi: 10.1029/2020JD033880

Armour, K. C., Bitz, C. M., & Roe, G. H. (2013). Time-Varying Climate Sensitivity from Regional Feedbacks. *Journal of Climate*, *26*, 4518–4534. doi: 10.1175/JCLI-D-12-00544.1

Bellomo, K., Angeloni, M., Corti, S., & von Hardenberg, J. (2021). Future climate change shaped by inter-model differences in Atlantic meridional overturning circulation response. *Nature Communications*, *12*, 3659. doi: 10.1038/s41467-021-24015-w

Bloch-Johnson, J., Pierrehumbert, R. T., & Abbot, D. S. (2015). Feedback temperature dependence determines the risk of high warming. *Geophysical Research Letters*, *42*(12), 4973-4980. doi: 10.1002/2015GL064240

Bloch-Johnson, J., Rugenstein, M., Stolpe, M. B., Rohrschneider, T., Zheng, Y., & Gregory, J. M. (2021). Climate Sensitivity Increases Under Higher CO_2 Levels Due to Feedback Temperature Dependence. *Geophysical Research Letters*, *48*, e2020GL089074. doi: 10.1029/2020GL089074

Caldeira, K., & Myhrvold, N. P. (2013). Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration. *Environmental Research Letters*, *8*(3), 034039. doi: 10.1088/1748-9326/8/3/034039

Colman, R., & McAvaney, B. (2009). Climate feedbacks under a very broad range of forcing. *Geophysical Research Letters*, *36*(1). doi: 10.1029/2008GL036268

Cummins, D. P., Stephenson, D. B., & Stott, P. A. (2020). Optimal Estimation of Stochastic Energy Balance Model Parameters. *Journal of Climate*, *33*(18), 7909–7926. doi: 10.1175/JCLI-D-19-0589.1

Edwards, C., & Penney, D. (2007). *Differential equations and boundary value problems: Computing and modelling (Fourth edition)*. Pearson.

Eiselt, K.-U., & Graversen, R. G. (2023). On the Control of Northern Hemispheric Feedbacks by AMOC: Evidence from CMIP and Slab Ocean Modeling. *Journal of Climate*, *36*(19), 6777–6795. doi: 10.1175/JCLI-D-22-0884.1

Etminan, M., Myhre, G., Highwood, E. J., & Shine, K. P. (2016). Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing. *Geophysical Research Letters*, *43*(24), 12,614–12,623. doi: 10.1002/2016GL071930

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., ... Zhang, H. (2021). The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity [Book Section]. In V. Masson-Delmotte et al. (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (p. 923–1054). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: 10.1017/9781009157896.009

Forster, P. M., Richardson, T., Maycock, A. C., Smith, C. J., Samset, B. H., Myhre, G., ... Schulz, M. (2016). Recommendations for diagnosing effective radiative forcing from climate models for CMIP6. *Journal of Geophysical Research: Atmospheres*, *121*(20), 12,460–12,475. doi: 10.1002/2016JD025320

Fredriksen, H.-B., Rugenstein, M., & Graversen, R. (2021). Estimating Radiative Forcing With a Nonconstant Feedback Parameter and Linear Response. *Journal of Geophysical Research: Atmospheres*, *126*(24), e2020JD034145. doi: 10.1029/2020JD034145

Fredriksen, H.-B., & Rypdal, M. (2017). Long-range persistence in global surface temperatures explained by linear multibox energy balance models. *Journal of Climate*, *30*, 7157–7168. doi: 10.1175/JCLI-D-16-0877.1

Fredriksen, H.-B., Smith, C. J., Modak, A., & Rugenstein, M. (2023). 21st Century Scenario Forcing Increases More for CMIP6 Than CMIP5 Models. *Geophysical Research Letters*, *50*(6), e2023GL102916. doi: 10.1029/2023GL102916

Geoffroy, O., & Saint-Martin, D. (2020). Equilibrium- and Transient-State Dependencies of Climate Sensitivity: Are They Important for Climate Projections? *Journal of Climate*, *33*(5), 1863 – 1879. doi: 10.1175/JCLI-D-19-0248.1

Geoffroy, O., Saint-Martin, D., Bellon, G., Voldoire, A., Olivié, D., & Tytéca, S. (2013). 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*(6), 1859–1876. doi: 10.1175/JCLI-D-12-00196.1

Geoffroy, O., Saint-Martin, D., Olivié, D. J. L., Voldoire, A., Bellon, G., & Tytéca, S. (2013). Transient Climate Response in a Two-Layer Energy-Balance Model. Part I: Analytical Solution and Parameter Calibration Using CMIP5 AOGCM Experiments. *Journal of Climate*, *26*, 1841–1857. doi: 10.1175/JCLI-D-12-00195.1

Good, P., Andrews, T., Chadwick, R., Dufresne, J.-L., Gregory, J. M., Lowe, J. A., … Shiogama, H. (2016). nonlinMIP contribution to CMIP6: model intercomparison project for non-linear mechanisms: physical basis, experimental design and analysis principles (v1.0). *Geoscientific Model Development*, *9*(11), 4019–4028. doi: 10.5194/gmd-9-4019-2016

Good, P., Gregory, J. M., & Lowe, J. A. (2011). A step-response simple climate model to reconstruct and interpret AOGCM projections. *Geophysical Research Letters*, *38*, L01703. doi: 10.1029/2010GL045208

Good, P., Gregory, J. M., Lowe, J. A., & Andrews, T. (2013). Abrupt CO2 experiments as tools for predicting and understanding CMIP5 representative concentration pathway projections. *Climate Dynamics*, *40*(3), 1041–1053. doi: 10.1007/s00382-012-1410-4

Gregory, J. M., Andrews, T., & Good, P. (2015). The inconstancy of the transient climate response parameter under increasing CO_2. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*, 20140417. doi: 10.1098/rsta.2014.0417

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., … Williams, K. D. (2004). A new method for diagnosing radiative forcing and climate sensitivity. *Geophysical Research Letters*, *31*, L03205. doi: 10.1029/2003GL018747

Hansen, J., Sato, M., Ruedy, R., Nazarenko, L., Lacis, A., Schmidt, G. A., … Zhang, S. (2005). Efficacy of climate forcings. *Journal of Geophysical Research: Atmospheres*, *110*(D18). doi: 10.1029/2005JD005776

Hasselmann, K., Sausen, R., Maier-Reimer, E., & Voss, R. (1993). On the cold start problem in transient simulations with coupled atmosphere-ocean models. *Climate Dynamics*, *9*(6), 53–61. doi: 10.1007/BF00210008

He, H., Kramer, R. J., Soden, B. J., & Jeevanjee, N. (2023). State dependence of CO2 forcing and its implications for climate sensitivity. *Science*, *382*(6674), 1051–1056. doi: 10.1126/science.abq6872

Held, I., Winton, M., Takahashi, K., Delworth, T. L., Zeng, F., & Vallis, G. (2010). Probing the Fast and Slow Components of Global Warming by Returning Abruptly to Preindustrial Forcing. *Journal of Climate*, *23*, 2418 – 2427. doi: 10.1175/2009JCLI3466.1

Hoffert, M. I., Callegari, A. J., & Hsieh, C.-T. (1980). The role of deep sea heat storage in the secular response to climatic forcing. *Journal of Geophysical Research: Oceans*, *85*, 6667-6679. doi: 10.1029/JC085iC11p06667

Jackson, L. S., Maycock, A. C., Andrews, T., Fredriksen, H.-B., Smith, C. J., & Forster, P. M. (2022). Errors in Simple Climate Model Emulations of Past and Future Global Temperature Change. *Geophysical Research Letters*, *49*(15), e2022GL098808. doi: 10.1029/2022GL098808

Jiang, W., Gastineau, G., & Codron, F. (2023). Climate Response to Atlantic Meridional Energy Transport Variations. *Journal of Climate*, *36*(16), 5399 – 5416. doi: 10.1175/JCLI-D-22-0608.1

Larson, E. J. L., & Portmann, R. W. (2016). A Temporal Kernel Method to Compute Effective Radiative Forcing in CMIP5 Transient Simulations. *Journal of Climate*, *29*(4), 1497-1509. doi: 10.1175/JCLI-D-15-0577.1

Leach, N. J., Jenkins, S., Nicholls, Z., Smith, C. J., Lynch, J., Cain, M., ... Allen, M. R. (2021). FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development*, *14*(5), 3007–3036. doi: 10.5194/gmd-14-3007-2021

Lin, Y.-J., Hwang, Y.-T., Ceppi, P., & Gregory, J. M. (2019). Uncertainty in the Evolution of Climate Feedback Traced to the Strength of the Atlantic Meridional Overturning Circulation. *Geophysical Research Letters*, *46*(21), 12331-12339. doi: 10.1029/2019GL083084

Liu, W., Fedorov, A., & Sévellec, F. (2019). The Mechanisms of the Atlantic Meridional Overturning Circulation Slowdown Induced by Arctic Sea Ice Decline. *Journal of Climate*, *32*(4), 977 – 996. doi: 10.1175/JCLI-D-18-0231.1

Liu, W., Fedorov, A. V., Xie, S.-P., & Hu, S. (2020). Climate impacts of a weakened Atlantic Meridional Overturning Circulation in a warming climate. *Science Advances*, *6*(26), eaaz4876. doi: 10.1126/sciadv.aaz4876

Madan, G., Gjermundsen, A., Iversen, S. C., & LaCasce, J. H. (2023). The weakening AMOC under extreme climate change. *Climate Dynamics*. doi: 10.1007/s00382-023-06957-7

Manabe, S., & Stouffer, R. J. (1993). Century-scale effects of increased atmospheric C02 on the ocean–atmosphere system. *Nature*, *364*, 215 – 218. doi: 10.1038/364215a0

Manabe, S., & Stouffer, R. J. (1994). Multiple-Century Response of a Coupled Ocean-Atmosphere Model to an Increase of Atmospheric Carbon Dioxide. *Journal of Climate*, *7*(1). doi: 10.1175/1520-0442(1994)007⟨0005:MCROAC⟩2.0.CO;2

Meraner, K., Mauritsen, T., & Voigt, A. (2013). Robust increase in equilibrium climate sensitivity under global warming. *Geophysical Research Letters*, *40*(22), 5944–5948. doi: 10.1002/2013GL058118

Millar, R. J., Nicholls, Z. R., Friedlingstein, P., & Allen, M. R. (2017). 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*(11), 7213–7228. doi: 10.5194/acp-17-7213-2017

Mitevski, I., Orbe, C., Chemke, R., Nazarenko, L., & Polvani, L. M. (2021). Non-Monotonic Response of the Climate System to Abrupt CO2 Forcing. *Geophysical Research Letters*, *48*(6), e2020GL090861. doi: 10.1029/2020GL090861

Mitevski, I., Polvani, L. M., & Orbe, C. (2022). Asymmetric Warming/Cooling Response to CO2 Increase/Decrease Mainly Due To Non-Logarithmic Forcing, Not Feedbacks. *Geophysical Research Letters*, *49*(5), e2021GL097133. doi: 10.1029/2021GL097133

Pincus, R., Forster, P. M., & Stevens, B. (2016). The Radiative Forcing Model Intercomparison Project (RFMIP): experimental protocol for CMIP6. *Geoscientific Model Development*, *9*(9), 3447–3460. doi: 10.5194/gmd-9-3447-2016

Proistosescu, C., & Huybers, P. J. (2017). Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Sciences Advances*, *3*, e1602821. doi: 10.1126/sciadv.1602821

Richardson, T. B., Forster, P. M., Smith, C. J., Maycock, A. C., Wood, T., Andrews, T., ... Watson-Parris, D. (2019). Efficacy of Climate Forcings in PDRMIP Models. *Journal of Geophysical Research: Atmospheres*, *124*(23), 12824-12844. doi: 10.1029/2019JD030581

Ridley, J. K., Blockley, E. W., & Jones, G. S. (2022). A Change in Climate State During a Pre-Industrial Simulation of the CMIP6 Model HadGEM3 Driven by Deep Ocean Drift. *Geophysical Research Letters*, *49*(6), e2021GL097171. doi: 10.1029/2021GL097171

Rohrschneider, T., Stevens, B., & Mauritsen, T. (2019). On simple representations of the climate response to external radiative forcing. *Climate Dynamics*, *53*(5), 3131–3145. doi: 10.1007/s00382-019-04686-4

Rugenstein, M., Bloch-Johnson, J., Abe-Ouchi, A., Andrews, T., Beyerle, U., Cao, L., ... Yang, S. (2019). LongRunMIP: Motivation and Design for a Large Collection of Millennial-Length AOGCM Simulations. *Bulletin of the American Meteorological Society*, *100*(12), 2551-2570. doi: 10.1175/BAMS-D-19-0068.1

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., & Regayre, L. A. (2018). FAIR v1.3: a simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*, *11*(6), 2273–2297. doi: 10.5194/gmd-11-2273-2018

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., ... Forster, P. M. (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16), 9591–9618. doi: 10.5194/acp-20-9591-2020

Stevens, B., Sherwood, S. C., Bony, S., & Webb, M. J. (2016). Prospects for narrowing bounds on Earth's equilibrium climate sensitivity. *Earth's Future*, *4*(11), 512-522. doi: 10.1002/2016EF000376

Stouffer, R. J., & Manabe, S. (2003). Equilibrium response of thermohaline circulation to large changes in atmospheric CO2 concentration. *Climate Dynamics*, *20*, 759 – 773. doi: 10.1007/s00382-002-0302-4

Sévellec, F., Fedorov, A. V., & Liu, W. (2017). Arctic sea-ice decline weakens the Atlantic Meridional Overturning Circulation. *Nature Climate Change*, *7*, 604 – 610. doi: 10.1038/nclimate3353

Tang, T., Shindell, D., Faluvegi, G., Myhre, G., Olivié, D., Voulgarakis, A., ... Smith, C. (2019). Comparison of Effective Radiative Forcing Calculations Using Multiple Methods, Drivers, and Models. *Journal of Geophysical Research: Atmospheres*, *124*(8), 4382-4394. doi: 10.1029/2018JD030188

Winton, M., Takahashi, K., & Held, I. M. (2010). Importance of Ocean Heat Uptake Efficacy to Transient Climate Change. *Journal of Climate*, *23*(9), 2333-2344. doi: 10.1175/2009JCLI3139.1

Yeager, S. G., Karspeck, A. R., & Danabasoglu, G. (2015). Predicted slowdown in the rate of Atlantic sea ice loss. *Geophysical Research Letters*, *42*, 10704 – 10713. doi: 10.1002/2015GL065364

Zhou, C., Wang, M., Zelinka, M. D., Liu, Y., Dong, Y., & Armour, K. C. (2023). Explaining Forcing Efficacy With Pattern Effect and State Dependence. *Geophysical Research Letters*, *50*(3), e2022GL101700. doi: 10.1029/2022GL101700

10 Open Research

Code is available in github (`https://github.com/Hegebf/Testing-Linear-Responses`), and will be deployed in zenodo to get a doi when the manuscript is accepted. The CMIP6 data are available through ESGF (`https://aims2.llnl.gov/search/?project=CMIP6/`), and the processed version used here is deployed in `https://doi.org/10.5281/zenodo.7687534`. LongRunMIP data can be accessed through `https://www.longrunmip.org/`.

Acknowledgments

We thank Jeff Ridley for discussions that helped us understand the behaviour of the model HadGEM-GC31-LL. We would also like to thank everyone who contributed to producing the LongRunMIP and CMIP6 model data used in this study. The work of author Hege-Beate Fredriksen was partly funded by the European Union as part of the EPOC project (Explaining and Predicting the Ocean Conveyor). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The work of Kai-Uwe Eiselt was part of the project UiT - Climate

Initiative, Ice-ocean-atmosphere interactions in the Arctic - from the past to the future, funded by the Faculty of Science and Technology, UiT the Arctic University of Norway. Peter Good was supported by the Met Office Hadley Centre Climate Programme funded by DSIT.

Appendix A Solution of generalized box model

Here we will derive the solution of a generalized box model, based on theory from Edwards and Penney (2007).

The general box model is given by the linear system:

$$\frac{d\mathbf{T}(t)}{dt} = \mathbf{C}^{-1}\mathbf{K}\mathbf{T}(t) + \mathbf{C}^{-1}\mathbf{F}(t) \tag{A1}$$

We consider first the homogeneous problem

$$\frac{d\mathbf{T}_h(t)}{dt} = \mathbf{A}\mathbf{T}_h(t)$$

where $\mathbf{A} = \mathbf{C}^{-1}\mathbf{K}$. We note that the matrix of possible solutions (the fundamental matrix) is:

$$\mathbf{\Phi}(t) = [\mathbf{v_1}e^{\gamma_1 t} \,|\, \mathbf{v_2}e^{\gamma_2 t} \,|\, \ldots \,|\, \mathbf{v_n}e^{\gamma_n t}].$$

where $\mathbf{v_n}$ are the eigenvectors corresponding to the eigenvalues γ_n of the matrix $\mathbf{A}$. If we also set an initial condition $\mathbf{T}(0) = \mathbf{T}_0$, the homogeneous solution takes the form:

$$\mathbf{T}_h(t) = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}\mathbf{T}_0 \tag{A2}$$

An alternative notation when $\mathbf{A}$ consists of constant coefficients is the matrix exponential $e^{\mathbf{A}t} = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}$, since

$$\frac{d\mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}}{dt} = \frac{de^{\mathbf{A}t}}{dt} = \mathbf{A}e^{\mathbf{A}t} = \mathbf{A}\mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}.$$

We note that the elements of $e^{\mathbf{A}t}$ are a linear combination of elements of $\mathbf{\Phi}(t)$.

Consider the case where we have a pair of complex conjugate eigenvalues, $\gamma_1 = \overline{\gamma_2}$, $\mathbf{v_1} = \overline{\mathbf{v_2}}$. Let $\mathbf{v_2} = \mathbf{a} + i\mathbf{b}$ and $\gamma_2 = p + iq$, such that

$$\begin{aligned}
\mathbf{v_2}e^{\gamma_2 t} &= (\mathbf{a} + i\mathbf{b})e^{(p+iq)t} \\
&= (\mathbf{a} + i\mathbf{b})e^{pt}(\cos qt + i\sin qt) \\
&= e^{pt}(\mathbf{a}\cos qt - \mathbf{b}\sin qt) + ie^{pt}(\mathbf{b}\cos qt + \mathbf{a}\sin qt)
\end{aligned}$$

Then the pair of complex eigenvalue solutions can instead be given by the real and complex part of the expression above, such that:

$$\mathbf{\Phi}(t) = [e^{pt}(\mathbf{a}\cos qt - \mathbf{b}\sin qt) \,|\, e^{pt}(\mathbf{b}\cos qt + \mathbf{a}\sin qt) \,|\, \mathbf{v_3}e^{\gamma_3 t} \,|\, \ldots \,|\, \mathbf{v_n}e^{\gamma_n t}].$$

The fundamental matrix of the homogeneous problem is also used to describe the particular solution to the original nonhomogeneous system:

$$\mathbf{T}_p(t) = e^{\mathbf{A}t}\int e^{-\mathbf{A}t}\mathbf{C}^{-1}\mathbf{F}(t)dt = \int e^{\mathbf{A}(t-s)}\mathbf{C}^{-1}\mathbf{F}(s)ds.$$

We assume that the forcing vector $\mathbf{F}(t)$ is a vector of constants $\mathbf{w}$ multiplied by the global mean forcing $F(t)$. Further, we note that computing the matrix product $e^{\mathbf{A}(t-s)}\mathbf{C}^{-1}$ only results in extra constant factors to each entry of $e^{\mathbf{A}(t-s)}$, such that the resulting column vector obtained from $e^{\mathbf{A}(t-s)}\mathbf{C}^{-1}\mathbf{w}$ will therefore be a linear combination of the entries of $e^{\mathbf{A}(t-s)}$ (or $\mathbf{\Phi}(t)$).

Finally, the global mean surface temperature $T(t)$ can be described as a linear combination (area-weighted average) of the components of the vector $\mathbf{T}_p(t) + \mathbf{T}_h(t)$,

$$T(t) = G^*(t)T_0 + \int_0^t G(t-s)F(s)ds \tag{A3}$$

where

$$\begin{aligned} G(t) &= e^{pt}(c_1 \cos qt - c_2 \sin qt) + e^{pt}(c_3 \cos qt + c_4 \sin qt) + \sum_{n=3}^{K} k_n e^{\gamma_n t} && \text{(A4)} \\ &= k_1 e^{pt} \cos qt + k_2 e^{pt} \sin qt + \sum_{n=3}^{K} k_n e^{\gamma_n t} && \text{(A5)} \end{aligned}$$

and $G^*(t)$ takes the same form as $G(t)$, but has different coefficients k_n. In case of more pairs of complex solutions, we can replace more pairs from $\sum_{n=3}^{K} k_n e^{\gamma_n t}$ by oscillatory solutions of the same form as $k_1 e^{pt} \cos qt + k_2 e^{pt} \sin qt$. For the system to be stable we must require the real part of each eigenvalue to be negative. And in the case of only real negative eigenvalues, all terms including cosines and sines are dropped from $G(t)$.

If we know the full history of the system instead of setting an initial value, the solution is given by

$$T(t) = \int_{-\infty}^t G(t-s)F(s)ds \tag{A6}$$

Step-response

When studying the response to a unit-step forcing, we first decompose the response:

$$T(t) = \int_0^t G(t-s) \cdot 1 \; ds = \sum_{n=1}^{K} \int_0^t G_n(t-s)ds \tag{A7}$$

where $G_1(t) = k_1 e^{pt} \cos qt$ and $G_2(t) = k_2 e^{pt} \sin qt$ describe the damped oscillatory responses, and $G_n(t) = k_n e^{\gamma_n t}$ describe responses associated with real negative eigenvalues. For the latter, we have the temperature responses

$$T_n(t) = \int_0^t G_n(t-s)ds = \int_0^t k_n e^{\gamma_n (t-s)} ds = S_n(1 - e^{\gamma_n t}) \tag{A8}$$

where $S_n = -k_n/\gamma_n$. For $G_1(t)$, we find the step-response

$$\begin{aligned} T_1(t) = \int_0^t G_1(t-s)ds &= \int_0^t k_1 e^{p(t-s)} \cos q(t-s) \; ds \\ &= k_1 \left[\frac{e^{pt}\left(p \cos qt + q \sin qt\right) - p}{p^2 + q^2} \right] \\ &= S_{osc1} - S_{osc1} e^{pt} \cos qt + \frac{k_1 q}{p^2 + q^2} e^{pt} \sin qt \\ &= S_{osc1} \left[1 - e^{pt} \left(\cos qt - \frac{q}{p} \sin qt \right) \right] && \text{(A9)} \end{aligned}$$

where $S_{osc1} = -\frac{k_1 p}{p^2 + q^2}$, and similarly for $G_2(t)$, we find

$$\begin{aligned} T_2(t) = \int_0^t G_2(t-s)ds &= \int_0^t k_2 e^{p(t-s)} \sin q(t-s) \; ds \\ &= k_2 \left[\frac{e^{pt}\left(p \sin qt - q \cos qt\right) + q}{p^2 + q^2} \right] \\ &= S_{osc2} - S_{osc2} e^{pt} \cos qt + \frac{k_2 p}{p^2 + q^2} e^{pt} \sin qt \\ &= S_{osc2} \left[1 - e^{pt} \left(\cos qt + \frac{p}{q} \sin qt \right) \right] && \text{(A10)} \end{aligned}$$

where $S_{osc2} = \frac{k_2 q}{p^2+q^2}$. The total step-response is therefore,

$$T(t) = S_{osc1}\left[1 - e^{pt}\left(\cos qt - \frac{q}{p}\sin qt\right)\right] + S_{osc2}\left[1 - e^{pt}\left(\cos qt + \frac{p}{q}\sin qt\right)\right] + \sum_{n=3}^{K} S_n(1 - e^{\gamma_n t}) \tag{A11}$$

Finally, we note that if the forcing was stepped up to a different value than 1, this value will be a factor included in $S_{osc1}, S_{osc2}, \ldots, S_n$.

Using step-response to derive other responses

If we have estimates of the parameters $S_{osc1}, S_{osc2}, \ldots, S_n, p, q, \gamma_n$, we find that $k_1 = \frac{-S_{osc1}(p^2+q^2)}{p}$, $k_2 = \frac{S_{osc2}(p^2+q^2)}{q}$, $k_n = -S_n\gamma_n$, which we can plug into the expression for $G(t)$ and compute the response to other forcings.

Testing linearity and comparing linear response models for global surface temperatures

Hege-Beate Fredriksen[1,2], **Kai-Uwe Eiselt**[1] **and Peter Good**[3]

[1]UiT the Arctic University of Norway, Tromsø, Norway
[2]Norwegian Polar Institute, Tromsø, Norway
[3]Met Office Hadley Centre, Exeter, United Kingdom

Key Points:

- We systematically compare different abrupt CO_2 change experiments from the Coupled Model Intercomparison Project 6 and LongRunMIP archives
- Linear response is overall a good assumption, but there is some uncertainty in how forcing varies with CO_2
- We derive a linear response model that can reproduce oscillations found in some models, linked to ocean circulation and sea ice changes

Corresponding author: Hege-Beate Fredriksen, `hege.fredriksen@npolar.no`

Abstract

Global temperature responses from different abrupt CO_2 change experiments participating in Coupled Model Intercomparison Project Phase 6 (CMIP6) and LongRunMIP are systematically compared in order to study the linearity of the responses. For CMIP6 models, abrupt-4xCO2 experiments warm on average 2.2 times more than abrupt-2xCO2 experiments. A factor of about 2 can be attributed to the differences in forcing, and the rest is likely due to nonlinear responses. Abrupt-0p5xCO2 responses are weaker than abrupt-2xCO2, mostly because of weaker forcing. CMIP6 abrupt CO_2 change experiments respond linearly enough to well reconstruct responses to other experiments, such as 1pctCO2, but uncertainties in the forcing can give uncertain responses. We derive also a generalised energy balance box model that includes the possibility of having oscillations in the global temperature responses. Oscillations are found in some models, and are connected to changes in ocean circulation and sea ice. Oscillating components connected to a cooling in the North Atlantic can counteract the long-term warming for decades or centuries and cause pauses in global temperature increase.

Plain Language Summary

We compare the global surface temperature responses in climate model experiments where the CO_2 concentration is abruptly changed from preindustrial levels and thereafter held constant. A quadrupling of CO_2 is expected to result in approximately twice the response to a doubling of CO_2. The ratio varies with time, but is on average 2.2 over the first 150 years. A factor 2 can be attributed to the radiative forcing, that is, how much the energy budget changes due to the change in CO_2. The remaining increase is likely due to stronger feedbacks. Experiments with half the CO_2 level are expected to have approximately the opposite response of a doubling, but we find their responses to be weaker. The reason appears to be a weaker radiative forcing. The evolution of the global temperature with time is also affected by changes in ocean heat uptake, ocean circulation, sea ice, cloud changes, etc., and these effects may be different with a stronger warming. Changes in the ocean circulation can also lead to oscillations appearing in addition to the warming. In some models, this effect may be strong enough to pause the long-term warming for decades or centuries, before it catches up again.

1 Introduction

Linear response is assumed for global surface temperature in many papers, resulting from e.g. box models (Geoffroy, Saint-Martin, Olivié, et al., 2013; Fredriksen & Rypdal, 2017; Caldeira & Myhrvold, 2013), and used in emulators like FaIR (Millar et al., 2017; Smith et al., 2018; Leach et al., 2021). It is based on the assumption that the global temperature response is independent of the climate state, and we can think of it as a powerful first-order approximation of the temperature response to a perturbation of the top-of-atmosphere (TOA) energy budget. For strong enough responses, state-dependent mechanisms like the albedo feedback will become important, so the question is: In what range of climate states can a linear response be considered a good assumption?

With a linear/impulse response model we can emulate the response to any known forcing within a few seconds, given knowledge about how the global temperature responds to an impulse. Alternatively, we can also gain this knowledge from step responses, since these are the integral of the impulse responses. The step-responses from experiments with abrupt quadrupling of the CO_2 concentration are typically used. This experiment is one of the DECK experiments required to participate in the Coupled Model Intercomparison Project (CMIP), and is therefore widely available.

Until recently, step-experiments with other CO_2 levels have only been available for a few models. Following the requests of nonlinMIP (Good et al., 2016), several CMIP6 models now make abrupt-2xCO2 and abrupt-0p5xCO2 experiments available. In addition, various abrupt CO_2 experiments are published through LongRunMIP (Rugenstein et al., 2019). The main motivation of this paper is to investigate the linearity of the temperature response by systematically comparing these different step experiments. That is, we want to test if the impulse response function derived from abrupt doubling of CO_2 experiments is equal (within expected uncertainties) to that derived from e.g. quadrupling of CO_2. This has implications for the concept of climate sensitivity – will the response to another doubling of CO_2 be similar to the first doubling?

In addition, we will discuss commonly used linear response models, derive the solution to a generalised box model, and study how well we can reconstruct the results of experiments that gradually increase the CO_2 concentration. With the generalised box model we demonstrate also how oscillations can appear in linear response models. The negative phase of oscillatory solutions may counteract the long-term warming for several decades, and these solutions can therefore be useful tools in understanding how plateaus or oscillations can appear in the global temperature responses to a step forcing, and how it is linked to changes in the ocean circulation and sea ice.

The generalised box model is described in Section 2. In Section 3 we discuss separation of forcing and response, and the linearity of global surface temperature response in the context of modifying the forcing-feedback framework to account for the non-constancy (or implicit time-dependence (Rohrschneider et al., 2019)) of global feedbacks. A non-constant feedback parameter just due to the pattern effect (a modulation of the global feedback from different paces of warming in different regions (Armour et al., 2013; Stevens et al., 2016; Andrews et al., 2015)) can be consistent with a linear response model, while state-dependent feedbacks imply a nonlinear response model. Section 4 describes the data included in this study and Section 5 describes estimation methods. Results are presented in sections 6 and 7, followed by a discussion in Section 8 and conclusions in Section 9.

2 Different linear response models, and their physical motivation

Generally, a linear response model for a climate state variable $\Phi(t)$ responding to a forcing $F(t)$ takes the form

$$\Phi(t) = G(t) * F(t) = \int_0^t G(t-s)F(s)ds, \tag{1}$$

assuming $F(t) = 0$ for $t \leq 0$ (Hasselmann et al., 1993). $G(t)$ is the Green's function, and $*$ denotes a convolution.

For global surface temperature, this integral can be interpreted as a part of the solution of a multibox energy balance model (see Fredriksen et al. (2021) and Appendix A),

$$\mathbf{C}\frac{d\mathbf{T}(t)}{dt} = \mathbf{K}\mathbf{T}(t) + \mathbf{F}(t) \tag{2}$$

where $\mathbf{C}$ is a diagonal matrix of heat capacities of different components of the climate system, $\mathbf{K}$ is a matrix of heat exchange coefficients, $\mathbf{T}$ is a vector of temperature responses, and $\mathbf{F}$ is a forcing vector. The two-box model (e.g. Geoffroy, Saint-Martin, Olivié, et al., 2013; Geoffroy, Saint-Martin, Bellon, et al., 2013; Held et al., 2010) is a widely used example. In appendix A we derive a general solution that can be applied to any linear K-box model, and find that in the case of only negative eigenvalues γ_n in the matrix $\mathbf{C}^{-1}\mathbf{K}$,

$$G(t) = \sum_{n=1}^{K} k_n e^{\gamma_n t}. \tag{3}$$

Hasselmann et al. (1993) notes that eigenvalues can also appear in complex pairs, where k_n and γ_n from one term of the pair are complex conjugates of the other term. To our knowledge, complex eigenvalues have never been used for estimating response functions in this field before. If pairs of complex eigenvalues are present, pairs from the sum above can be replaced by damped oscillatory responses on the form $k_1 e^{pt}\cos qt + k_2 e^{pt}\sin qt$ (see Appendix A). For these solutions to be stable, the real part of the eigenvalues (p) should be negative.

The step-forcing responses for negative eigenvalue solutions take the form:

$$T(t) = \sum_{n=1}^{K} S_n(1 - e^{\gamma_n t}) \tag{4}$$

and for complex eigenvalues, pairs from this sum are replaced by pairs on the form:

$$S_{osc1}\left[1 - e^{pt}\left(\cos qt - \frac{q}{p}\sin qt\right)\right] + S_{osc2}\left[1 - e^{pt}\left(\cos qt + \frac{p}{q}\sin qt\right)\right] \tag{5}$$

In these terms, the exponentially relaxing responses are modulated by sines and cosines.

So why do we want to expand the method to allow oscillatory responses for some models? It is not given that all eigenvalues of the linear model have to be negative if we allow the matrix $\mathbf{K}$ to have asymmetric terms. Asymmetric terms could for instance explain anomalies in energy fluxes following the ocean circulation, going only in one direction between two boxes. So if for instance the Atlantic Meridional Overturning Circulation (AMOC) has a strong response, this might require complex eigenvalues in a linear model for the surface temperature. And as we show in this paper, there are indeed models showing oscillations that can be described with such an oscillatory response function.

Since there could be many configurations of the box model (with different physical interpretations) leading to the same solution, from now on we will just work with the parameters in Eqs. (3, 4, 5) and not convert these to the parameters in the original box model in Eq. (2). When doing this we only have to specify the number of boxes used, and not worry about what is the best configuration of the boxes.

3 Distinguishing between forcing and response

The temperature response $T(t) = G(t) * F(t)$ cannot alone tell us how to distinguish between what is forcing and what is response to the forcing, since we can just move a factor between G and F without changing T. This separation is often done using the linear forcing - feedback framework, expressing the global top-of-the-atmosphere radiation imbalance (N) as

$$N = F + \lambda T \tag{6}$$

where $\lambda < 0$ is the feedback parameter, T is the global temperature response and F is the radiative forcing. This tells us how we can use the additional knowledge about the time series N to distinguish between F and T. However, it is now well known that the feedback parameter is not well approximated by a constant, so several modifications to this framework have been proposed to account for this. Note that how N relates to T does not impact the mathematical structure of the temperature response (as long as it is a linear relation), only how the forcing and feedbacks should be defined.

We can distinguish between three main classes of modifications:
(1) Assuming that N is a nonlinear function of T, e.g:

$$N = F + c_1 T + c_2 T^2 \tag{7}$$

This describes how λ could change with state (temperature) (Bloch-Johnson et al., 2015, 2021). Some examples of feedbacks that are well known to depend on temperature are the ice-albedo feedback and the water vapour feedback.

(2) Decomposing the surface temperature as

$$T = \sum_{n=1}^{K} T_n \tag{8}$$

and associate a feedback parameter λ_n with each component T_n, such that:

$$N = F + \sum_{n=1}^{K} \lambda_n T_n. \tag{9}$$

This can describe the pattern effect, if assuming different regions have different feedbacks and different amplitudes of the temperature response, which modulates the global value of λ with time (Armour et al., 2013). Proistosescu and Huybers (2017); Fredriksen et al. (2021, 2023) use such a decomposition of the temperature into linear responses with different time-scales.

Extending the decomposition of N in Eq. (9) to include oscillatory components may not be straight-forward if oscillations are in fact connected to the North Atlantic temperatures and changes in AMOC. The troposphere is very stable in this region and surface temperature changes are therefore confined in the lower troposphere, and not necessarily causing much change in the TOA radiation (Eiselt & Graversen, 2023; Jiang et al., 2023). Increasing surface temperatures in such stable regions lead to increased estimates of the climate sensitivity, interpreted as a positive lapse rate feedback (Lin et al., 2019). In the framework of Eq. (9) a possibility is to ignore or put less weight on the North Atlantic temperature component, due to the weaker connection between T and N here, but this needs to be further investigated in a future paper. Related effects can also play a role, for instance can AMOC changes lead to TOA radiation changes in surrounding areas, such as through low cloud changes in the tropics (Jiang et al., 2023). Such effects are likely model dependent.

(3) Descriptions using a heat-uptake efficacy factor ε, that describe how N depends on the heat uptake in the deeper ocean exist as well. This is mathematically equivalent to the second class for global quantities (Rohrschneider et al., 2019). In this description, the sum $T = \sum_{n=1}^{K} T_n$ is not necessarily considered a decomposition of the surface temperature, but includes also components describing temperature anomalies in the deeper ocean. If these temperatures are part of a linear model, typically a two- or three- box model, N can still be expressed as in Eq. (9). As these temperature components are just linear combinations of the components in Fredriksen et al. (2021); Proistosescu and Huybers (2017), it is only a matter of choice if expressing N using the temperatures in each box, or using the components of the diagonalized system, associated with different time scales of the system.

Descriptions with heat-uptake efficacy take slightly different forms in different papers. Winton et al. (2010) describes efficacy without specifying a model for the ocean heat uptake, while Held et al. (2010); Geoffroy, Saint-Martin, Bellon, et al. (2013) include it in the two-box model:

$$c_F \frac{dT}{dt} = -\beta T - \varepsilon H + F \tag{10}$$

$$c_D \frac{dT_D}{dt} = H \tag{11}$$

where T and T_D are the temperature anomalies of the surface and deep ocean boxes, respectively, and $H = \gamma(T - T_D)$ is the heat uptake of the deep ocean. The sum of the heat uptake in both layers equals N, leading to:

$$N = F - \beta T - (\varepsilon - 1)\gamma(T - T_D) \tag{12}$$

The concept of efficacy can be considered a way of retaining a "pattern effect" in box models with only one box connected to the surface, by relating the evolving spatial pattern of surface temperature change to the oceanic heat uptake (Held et al., 2010; Geoffroy & Saint-Martin, 2020). Similarly, efficacy of forcing (Hansen et al., 2005) has also been shown to be related to a "pattern effect" (Zhou et al., 2023), since forcing in different regions can trigger different atmospheric feedbacks.

Cummins et al. (2020); Leach et al. (2021) have modified this description to use it with a 3-box model, and use the heat uptake from the middle box to the deep ocean box to modify the radiative response

$$N(t) = F(t) - \lambda T_1(t) + (1 - \varepsilon)\kappa_3[T_2(t) - T_3(t)] \tag{13}$$

If writing this equation in the form of Eq. (9), we find that the feedback parameters associated with $T_2(t)$ and $T_3(t)$ have equal magnitudes and opposite signs. This could put unfortunate constraints on parameters in this system, like net positive regional feedbacks, if interpreted as a pattern effect. We suggest avoiding this indirect description of the pattern effect with an efficacy parameter when using more than two boxes, and instead use a more direct interpretation of the parameters as describing a spatial pattern, such as Eq. (9).

3.1 Forcing defined using fixed-SST experiments

An alternative, that is not based on assumptions about the evolution of the feedbacks, is to run additional model experiments where sea-surface temperatures are kept fixed (Hansen et al., 2005; Pincus et al., 2016). These experiments aim to simulate close to 0 surface temperature change, such that $N \approx F$. Forcing estimated from these experiments have less uncertainty than regression methods based on the above-mentioned relationships between N, T and F (P. M. Forster et al., 2016), but are contaminated by land temperature responses. A forcing definition that includes all adjustements in N due to the forcing, but no adjustments due to surface temperature responses is the effective radiative forcing (ERF). This is considered the best predictor of surface temperatures, since it has forcing efficacy factors closest to 1 (Richardson et al., 2019). Ideally ERF should be estimated in models by fixing all surface temperatures, but this is technically challenging (Andrews et al., 2021). Instead, it is more common to correct the fixed-SST estimates for the land response (Richardson et al., 2019; Tang et al., 2019; Smith et al., 2020). We have not used these estimates in this paper, since they are not available for many models.

4 Choice of data

We compare abrupt-4xCO2 global temperature responses to all other abrupt CO2 experiments we can find. In the CMIP6 archive we have 12 models with abrupt-2xCO2 and 9 models with abrupt-0p5xCO2. In LongRunMIP we find 6 models with at least two different abrupt CO2 experiments, and we use the notation abruptNx to describe these, where N could be 2, 4, 6, 8 or 16. The advantage of models in LongRunMIP is that we can study responses also on millennial time scales, while for CMIP6 models the experiments are typically 150 years long.

There exist also similar comparisons of abrupt CO_2 experiments for a few other models outside of these larger data archives (e.g., Mitevski et al., 2021, 2022; Meraner et al., 2013; Rohrschneider et al., 2019). These data are not analysed in this study, but will be included in our discussion.

CMIP6 abrupt CO_2 experiments are used to reconstruct 1pctCO2 experiments, and the reconstructions are compared to the coupled model output of CMIP6 models. The reason for choosing this experiment is that the forcing is relatively well known. If assuming the forcing scales like the superlogarithimic formula of Etminan et al. (2016), it should increase slightly more than linearly until CO_2 is quadrupled, and end up at the same forcing level as the abrupt-4xCO2 experiments. The Etminan et al. (2016) forcing includes stratospheric adjustments, but not tropospheric and cloud adjustments like the ERF. However, we don't use the absolute values of this forcing, only the forcing ratios. We may also take these ratios as approximate ERF ratios if assuming the Etminan et al. (2016) forcing can be converted to ERF with a constant factor.

For other experiments, the uncertainty in forcing estimates is an even more important contribution to uncertainties in the responses. Jackson et al. (2022) test emulator responses to the Radiative Forcing Model Intercomparison Project (RFMIP) forcing for 8 models, and find large model differences in emulator performance. Using a different forcing estimation method (Fredriksen et al., 2021) for the CMIP6 models, Fredriksen et al. (2023) find a generally good emulator performance for historical and SSP experiments. An important difference between the forcing estimates is that the RFMIP forcing used by Jackson et al. (2022) is not corrected for land temperature responses, while the regression-based forcing in Fredriksen et al. (2023) is defined for no surface temperature response. The method described in Fredriksen et al. (2021, 2023) is actually designed to make forcing estimates compatible with a linear temperature response, and we therefore refer to these results for performance of linear response models for historical and future scenario forcing. However, if the linear response assumption is poor for the temperatures, this influences performance of the forcing estimation method as well. For this reason it is important to test the linear response hypothesis with idealized experiments, which is the focus of this paper.

4.1 AMOC and sea ice

In our discussion of oscillatory responses and plateaus in global temperature, we consider also AMOC and sea ice changes in the models. The AMOC index is calculated as the maximum of the meridional overturning stream function (*mstfmz* or *mstfyz* in CMIP6 and *moc* in LongRunMIP) north of 30°N in the Atlantic basin below 500 m depth.

The sea-ice area is calculated by multiplying the sea-ice concentration (*siconc* or *siconca* in CMIP6 and *sic* in LongRunMIP) with the cell area (*areacello* or *areacella*) and then summing separately over the northern and southern hemispheres.

5 Estimation

5.1 Forcing ratios for step experiments

A linear temperature response assumption predicts the response in any abrupt CO_2 experiment to be a scaled version of that of the abrupt-2xCO2 experiment, since only the forcing is different in these experiments. So when comparing abrupt CO2 experiments, they are all scaled to correspond to the abrupt-2xCO2 experiment. However, choosing the best scaling factor is challenging, since the forcing is uncertain, and it is not easy to distinguish between differences due to forcing and possible nonlinear temperature responses. Therefore, we have used three different types of scaling factors in our analysis:

1) Use the same scaling factor for all models, and assume a forcing scaling like the superlogarithmic radiative forcing (RF) formula in Etminan et al. (2016) in the CO_2 range where this formula is valid, and logarithmic forcing outside this range (just to have something in lack of a valid non-logarithmic description). The factors used are 0.478 for abrupt-4xCO2 and 0.363 for abrupt-6xCO2. A logarithmic dependence on the CO_2 concentrations results in the factors -1, 1/4 and 1/8 for the abrupt- 0p5xCO2, 8xCO2 and 16xCO2 experiments.

2) Estimate ratios by performing Gregory regressions (Gregory et al., 2004) of the first 5, 10, 20 and 30 years of the experiments.

3) Use the mean temperature ratio to the abrupt-2xCO2 experiment over the first 150 years as the scaling factor. This is not meant to be an unbiased estimate of the forcing ratio, but investigates the forcing ratios in the hypothetical case of per-

fectly linear responses. However, some degree of nonlinear response is expected e.g. from differences in feedbacks (Bloch-Johnson et al., 2021). After scaling temperature responses with this factor, it is easier to visualise how nonlinear responses affect different time scales of the response.

5.2 Reconstructing 1pctCO2 experiments

Performing an integration by parts of Eq. (1) leads to

$$T(t) = \int_0^t \frac{dF}{ds} R(t-s) ds, \tag{14}$$

where $R(t) = \int_0^t G(t-s)ds$ is the response to a unit-step forcing. Discretising this equation leads to the expression used to compute impulse responses in Good et al. (2011, 2013, 2016); Larson and Portmann (2016):

$$T_i = \sum_{j=0}^{i} \frac{\Delta F_j R_{i-j}}{\Delta F_s} \tag{15}$$

where ΔF_j are annual forcing increments, and the discretised step response R_{i-j} is a response to a general step forcing ΔF_s, and must therefore be normalised with this forcing. Further details of the derivation are provided in Fredriksen et al. (2021) Supplementary Text S2.

With Eq. (15) we can use datapoints from abrupt CO_2 experiments and knowledge of forcing to directly compute the responses to other experiments. Then we can avoid the additional uncertainty related to what model to fit and its parameter uncertainties. Fitting a box model first would smooth out internal variability from the step response function, which could be an advantage when studying responses to experiments with more variable forcing. Another advantage of box models is that the response function can be extrapolated into the future, while with Eq. (15) the length of the reconstruction is restricted by the length of the step experiment. Here we will use 140 years of data for the reconstruction of 1pctCO2 experiments, and as we will see, the reconstructed responses to 1pctCO2 experiments are already very smooth, so smoothing the response function with exponential responses should not change the results significantly, as long as the smoothed model provides a good fit to the datapoints.

To test this reconstruction, we will use CMIP6 annual anomalies from the experiments abrupt-4xCO2, abrupt-2xCO2 and abrupt-0p5xCO2. The input forcing ratio starts at 0, and increases either linearly, consistent with a logarithmic dependence on CO_2 concentration, or as a ratio scaling like the superlogarithmic formula (Etminan et al., 2016). For abrupt-4xCO2, we assume the ratio becomes 1 in year 140, the time of quadrupling, and for abrupt-2xCO2, we assume the ratio is 1 in year 70, the time of doubling. The positive 1pctCO2 forcing does not equal the negative abrupt-0p5xCO2 forcing at any time point, so we just assume the abrupt-0p5xCO2 forcing to be the negative of the abrupt-2xCO2 forcing.

5.3 Fitting response functions

We will compare estimated response models from a two-box model, three-box model, and a four-box model with one pair of complex eigenvalues. These response models consist of two or three exponential responses, or two exponential plus two damped oscillatory responses. Decomposing the response using box models may also help us gain insight into the physical reasons why a linear response model works or not.

We apply the python package lmfit to estimate the parameters of the response models. It takes in an initial parameter guess, and then searches for a solution that minimizes the least-squared errors. The final parameter estimates can be sensitive to the initial guesses,

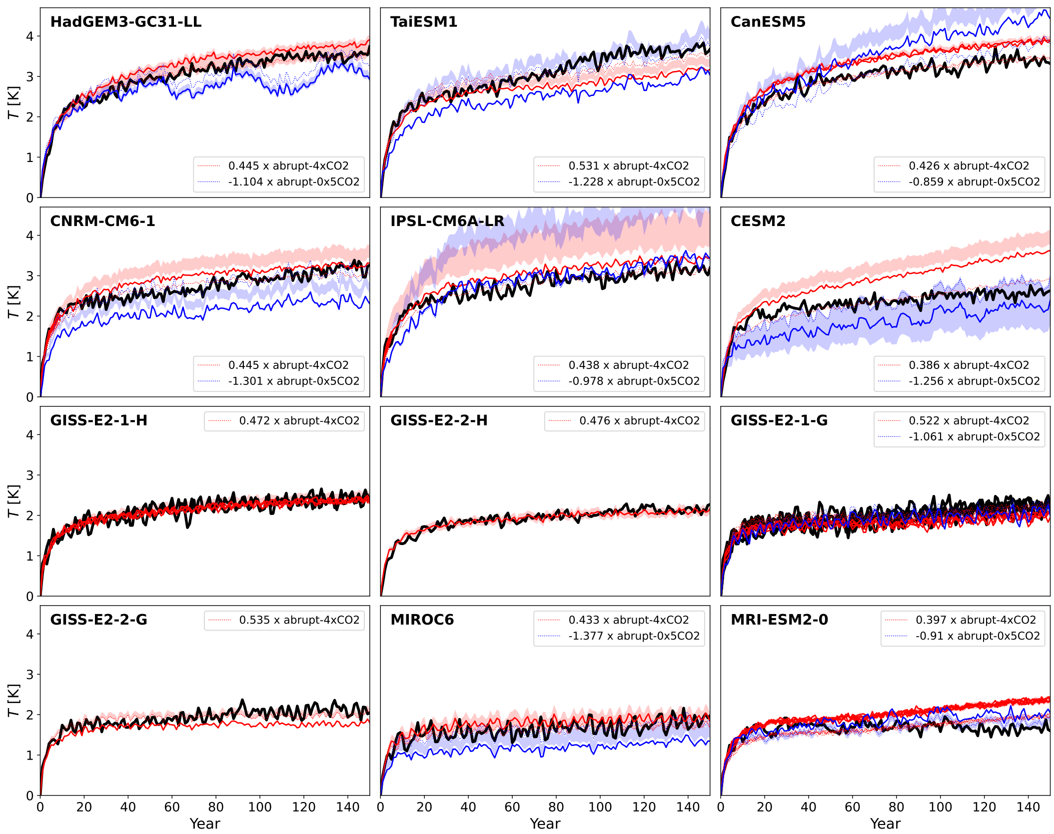

Figure 1. Comparing abrupt CO_2 experiments for CMIP6 models, where the abrupt-4xCO2 and abrupt-0p5xCO2 experiments are scaled in three different ways to correspond to the abrupt-2xCO2 experiment. Models are sorted by their abrupt-2xCO2 response in year 150. The black curves are abrupt-2xCO2 experiments, the red are scaled abrupt-4xCO2 and the blue are scaled abrupt-0p5xCO2 experiments. Solid curves use the same scaling factor for all models: 0.478 for abrupt-4xCO2 and -1 for abrupt-0p5xCO2. Thin dotted curves use the mean temperature ratio as the scaling factor (shown in legends and supplementary figure S1), and shading shows the range of the ratios of the Gregory regressions given in Supporting Tables S1 and S2.

since the optimization algorithm may just have found a local minimum. The more parameters we have in the model, the less we can trust the estimates. We see this in particular when including oscillatory responses; then we need to estimate 8 parameters, and are at risk of overfitting for the typical 150 year long experiments. As we will see, there could be different solutions containing oscillations that all provide good fits to the data. Longer time series (or some physical reasoning) would be needed in order to select the optimal fit for these records. For longer time series such as those from LongRunMIP we obtain more useful estimates.

6 Linear response results

6.1 Comparing abrupt CO_2 experiments

The curves in Figures 1 and 2 are all scaled to correspond to the abrupt-2xCO2 experiment, where the different scaling factors used illustrate the problem with the forcing uncertainty. The thick solid curves use the same scaling factor for all models (method 1), while the factors from the second and third method are model specific. The shading shows the range using the four different forcing ratios computed with Gregory regressions (method

2), that is, the minimum and maximum values from Tables S1 - S3. The thin dashed curves use the mean temperature ratios (method 3). These values are given in the subfigure legends, and shown in supporting figures S1 - S2. By definition, the black curves and the dotted red and blue curves all have the same time mean. Model specific factors can be explained by their different fast adjustments to the instantaneous radiative forcing. In addition, models can have different instantaneous forcing values, as this is shown to depend on the climatological base state (He et al., 2023). From the mean temperature ratios of the first 150 years of CMIP6 we find also that abrupt-4xCO2 warms on average 2.2 times more than abrupt-2xCO2, and abrupt-0p5xCO2 cools on average 9 % less than abrupt-2xCO2 warms (see Table S4). For LongRunMIP, abrupt4x warms 2.13 times abrupt2x when averaging all available years, or 2.18 times if averaging just the first 150 years (see Table S5, and both estimates exclude FAMOUS).

Significant differences between the curves in Figures 1 and 2 that cannot be explained by their different forcing must be explained by a nonlinear/state-dependent response. A first order assumption could be that models that warm more should tend to be more nonlinear. To investigate this we have ordered the models by their abrupt-2xCO2 response in year 150 in Figure 1 and year 500 for the longer experiments in Figure 2. We find that there are some clear differences for the warmest CMIP6 models, but also for the coldest (MRI-ESM2-0). The four different GISS models appear to be very linear.

For the two LongRunMIP models with the strongest 2xCO2 warming (CNRM-CM6-1 and FAMOUS) there are some clear differences between the curves (see Figure 2). The initial warming for CNRM-CM6-1 is halted in the 2xCO2 compared to the 4xCO2 experiment. For FAMOUS the scaling factor is particularly uncertain, and after a few centuries the pace of warming is slower in the scaled abrupt-4xCO2 experiment than in the abrupt-2xCO2 experiment. We observe only minor differences for MPI-ESM1-2, HadCM3L and CCSM3 when scaling with the mean temperature ratios. For CESM104 we observe that the abrupt2x experiment has some oscillations that are not seen in the other experiments, in addition to an abrupt change in the abrupt8x experiment.

If more warming increases the likelihood of finding nonlinear responses, we should also expect nonlinear responses to become more apparent towards the end of the simulations. We can then hypothesize that differences in forcing should explain initial differences (maybe up to a decade), and nonlinear responses explain differences at later stages. Following this, we should put more trust in the forcing scaling factors that make the initial temperature increase most similar to the abrupt-2xCO2 experiment. Which factor this is differs between models. In general, method 2 should put more emphasis on describing the first years correctly, while method 3 emphasises a good fit on all scales.

Although the individual forcing estimates are uncertain, it is a noteworthy result that the abrupt-2xCO2 regression forcing (method 2) is on average half of the abrupt-4xCO2 forcing (see Tables S1 and S3). The uncertainty of this mean is however too large to rule out that the forcing for a second CO_2 doubling is in fact larger than the first doubling, according to the findings of Etminan et al. (2016); He et al. (2023). And consistent with these expectations, for CMIP6 abrupt-0p5xCO2 we find a weaker negative forcing than logarithmic (Table S2). Our forcing ratios based on the LongRunMIP simulations for abrupt 6x, 8x and 16x CO_2 indicate that the forcing is weaker than logarithmic for higher CO_2 concentrations. Although based on very few simulations, this result is the opposite of the expectation that each CO_2 doubling produces stronger forcing (He et al., 2023).

An average forcing factor of 2 means the forcing alone is unlikely to explain the 2.2 factor difference in warming between CMIP6 abrupt-2xCO2 and abrupt-4xCO2. This conclusion is also supported by the differences in the pace of warming between abrupt-2xCO2 and abrupt-4xCO2 for several models (best visualised with the dotted curves from method 3 in Figure 1). The abrupt-4xCO2 temperatures scaled using method 2 in Figure 1 are

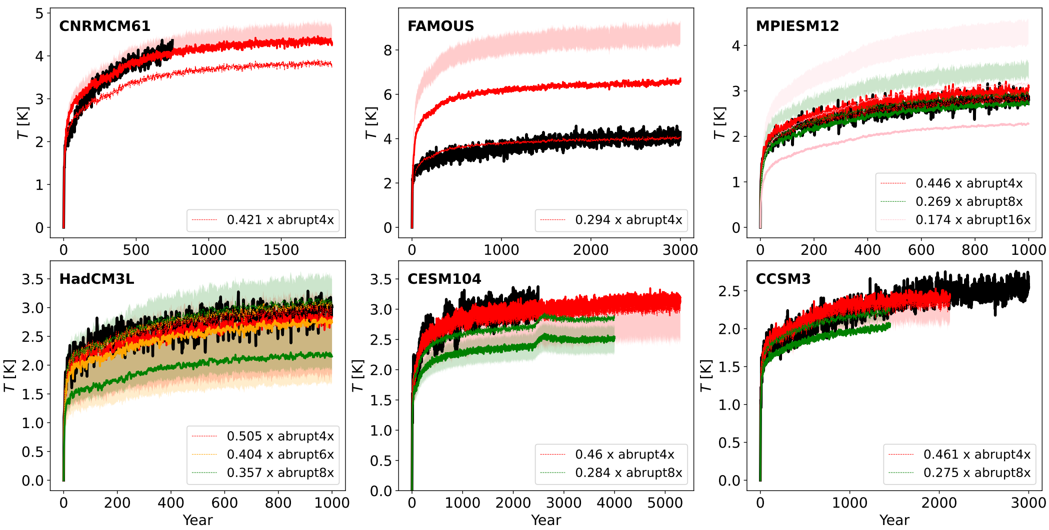

Figure 2. Comparing abrupt CO_2 experiments for LongRunMIP models. The scaling factors for the thick curves are 0.478 for 4x, 0.363 for 6x, 1/4 for 8x, 1/8 for 16x. For the thin dashed curves, the factors are computed from the mean T ratios to the first 150 years of abrupt2x, shown in Supporting figure S2, and shown in the legends here. The models are sorted by their abrupt2x temperature response in year 500. Note their different lengths and temperature scales.

on average 10 % stronger than the abrupt-2xCO2 experiments (computed from the ratio 2.2/2). The scaled abrupt-0p5xCO2 temperatures are on average 2 % stronger than the abrupt-2xCO2 temperatures (see Table S4), suggesting that the weak forcing can explain much of the weak response for abrupt-0p5xCO2. For LongRunMIP models, the average forcing ratio between 2x and 4x CO_2 reduces to 0.46 when excluding FAMOUS, making differences in the scaled temperatures over the first 150 years vanish (computed with method 2, see Table S5). For some models (CESM104 and CCSM3) the scaled temperatures deviate more from abrupt2x on millennial time scales.

Bloch-Johnson et al. (2021) suggests that feedback temperature dependence is the main reason why abrupt-4xCO2 warms more than twice the abrupt-2xCO2. This is consistent with the nonlinear responses we observe for several models. If the mean temperature ratio was a valid estimate of the forcing ratio, then in a linear framework, the same factors we found for the temperature ratios should be able to explain the ratios in top-of-atmosphere radiative imbalance. For some models this is not a good approximation (see supporting figures S1 and S2), consistent with the findings of Bloch-Johnson et al. (2021). FAMOUS has a particularly large difference in T and N ratios. Its abrupt4x warming is also so extreme that the quadratic model in Bloch-Johnson et al. (2021) suggests a runaway greenhouse effect.

6.2 Reconstructing 1pctCO2 experiments

In general, we find that both abrupt-4xCO2 experiments (see Figure 3) and abrupt-2xCO2 experiments (see Figure 4) can reconstruct the 1pctCO2 experiment very well. The largest deviation we find for the model KIOST-ESM, but we suspect the 1pctCO2 experiment from this model may have errors in the branch time information or the model setup. For many models the abrupt-0p5xCO2 experiment can also be used to make a good reconstruction, but not all (see Figure 4). For several models where abrupt-0p5xCO2 makes a poor reconstruction (TaiESM1, CNRM-CM6-1, CESM2, MIROC6), our assumptions about the forcing seems to be the limiting factor. If upscaling the negative of the abrupt-0p5xCO2 response for these models with a different factor than -1 to correspond better with the abrupt-2xCO2 experiment, we would have obtained a better reconstruction of 1pctCO2.

For many models we find that reconstructions with abrupt-4xCO2 slightly overestimates the 1pctCO2 response in the middle parts of the experiment, similar to earlier findings by Good et al. (2013); Gregory et al. (2015). In Figure 3 we compare reconstructions with a linear forcing (from logarithmic dependence on CO_2) and a forcing scaling like the superlogarithimic formula (Etminan et al., 2016). We find that reconstructions using the superlogarithmic forcing (shown in brown) explains the middle part of the 1pctCO2 experiment a little better than the logarithmic forcing (shown in red), since this forcing is slightly weaker in the middle. Even with the superlogarithmic forcing ratio, the model average reconstruction with abrupt-4xCO2 is a little overestimated in the middle part of the experiment (Figure 5). The average reconstruction with abrupt-2xCO2 explains the middle part of the experiment well, but slightly underestimates the latter part.

Which of abrupt-2xCO2 or abrupt-4xCO2 make the best reconstruction is model dependent. The 1pctCO2 experiment goes gradually to $4xCO_2$, and if there is a state-dependence involved in the response, we might expect something in between abrupt-2xCO2 and abrupt-4xCO2 responses to make the best prediction. MRI-ESM2-0 is a good example where this might be the case. For this model we observe a small underestimation with abrupt-2xCO2 and a small overestimation with abrupt-4xCO2. The reconstruction is very good with abrupt-0p5xCO2, which has an absolute response looking like an average of abrupt-2xCO2 and abrupt-4xCO2 (see Figure 1). CESM2 is also a good example where state-dependent effects are visible, since the abrupt-2xCO2 underestimates and abrupt-4xCO2 overestimates the response in the latest decades of the 1pctCO2 experiment.

For TaiESM1 and CNRM-CM6-1 the paces of warming differ a little for abrupt-2xCO2 and abrupt-4xCO2 during the middle/late stages of the experiments. Although the differences are not very significant, this is an indication of a nonlinear response. For some models (CanESM5, CNRM-CM6-1, HadGEM3-GC31-LL, IPSL-CM6A-LR, MIROC6) it is unclear if the small errors in the reconstructions are due to incorrect scaling of the forcing or nonlinear responses. The four GISS models are the most linear models, where we make good and very similar reconstructions with both abrupt-4xCO2 and abrupt-2xCO2. We observe just a small underestimation in the end of the experiment for GISS-E2-2-G abrupt-4xCO2.

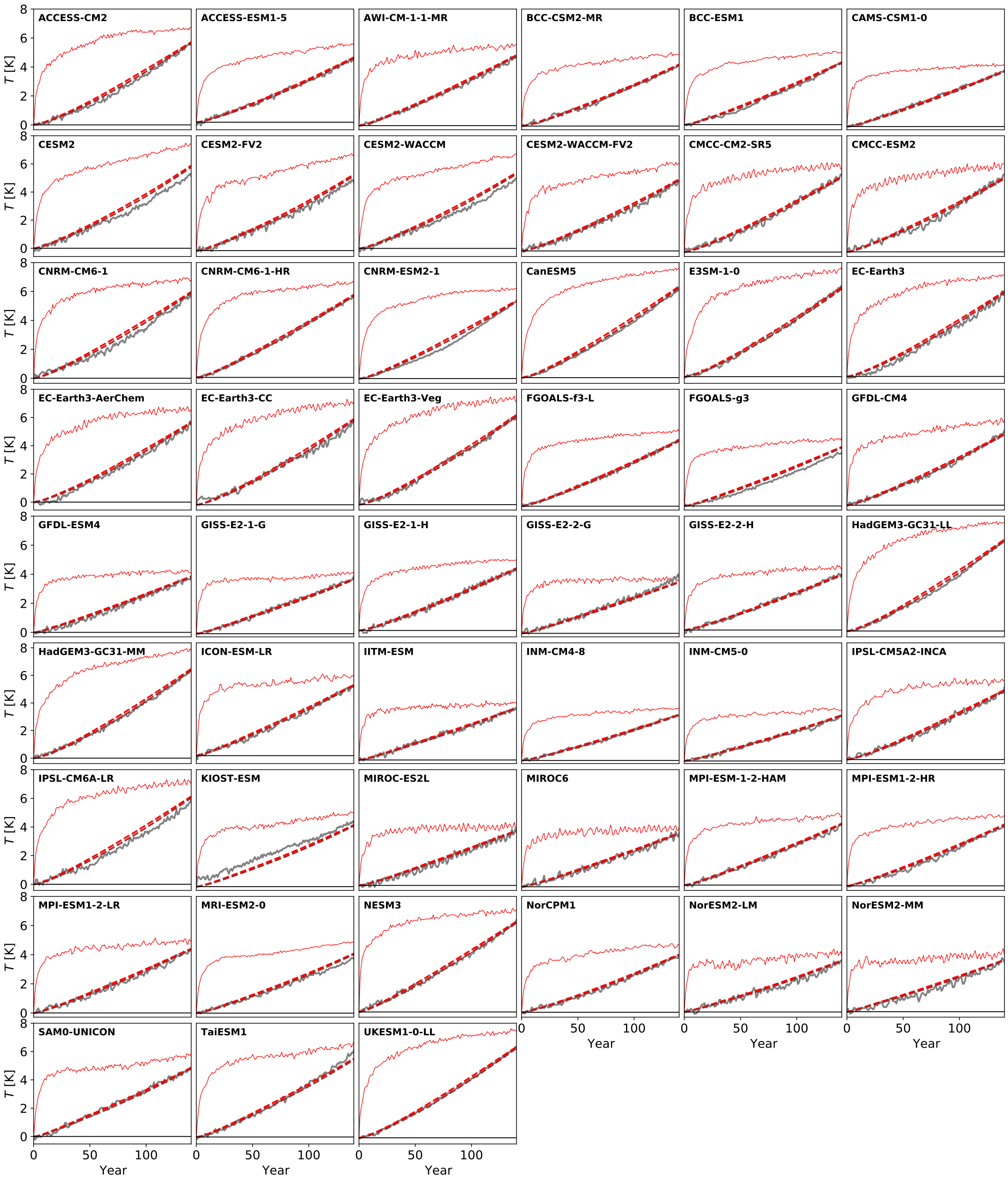

Figure 3. Red/brown dashed curves show reconstructions of the 1pctCO2 experiment (gray) using the data from the abrupt-4xCO2 experiment (red). The dashed red curve is a reconstruction based on a linearly increasing forcing, and the dashed brown curve is a reconstruction based on a forcing scaling like the superlogarithmic (Etminan et al., 2016) formula. For the experiments where several members exist, we have plotted the ensemble mean.

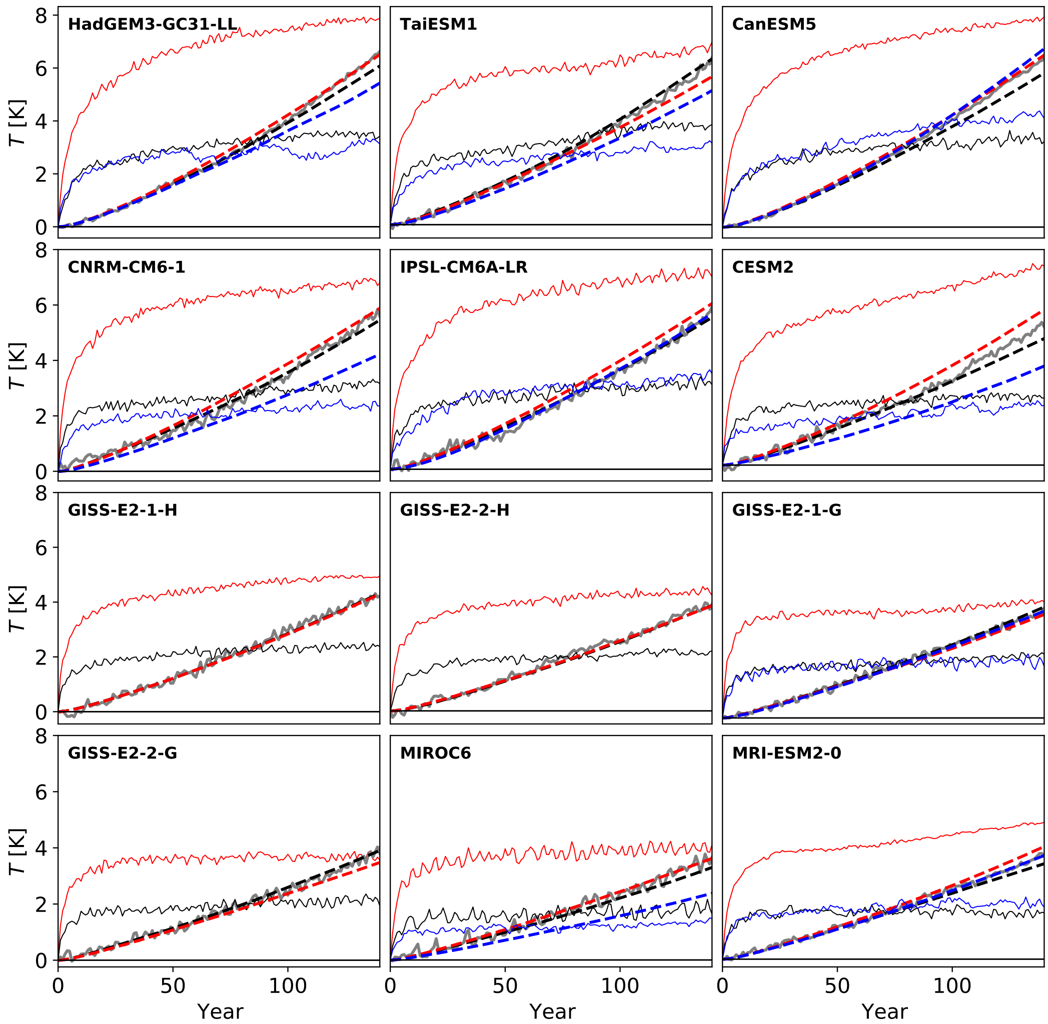

Figure 4. The dashed curves are reconstructions of the 1pctCO2 experiment (gray) using data from the abrupt-4xCO2 (red), abrupt-2xCO2 (black) and abrupt-0p5xCO2 (blue) experiments (solid curves). The forcing is assumed to scale like the superlogarithmic forcing in the reconstruction. The sign is flipped when plotting data from the abrupt-0p5xCO2 experiment. For the experiments where several members exist, we have plotted the ensemble mean.

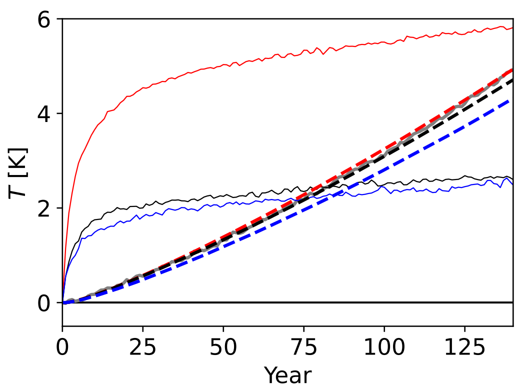

Figure 5. The model means of all curves in Figure 4.

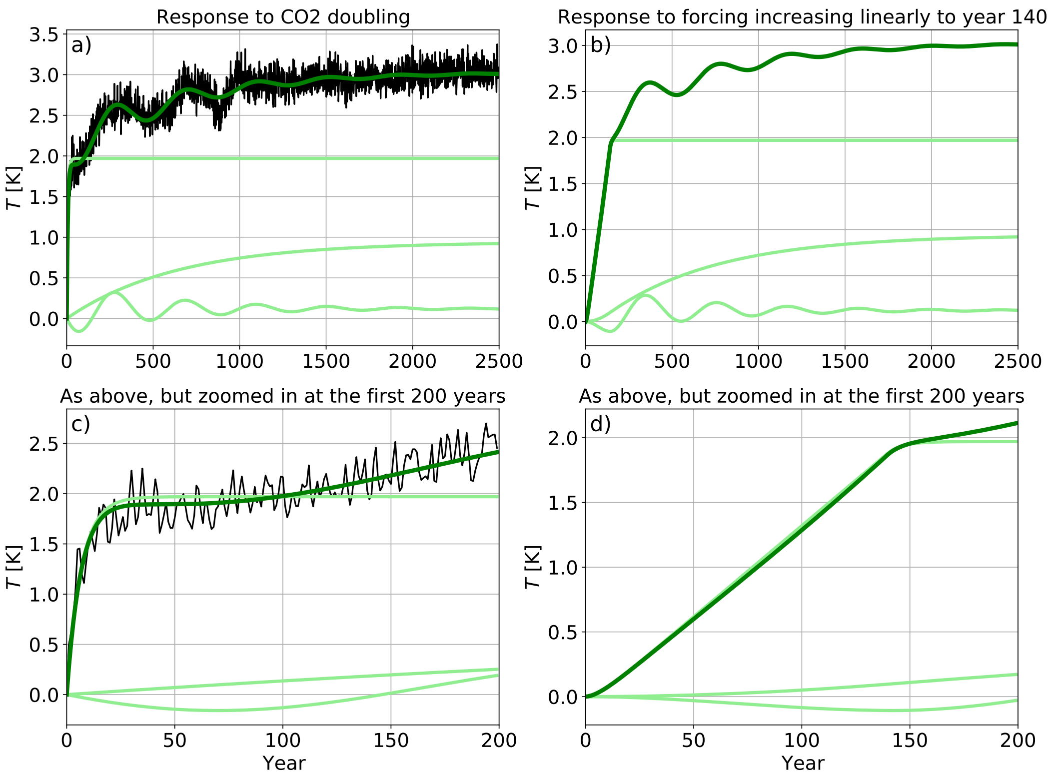

Figure 6. a) Result of fitting a two-exp and a pair of oscillatory responses to CESM104 abrupt2x. The dark green curves are the total responses to either an abrupt doubling of CO_2 (left) or a forcing increasing linearly to doubling of CO_2 in year 140, and is thereafter kept constant (right). The light green curves are components of the total response: Two exponential responses with time scales of approximately 7 and 639 years, and one oscillatory response with a period of approximately 410 years and damping time scale of 619 years.

6.3 Comparing different response functions

We fit two-exp, three-exp and two-exp + oscillatory response for all CMIP6 models. The resulting root mean squared error (RMSE) of these fits are summarised in Tables S6 and S7 for abrupt-4xCO2, Table S8 for abrupt-2xCO2 and Table S9 for abrupt-0p5xCO2. The results for LongRunMIP experiments are listed in Table S10. As expected, RMSE is always smaller or unchanged for the three-exp model compared to the two-exp model. With an ideal estimation method, the two-exp + osc. should be reduced to a three-exp (by setting $q = 0$ and $S_2 = 0$) if the oscillatory solution is not a better description than the three-exp. Hence all results here with increased RMSE are just the results of not finding the optimal parameters. However, for the models where we estimate higher RMSE values for the two-exp + osc, this model is very unlikely to be a good description. Going further, we will therefore just focus on the models where adding oscillations provides a better description.

Including oscillations provides a smaller RMSE compared to the three-exp model for 11/22 LongRunMIP abrupt experiments. For most of these experiments, the improvement is very minor, and probably not worth the additional parameters. However, for one of these simulations an oscillatory response provides a visually significant better description: the CESM104 abrupt2x, shown in Figure 6 a). This experiment is also studied in further detail in Section 7.1 and Figure 8.

In Figure 6 b) and d) we estimate the temperature response to a forcing that increases linearly until doubling (in year 140), and is then kept constant thereafter. This will be approximately half the output of 1pctCO2 experiments, and demonstrates that with this linear oscillatory model, the oscillations cannot be seen during the 140 years with linear forcing. The negative response of the oscillatory part is to a large degree cancelled out by the slow exponential part, and the majority of the temperature response is described by the fastest exponential response.

42/71 runs for CMIP6 abrupt-4xCO2 have smaller RMSE if including oscillations (note that we count different members from the same model). Also for these models, most improvements are so minor that we cannot really argue that the extra parameters are needed. Despite large estimation uncertainties for these shorter runs, we find indications that there may be oscillations in many models. In the following, we highlight results for members from the 8 models where we have the largest improvements in RMSE for abrupt-4xCO2: ACCESS-CM2, GISS-E2-1-G, ICON-ESM-LR, KIOST-ESM, MRI-ESM2-0, NorESM2-LM, SAM0-UNICON, TaiESM1. We note the generally close resemblance between these runs (see Figure 7) and the first 150 years of the CESM104 abrupt2x run in Figure 6 c).

The two-exp and oscillatory fits in Figure 7 show that the oscillatory component can take various shapes. For most members (e.g. TaiESM1 r1i1p1f1), the best fit includes an oscillatory component that resembles the purely exponential components, but where the initial warming overshoots before stabilizing at a lower equilibrium temperature. In these cases the estimated oscillations have a quick damping time scale (τ_p), typically 20-30 years. For MRI-ESM2-0 members r7 and r10 we have instead an oscillation starting with an initial cooling, which is part of a slow oscillation that could develop as in the CESM104 abrupt2x run. When including this slow oscillation, we find only shorter time scales (annual and decadal) for the two purely exponential parts. For the members where the oscillation has a shorter period, we have a centennial-scale purely exponential part to explain the slow variations in the temperature. Since we know from longer runs that a centennial-millennial scale exponential component is necessary to explain the full path to equilibrium, the fits for MRI-ESM2-0 members r7 and r10 are unlikely to explain the future of these experiments. This could in theory be resolved by combining the two short time-scale exponential parts to one, and allowing the second exponential part to take a long time scale instead. However, with only 150 years of data, a fit containing several components varying on centennial to millennial scales will be poorly constrained. The take-home message from this is that we cannot really tell from the global surface temperature of these short experiments if we deal with a short-period and quickly damped out oscillation or an oscillation lasting for centuries. Longer experiments are needed, but a closer look at the AMOC evolution and the spatial pattern of warming may also give some hints.

Of these 8 models, 3 models have also run abrupt-2xCO2 and abrupt-0p5xCO2 experiments. We see no clear signs of oscillations in these abrupt-0p5xCO2 runs. For GISS-E2-1-G abrupt-2xCO2 we observe a small flattening out of the temperature as for abrupt-4xCO2, for MRI-ESM2-0 abrupt-2xCO2 the temperature flattens out, and does not start to increase again. For TaiESM1 abrupt-2xCO2, the temperature behaves similarly as for abrupt-4xCO2 (although our estimated decomposition looks a bit different). Hence there are hints that the same phenomenon appears also for abrupt-2xCO2, but the responses may not be perfectly linear.

7 Oscillations and plateaus in global temperatures

7.1 Oscillation in CESM1 warming experiments

The CESM1 abrupt CO_2 responses are further investigated (Figure 8) by looking at the Northern Hemisphere (NH) and Southern Hemisphere (SH) temperatures separately (a), and by comparing with the AMOC index (b) and NH and SH sea ice areas (c). We find that the oscillations happen only in the NH, and that the abrupt2x (blue) NH temper-

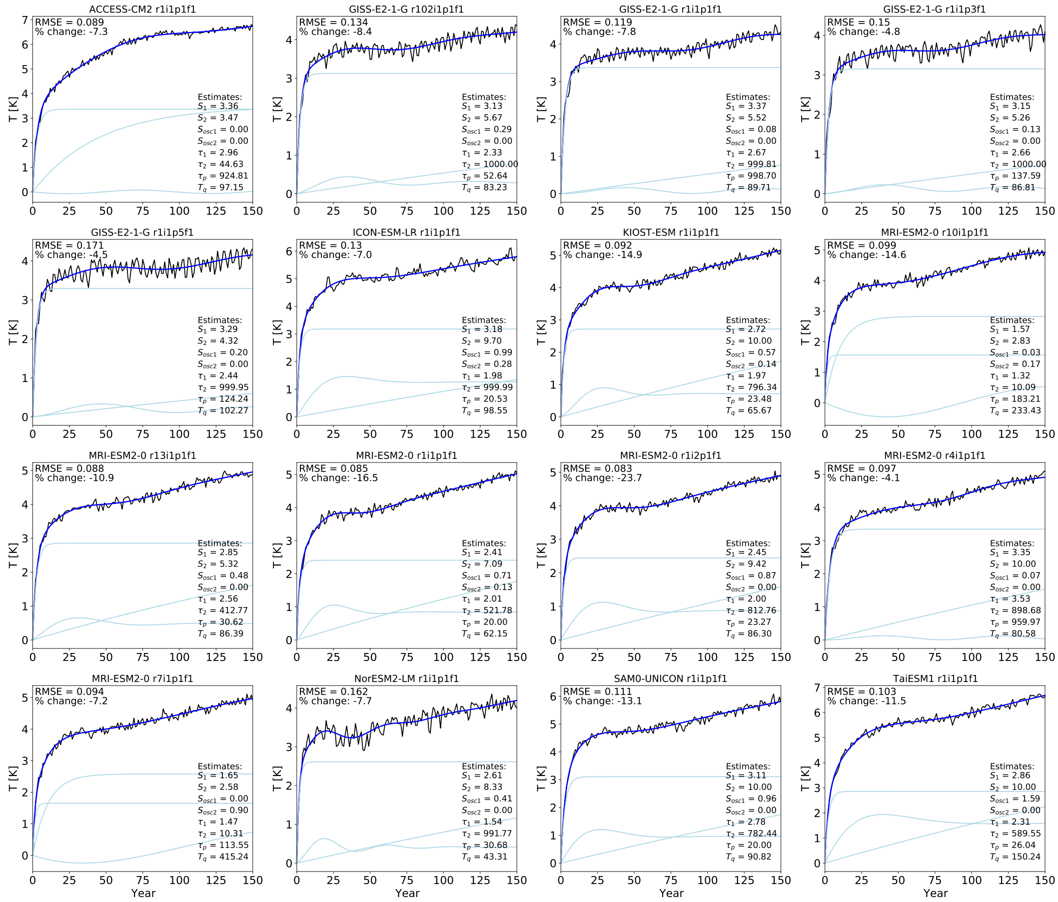

Figure 7. The two-exponential + oscillatory model fits (blue curves) for 16 different abrupt-4xCO2 runs (black curves). The light blue curves show the decomposition of the blue curve into two exponential components and one oscillatory component. The estimated parameters are listed in the figures, and the % change refers to the improvement in RMSE from three-exponential fit to the two-exponential + oscillatory model fit.

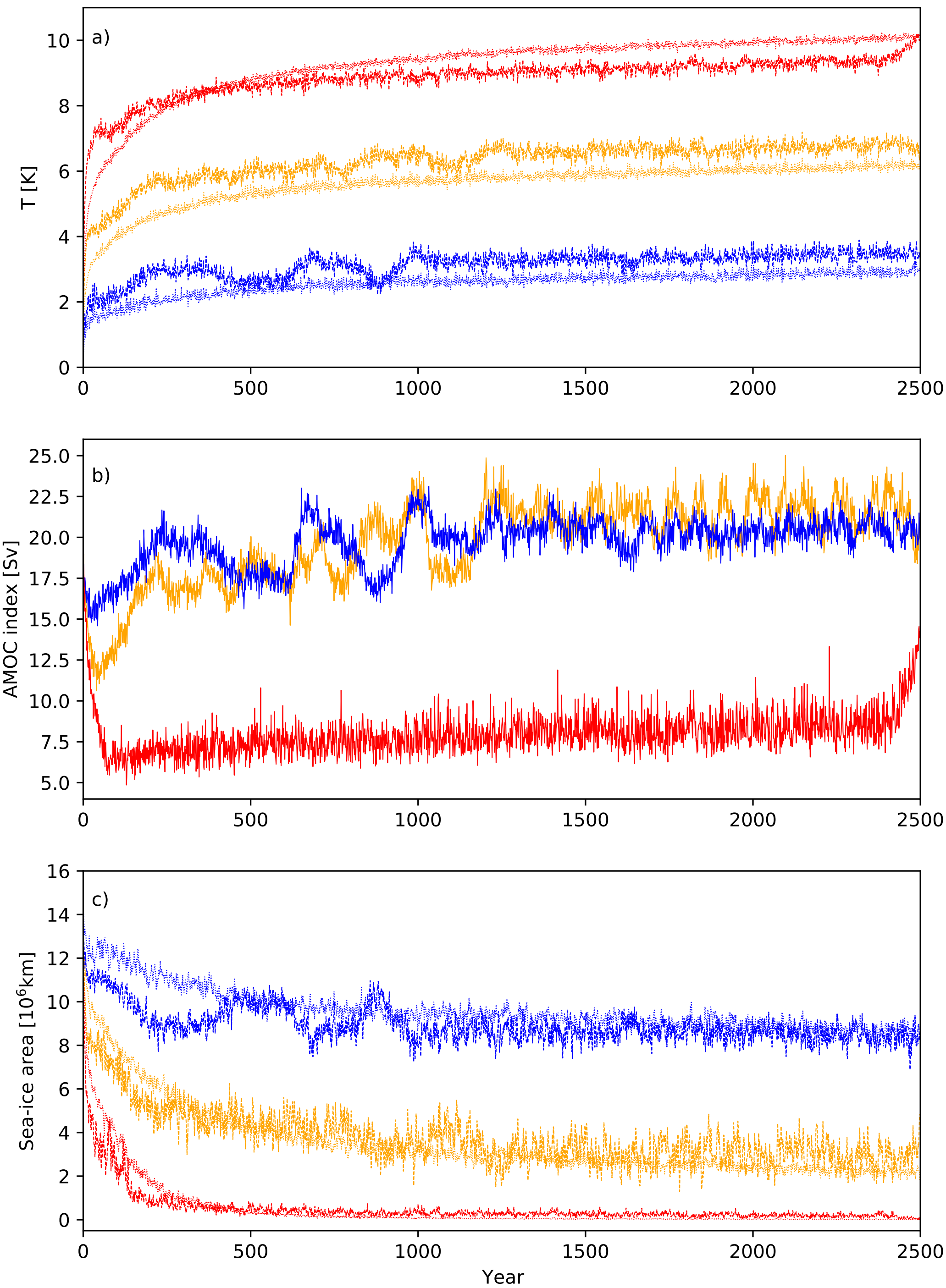

Figure 8. Mean surface temperature (a), AMOC index (b) and sea-ice area (c) for CESM104 abrupt2x (blue), abrupt4x (orange) and abrupt 8x (red). In a) and c), dashed curves are means over the Northern Hemisphere, and dotted (thinner) curves are means over the Southern Hemisphere.

ature is strongly correlated with the AMOC index (R = 0.796) and anticorrelated with the NH sea ice area (R = -0.919) if using all 2500 annual values for computation. If looking only at the first decades after the abrupt CO_2 doubling, we observe an anticorrelation between temperatures (which increase) and AMOC (which weakens). A plausible mechanism for this is that the strong initial warming inhibits the sinking of water in the North Atlantic by reducing its density. On longer time scales, AMOC changes also impact temperatures, by bringing more/less warm water northwards, which could explain the positive correlation.

The comparison with the abrupt 4x (orange) and 8x (red) simulations from the same model shows that all NH temperatures have a small plateau for some decades after the initial temperature increase, likely connected to their initial decrease in AMOC strength and sea-ice area. There are also some long-term variations later on in these experiments, but not following a similar oscillatory behaviour as the 2x experiment. We note for instance that the abrupt change around year 2500 in the abrupt8x experiment is strongly connected to an AMOC recovery. Hence, while linear response models estimated from the abrupt2x simulation may well describe the long-term responses to these other abrupt CO_2 experiments, the oscillatory behavior does not transfer to the same degree. In lack of more simulations with weaker forcing from this model, it is difficult to judge if the oscillatory phenomenon really is part of a linear model that can only be used for weaker forcings, or if it is a nonlinear effect or a random fluctuation.

7.2 Oscillation in cooling HadGEM experiment

Among models with abrupt-0p5xCO2 experiments, we find one (HadGEM-GC31-LL) with an interesting oscillation. This oscillation appears to have an increasing amplitude (see Figure 9 a)). To fit our model to these data, we need to allow the oscillatory part of the solution to have a positive real part eigenvalue, such that we get unstable/growing oscillations. This corresponds to a negative damping time scale τ_p. In b) we note that the oscillation appears mainly in the Southern Hemisphere, and is tightly connected to oscillations in the SH sea-ice extent. The Northern Hemisphere temperature is only slightly influenced by the oscillation, possibly through the atmosphere or because AMOC couples it to the SH. AMOC data are not provided for this experiment, but temperature changes in the North Atlantic (not shown) indicate that AMOC is changing. The estimated parameters are listed in the figure, and shows also that we have allowed negative values of S_{osc1} and S_{osc2}. The physical interpretation of this is that the SH sea ice actually decreases on average in extent, hence contributing to a warming on an otherwise cooling globe.

This oscillation seems to have a different physical origin than the oscillations/plateaus we observe in warming experiments. Similar changes in the SH were observed in the piControl experiment of this model (Ridley et al., 2022). In the piControl the deeper ocean has not yet reached an equilibrium state and the drifting temperatures eventually cause the water column in the Weddell and Ross seas to become unstable, and start to convect up warmer deeper ocean water that melts the sea ice. We suspect the oscillations in the abrupt-0p5xCO2 experiment is a similar phenomenon, except that in this run the cooling of the atmosphere and ocean surface layer brings the ocean column in the southern oceans faster into an unstable state. The more the surface is cooling, the larger the area can become where this instability and melting of sea ice happens, which can explain the growing oscillation and overall reduced sea ice cover.

7.3 Multidecadal pauses in global temperature increase

In Fig. 7 it can observed that the abrupt-4xCO2 simulations for several models (e.g., GISS-E2.1-G, MRI-ESM2.0, SAM0-UNICON) exhibit a plateau in their global mean surface temperature evolution after the initial fast-paced increase. This happens typically between years 30 and 70 and after year 70 the temperature starts increasing again. Averaging the temperature separately over northern and southern hemisphere (NH and SH,

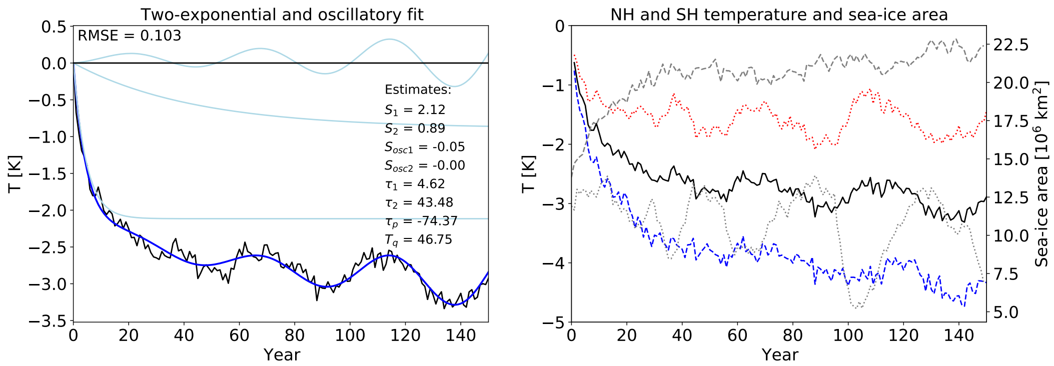

Figure 9. Results from HadGEM-GC31-LL abrupt-0p5xCO2 r1i1p1f3, where allowing an unstable (growing) oscillation makes a good fit. a) The black curve is the global surface air temperature change relative to piControl, the thick blue curve is the fitted model consisting of two exponential components (slowly varying light blue curves) and one oscillatory pair (plotted together as the oscillating light blue curve). Note that to make the fit the signs were flipped, such that the listed parameters $S_1, S_2, S_{osc1}, S_{osc2}$ are consistent with a positive response. b) The global temperature response (black) split up in Northern Hemisphere (NH, dashed blue) temperature and Southern Hemisphere (SH, dotted red) temperature. On the right axis we have the sea-ice area, which is plotted for the SH (dotted gray) and NH (dashed gray).

respectively; see Fig. 10 for the example of GISS-E2.1-G) reveals that the plateau of the global mean temperature results from a plateauing or even decrease of the NH temperature while the SH temperature increases monotonically. More specifically, maps of time slices of surface warming make clear that it is the North Atlantic that cools in response to the CO_2-forcing (Fig. 10, left column). Models that do not exhibit the plateauing global mean temperature typically exhibit neither the plateauing in the NH nor the cooling (or lack of warming) in the North Atlantic (E3SM-1.0 shown as an example in Fig. 10, right column). Though there may be models where the North Atlantic cools/warms less, but not enough to cause a significant slowdown of global temperature increase.

The difference in North Atlantic temperatures between models with and without plateau is found to be concomitant with a difference in the development of AMOC and the development of Arctic sea ice (see Figure 10), consistent with earlier studies (Bellomo et al., 2021; Mitevski et al., 2021). Models with plateauing global mean temperature tend to simulate a stronger AMOC decline in response to the CO_2-forcing (e.g. GISS-E2-1-G and SAM0-UNICON) than do the models without plateau. Notably, the pre-industrial AMOC also tends to be stronger in models with plateau than in those without plateau. Furthermore, models with plateau retain more of their Arctic sea ice than models without plateau. The connection between a plateauing global temperature, weakening AMOC, and enhanced NH sea ice cover was also noted by Held et al. (2010) for the GFDL Climate Model version 2.1.

A stronger decline in AMOC is consistent with lower North Atlantic temperatures (Bellomo et al., 2021) and less sea ice melt (Yeager et al., 2015; Liu et al., 2020; Eiselt & Graversen, 2023). The AMOC constitutes a part of the poleward energy transport in the climate system that is necessary to balance the differential energy input from solar radiation. The AMOC accomplishes northward energy transport by transporting warm water from the Tropics into the Arctic increasing the ocean heat release there and thus warming the North Atlantic. A decline of the AMOC will hence lead to a cooling or at least a hampering of the warming in response to a CO_2-forcing. Growing sea ice in response to a cooling

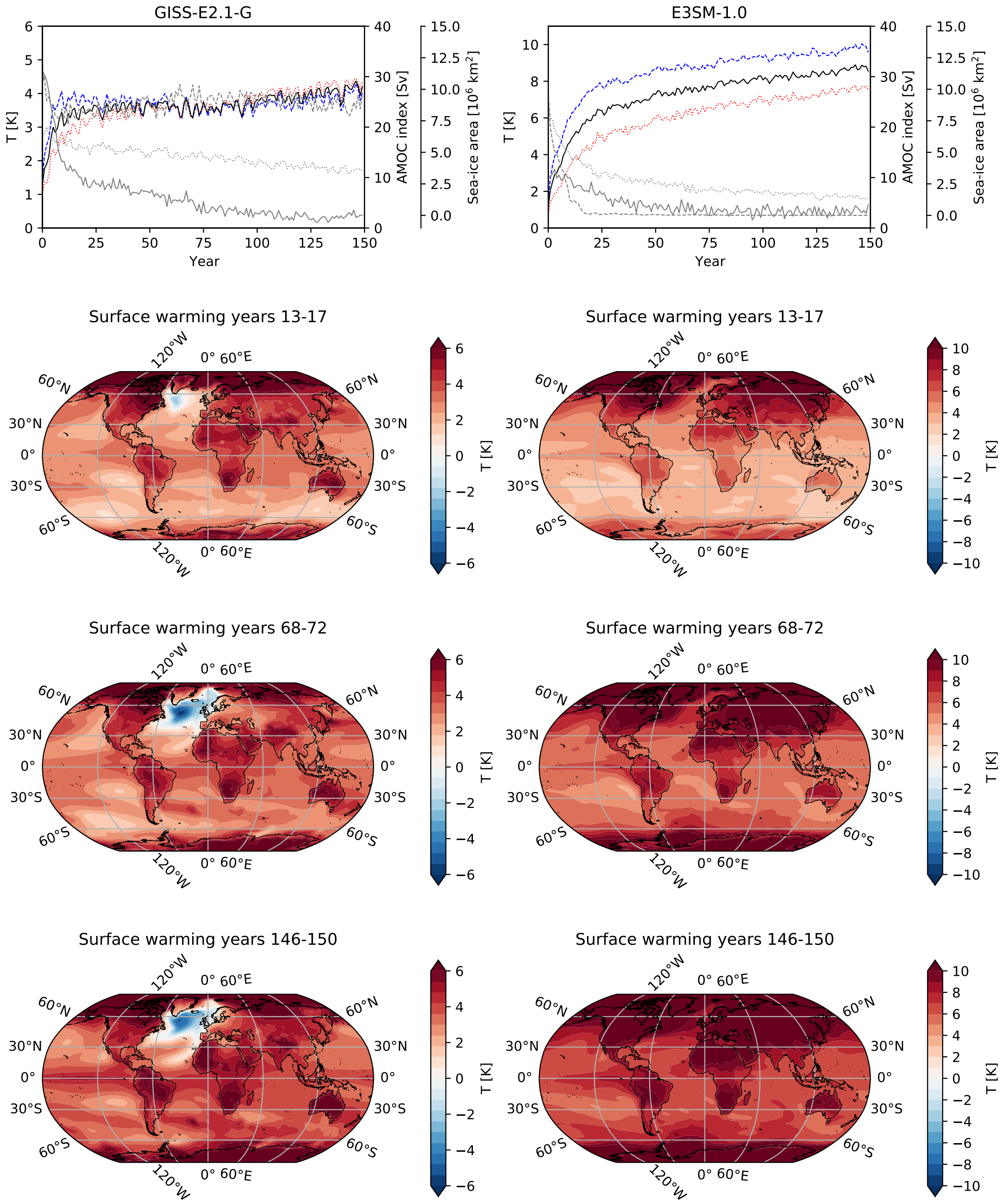

Figure 10. Example of models with and without plateaus in global temperature.

will contribute to keeping the temperature low for a while. Changes in sea ice has also been shown to affect AMOC (Sévellec et al., 2017; Liu et al., 2019; Madan et al., 2023). The growth of sea ice can therefore be an explanation for an eventual AMOC recovery, and finally lead to a decay of the oscillating component.

8 Discussion

Many earlier studies comparing different abrupt CO_2 experiments focus on experiments from single models, and are often mainly interested in the equilibrium response. Such studies find both decreasing and increasing climate sensitivities with stronger CO_2 forcing (see discussions in Meraner et al. (2013); Bloch-Johnson et al. (2021)), but the more comprehensive analysis by Bloch-Johnson et al. (2021) (including many of the same models as this paper) finds that climate sensitivity increases in most models.

Slab-ocean models are used in several studies (Colman & McAvaney, 2009; Meraner et al., 2013), and are useful tools for studying the temperature-dependence of atmospheric feedbacks. They are relatively cheap to run, and the pattern effect is somewhat suppressed in these models, partly because they go quicker to equilibrium and partly due to the lack of ocean dynamics that can change the pattern of the temperature response. This makes it easier to separate the nonlinear/temperature dependent feedbacks from the pattern effect, but ignores also possible permanent changes in feedbacks due to changes in the ocean circulation.

For a wide range of abrupt CO_2 increase experiments (1x to 8x), Mitevski et al. (2021) finds that the increase in effective climate sensitivity with increasing CO_2 is not monotonic in two fully coupled models (GISS-E2.1-G and CESM-LE), in contrast to the monotonic increase found in slab-ocean experiments (Meraner et al., 2013; Mitevski et al., 2021). The nonmonotonic increase is related to the decreasing temperatures in the North Atlantic and the weakening AMOC. For small enough abrupt CO_2 concentration increases (up to 2x and 3x CO_2 for GISS-E2.1-G and CESM-LE, respectively) the AMOC recovers after the initial decrease, while for higher concentrations it does not. For higher concentrations, the North Atlantic cools less however, because of the increased warming from CO_2.

Manabe and Stouffer (1993, 1994) also focused on studying the thermohaline circulation in the Atlantic Ocean in different abrupt CO_2 experiments. In their 2x and 4x experiments they observe a weakening of the thermohaline circulation. The circulation recovered again for $2xCO_2$, but remained weak for $4xCO_2$. For $0.5xCO_2$ Stouffer and Manabe (2003) finds a weak and shallow thermohaline circulation in the Atlantic.

The collapse of AMOC above a certain CO_2 level is an example of how a change in the ocean circulation can cause a nonlinear global temperature response. A change in circulation changes the surface temperature pattern, which further modulates which atmospheric feedbacks are triggered. In the case of a permanent collapse of AMOC, the new pattern and associated feedbacks are also permanently changed. In general, any change in effectiveness of deeper ocean heat uptake can depend on state, and therefore result in a nonlinear response. A warming of the surface can lead to a more stratified ocean with reduced vertical mixing. To some extent, however, the reduced heat uptake can still be approximated as a linear function of the surface temperature increase. We have also demonstrated the opposite effect here, that a cooling of the surface can lead to a linear oscillating response, as a result of ocean-sea ice dynamics in the Southern Ocean.

Linear response models can take many forms. Examples of physically motivated models are the upwelling-diffusion models (Hoffert et al., 1980) used in the First IPCC report, and the temperature component of the FaIR emulator (Millar et al., 2017; Smith et al., 2018; Leach et al., 2021) used in AR6 (P. Forster et al., 2021). They are powerful tools for e.g. the IPCC reports since they can be used to quickly explore a wider range of forcing scenarios than that simulated by coupled models. We suggest that a generalised box model is easier to interpret, test and generalise than box models using an efficacy factor, since temperature components and different feedback parameters are more directly associated with the pattern of surface temperature evolution, instead of being indirectly associated through an efficacy factor. We do not have to assume anything about the distribution of the boxes as long as we are interested in global quantities, but in order to better constrain the values of the different feedback parameters, the additional information about the pattern can be useful.

9 Conclusions

We find that linear response is overall a good assumption for global surface temperatures. However, good predictions with linear response models are crucially dependent on good forcing estimates. Distinguishing between forcing and response is a challenge, and the uncertainty of forcing estimates is the main limitation to determining if a model has a linear response or not.

Mitevski et al. (2022) and Geoffroy and Saint-Martin (2020) highlight the importance of taking into account the nonlogarithmic dependence of the forcing on the CO_2 concentration. This implies stronger forcing for each CO_2 doubling, also consistent with recent findings of (He et al., 2023). He et al. (2023) finds that the stratospheric temperature impacts CO_2 forcing, and that other forcing agents affecting the stratospheric temperature therefore can modulate the CO_2 forcing. Such nonlinear interaction between forcing agents should be studied in further detail, as this deviates from a linear framework. We hope also the effort initiated by RFMIP (Pincus et al., 2016) to better constrain forcing estimates will be continued for more models and experiments in the future.

For models with a plateau in the global temperature response to an abrupt increase in CO_2 stemming from a cooling of the North Atlantic, the cooling component (which can be modelled with an oscillatory part) can counteract the warming from the slow centennial-millennial scale component for a long time. For these models, a response model with a single exponential response can actually be sufficient for many short-term prediction purposes. In CESM104 abrupt2x a single exponential explains the majority of the first decades after abrupt doubling of CO_2, and for all 140 years with linearly increasing forcing.

Parameter estimation taking into account the possibility for centennial-scale oscillations is difficult for short time series, like the typical 150 year abrupt CO_2 experiments. We encourage more models to run longer abrupt CO_2 experiments, also for different levels of CO_2. Longer runs will help constrain linear response models better on the longer term, which can then further be used to quickly predict a wide range of other forcing scenarios. In particular, more and longer abrupt-2xCO2 would be useful, since these are very likely to be within the range where a linear response is a good approximation. Linear responses estimated from abrupt-4xCO2 are also quite good approximations, but there are some signs of nonlinear responses playing a role in these experiments (Fredriksen et al., 2023; Bloch-Johnson et al., 2021). CMIP6 abrupt-4xCO2 warms on average 2.2 times abrupt-2xCO2, and we estimate that about a factor 2 can be attributed to the forcing difference. The remaining 10% extra warming in abrupt-4xCO2 is likely attributed to nonlinear responses, such as feedback changes (Bloch-Johnson et al., 2021).

References

Andrews, T., Gregory, J. M., & Webb, M. J. (2015). The Dependence of Radiative Forcing and Feedback on Evolving Patterns of Surface Temperature Change in Climate Models. *Journal of Climate*, *28*(4), 1630–1648. doi: 10.1175/JCLI-D-14-00545.1

Andrews, T., Smith, C. J., Myhre, G., Forster, P. M., Chadwick, R., & Ackerley, D. (2021). Effective Radiative Forcing in a GCM With Fixed Surface Tempera-

tures. *Journal of Geophysical Research: Atmospheres*, *126*, e2020JD033880. doi: 10.1029/2020JD033880
Armour, K. C., Bitz, C. M., & Roe, G. H. (2013). Time-Varying Climate Sensitivity from Regional Feedbacks. *Journal of Climate*, *26*, 4518–4534. doi: 10.1175/JCLI-D-12-00544.1
Bellomo, K., Angeloni, M., Corti, S., & von Hardenberg, J. (2021). Future climate change shaped by inter-model differences in Atlantic meridional overturning circulation response. *Nature Communications*, *12*, 3659. doi: 10.1038/s41467-021-24015-w
Bloch-Johnson, J., Pierrehumbert, R. T., & Abbot, D. S. (2015). Feedback temperature dependence determines the risk of high warming. *Geophysical Research Letters*, *42*(12), 4973-4980. doi: 10.1002/2015GL064240
Bloch-Johnson, J., Rugenstein, M., Stolpe, M. B., Rohrschneider, T., Zheng, Y., & Gregory, J. M. (2021). Climate Sensitivity Increases Under Higher CO_2 Levels Due to Feedback Temperature Dependence. *Geophysical Research Letters*, *48*, e2020GL089074. doi: 10.1029/2020GL089074
Caldeira, K., & Myhrvold, N. P. (2013). Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration. *Environmental Research Letters*, *8*(3), 034039. doi: 10.1088/1748-9326/8/3/034039
Colman, R., & McAvaney, B. (2009). Climate feedbacks under a very broad range of forcing. *Geophysical Research Letters*, *36*(1). doi: 10.1029/2008GL036268
Cummins, D. P., Stephenson, D. B., & Stott, P. A. (2020). Optimal Estimation of Stochastic Energy Balance Model Parameters. *Journal of Climate*, *33*(18), 7909–7926. doi: 10.1175/JCLI-D-19-0589.1
Edwards, C., & Penney, D. (2007). *Differential equations and boundary value problems: Computing and modelling (Fourth edition)*. Pearson.
Eiselt, K.-U., & Graversen, R. G. (2023). On the Control of Northern Hemispheric Feedbacks by AMOC: Evidence from CMIP and Slab Ocean Modeling. *Journal of Climate*, *36*(19), 6777–6795. doi: 10.1175/JCLI-D-22-0884.1
Etminan, M., Myhre, G., Highwood, E. J., & Shine, K. P. (2016). Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing. *Geophysical Research Letters*, *43*(24), 12,614–12,623. doi: 10.1002/2016GL071930
Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., ... Zhang, H. (2021). The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity [Book Section]. In V. Masson-Delmotte et al. (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (p. 923–1054). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: 10.1017/9781009157896.009
Forster, P. M., Richardson, T., Maycock, A. C., Smith, C. J., Samset, B. H., Myhre, G., ... Schulz, M. (2016). Recommendations for diagnosing effective radiative forcing from climate models for CMIP6. *Journal of Geophysical Research: Atmospheres*, *121*(20), 12,460–12,475. doi: 10.1002/2016JD025320
Fredriksen, H.-B., Rugenstein, M., & Graversen, R. (2021). Estimating Radiative Forcing With a Nonconstant Feedback Parameter and Linear Response. *Journal of Geophysical Research: Atmospheres*, *126*(24), e2020JD034145. doi: 10.1029/2020JD034145
Fredriksen, H.-B., & Rypdal, M. (2017). Long-range persistence in global surface temperatures explained by linear multibox energy balance models. *Journal of Climate*, *30*, 7157–7168. doi: 10.1175/JCLI-D-16-0877.1
Fredriksen, H.-B., Smith, C. J., Modak, A., & Rugenstein, M. (2023). 21st Century Scenario Forcing Increases More for CMIP6 Than CMIP5 Models. *Geophysical Research Letters*, *50*(6), e2023GL102916. doi: 10.1029/2023GL102916
Geoffroy, O., & Saint-Martin, D. (2020). Equilibrium- and Transient-State Depen-

dencies of Climate Sensitivity: Are They Important for Climate Projections? *Journal of Climate*, *33*(5), 1863 – 1879. doi: 10.1175/JCLI-D-19-0248.1

Geoffroy, O., Saint-Martin, D., Bellon, G., Voldoire, A., Olivié, D., & Tytéca, S. (2013). 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*(6), 1859–1876. doi: 10.1175/JCLI-D-12-00196.1

Geoffroy, O., Saint-Martin, D., Olivié, D. J. L., Voldoire, A., Bellon, G., & Tytéca, S. (2013). Transient Climate Response in a Two-Layer Energy-Balance Model. Part I: Analytical Solution and Parameter Calibration Using CMIP5 AOGCM Experiments. *Journal of Climate*, *26*, 1841–1857. doi: 10.1175/JCLI-D-12-00195.1

Good, P., Andrews, T., Chadwick, R., Dufresne, J.-L., Gregory, J. M., Lowe, J. A., … Shiogama, H. (2016). nonlinMIP contribution to CMIP6: model intercomparison project for non-linear mechanisms: physical basis, experimental design and analysis principles (v1.0). *Geoscientific Model Development*, *9*(11), 4019–4028. doi: 10.5194/gmd-9-4019-2016

Good, P., Gregory, J. M., & Lowe, J. A. (2011). A step-response simple climate model to reconstruct and interpret AOGCM projections. *Geophysical Research Letters*, *38*, L01703. doi: 10.1029/2010GL045208

Good, P., Gregory, J. M., Lowe, J. A., & Andrews, T. (2013). Abrupt CO2 experiments as tools for predicting and understanding CMIP5 representative concentration pathway projections. *Climate Dynamics*, *40*(3), 1041–1053. doi: 10.1007/s00382-012-1410-4

Gregory, J. M., Andrews, T., & Good, P. (2015). The inconstancy of the transient climate response parameter under increasing CO_2. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *373*, 20140417. doi: 10.1098/rsta.2014.0417

Gregory, J. M., Ingram, W. J., Palmer, M. A., Jones, G. S., Stott, P. A., Thorpe, R. B., … Williams, K. D. (2004). A new method for diagnosing radiative forcing and climate sensitivity. *Geophysical Research Letters*, *31*, L03205. doi: 10.1029/2003GL018747

Hansen, J., Sato, M., Ruedy, R., Nazarenko, L., Lacis, A., Schmidt, G. A., … Zhang, S. (2005). Efficacy of climate forcings. *Journal of Geophysical Research: Atmospheres*, *110*(D18). doi: 10.1029/2005JD005776

Hasselmann, K., Sausen, R., Maier-Reimer, E., & Voss, R. (1993). On the cold start problem in transient simulations with coupled atmosphere-ocean models. *Climate Dynamics*, *9*(6), 53–61. doi: 10.1007/BF00210008

He, H., Kramer, R. J., Soden, B. J., & Jeevanjee, N. (2023). State dependence of CO2 forcing and its implications for climate sensitivity. *Science*, *382*(6674), 1051–1056. doi: 10.1126/science.abq6872

Held, I., Winton, M., Takahashi, K., Delworth, T. L., Zeng, F., & Vallis, G. (2010). Probing the Fast and Slow Components of Global Warming by Returning Abruptly to Preindustrial Forcing. *Journal of Climate*, *23*, 2418 – 2427. doi: 10.1175/2009JCLI3466.1

Hoffert, M. I., Callegari, A. J., & Hsieh, C.-T. (1980). The role of deep sea heat storage in the secular response to climatic forcing. *Journal of Geophysical Research: Oceans*, *85*, 6667-6679. doi: 10.1029/JC085iC11p06667

Jackson, L. S., Maycock, A. C., Andrews, T., Fredriksen, H.-B., Smith, C. J., & Forster, P. M. (2022). Errors in Simple Climate Model Emulations of Past and Future Global Temperature Change. *Geophysical Research Letters*, *49*(15), e2022GL098808. doi: 10.1029/2022GL098808

Jiang, W., Gastineau, G., & Codron, F. (2023). Climate Response to Atlantic Meridional Energy Transport Variations. *Journal of Climate*, *36*(16), 5399 – 5416. doi: 10.1175/JCLI-D-22-0608.1

Larson, E. J. L., & Portmann, R. W. (2016). A Temporal Kernel Method to Compute Effective Radiative Forcing in CMIP5 Transient Simulations. *Journal of Climate*, *29*(4), 1497-1509. doi: 10.1175/JCLI-D-15-0577.1

Leach, N. J., Jenkins, S., Nicholls, Z., Smith, C. J., Lynch, J., Cain, M., … Allen, M. R. (2021). FaIRv2.0.0: a generalized impulse response model for climate uncertainty and future scenario exploration. *Geoscientific Model Development*, *14*(5), 3007–3036. doi: 10.5194/gmd-14-3007-2021

Lin, Y.-J., Hwang, Y.-T., Ceppi, P., & Gregory, J. M. (2019). Uncertainty in the Evolution of Climate Feedback Traced to the Strength of the Atlantic Meridional Overturning Circulation. *Geophysical Research Letters*, *46*(21), 12331-12339. doi: 10.1029/2019GL083084

Liu, W., Fedorov, A., & Sévellec, F. (2019). The Mechanisms of the Atlantic Meridional Overturning Circulation Slowdown Induced by Arctic Sea Ice Decline. *Journal of Climate*, *32*(4), 977 – 996. doi: 10.1175/JCLI-D-18-0231.1

Liu, W., Fedorov, A. V., Xie, S.-P., & Hu, S. (2020). Climate impacts of a weakened Atlantic Meridional Overturning Circulation in a warming climate. *Science Advances*, *6*(26), eaaz4876. doi: 10.1126/sciadv.aaz4876

Madan, G., Gjermundsen, A., Iversen, S. C., & LaCasce, J. H. (2023). The weakening AMOC under extreme climate change. *Climate Dynamics*. doi: 10.1007/s00382-023-06957-7

Manabe, S., & Stouffer, R. J. (1993). Century-scale effects of increased atmospheric C02 on the ocean–atmosphere system. *Nature*, *364*, 215 – 218. doi: 10.1038/364215a0

Manabe, S., & Stouffer, R. J. (1994). Multiple-Century Response of a Coupled Ocean-Atmosphere Model to an Increase of Atmospheric Carbon Dioxide. *Journal of Climate*, *7*(1). doi: 10.1175/1520-0442(1994)007⟨0005:MCROAC⟩2.0.CO;2

Meraner, K., Mauritsen, T., & Voigt, A. (2013). Robust increase in equilibrium climate sensitivity under global warming. *Geophysical Research Letters*, *40*(22), 5944–5948. doi: 10.1002/2013GL058118

Millar, R. J., Nicholls, Z. R., Friedlingstein, P., & Allen, M. R. (2017). 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*(11), 7213–7228. doi: 10.5194/acp-17-7213-2017

Mitevski, I., Orbe, C., Chemke, R., Nazarenko, L., & Polvani, L. M. (2021). Non-Monotonic Response of the Climate System to Abrupt CO2 Forcing. *Geophysical Research Letters*, *48*(6), e2020GL090861. doi: 10.1029/2020GL090861

Mitevski, I., Polvani, L. M., & Orbe, C. (2022). Asymmetric Warming/Cooling Response to CO2 Increase/Decrease Mainly Due To Non-Logarithmic Forcing, Not Feedbacks. *Geophysical Research Letters*, *49*(5), e2021GL097133. doi: 10.1029/2021GL097133

Pincus, R., Forster, P. M., & Stevens, B. (2016). The Radiative Forcing Model Intercomparison Project (RFMIP): experimental protocol for CMIP6. *Geoscientific Model Development*, *9*(9), 3447–3460. doi: 10.5194/gmd-9-3447-2016

Proistosescu, C., & Huybers, P. J. (2017). Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Sciences Advances*, *3*, e1602821. doi: 10.1126/sciadv.1602821

Richardson, T. B., Forster, P. M., Smith, C. J., Maycock, A. C., Wood, T., Andrews, T., … Watson-Parris, D. (2019). Efficacy of Climate Forcings in PDRMIP Models. *Journal of Geophysical Research: Atmospheres*, *124*(23), 12824-12844. doi: 10.1029/2019JD030581

Ridley, J. K., Blockley, E. W., & Jones, G. S. (2022). A Change in Climate State During a Pre-Industrial Simulation of the CMIP6 Model HadGEM3 Driven by Deep Ocean Drift. *Geophysical Research Letters*, *49*(6), e2021GL097171. doi: 10.1029/2021GL097171

Rohrschneider, T., Stevens, B., & Mauritsen, T. (2019). On simple representations of the climate response to external radiative forcing. *Climate Dynamics*, *53*(5), 3131–3145. doi: 10.1007/s00382-019-04686-4

Rugenstein, M., Bloch-Johnson, J., Abe-Ouchi, A., Andrews, T., Beyerle, U., Cao, L., ... Yang, S. (2019). LongRunMIP: Motivation and Design for a Large Collection of Millennial-Length AOGCM Simulations. *Bulletin of the American Meteorological Society*, *100*(12), 2551-2570. doi: 10.1175/BAMS-D-19-0068.1

Smith, C. J., Forster, P. M., Allen, M., Leach, N., Millar, R. J., Passerello, G. A., & Regayre, L. A. (2018). FAIR v1.3: a simple emissions-based impulse response and carbon cycle model. *Geoscientific Model Development*, *11*(6), 2273–2297. doi: 10.5194/gmd-11-2273-2018

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., ... Forster, P. M. (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16), 9591–9618. doi: 10.5194/acp-20-9591-2020

Stevens, B., Sherwood, S. C., Bony, S., & Webb, M. J. (2016). Prospects for narrowing bounds on Earth's equilibrium climate sensitivity. *Earth's Future*, *4*(11), 512-522. doi: 10.1002/2016EF000376

Stouffer, R. J., & Manabe, S. (2003). Equilibrium response of thermohaline circulation to large changes in atmospheric CO2 concentration. *Climate Dynamics*, *20*, 759 – 773. doi: 10.1007/s00382-002-0302-4

Sévellec, F., Fedorov, A. V., & Liu, W. (2017). Arctic sea-ice decline weakens the Atlantic Meridional Overturning Circulation. *Nature Climate Change*, *7*, 604 – 610. doi: 10.1038/nclimate3353

Tang, T., Shindell, D., Faluvegi, G., Myhre, G., Olivié, D., Voulgarakis, A., ... Smith, C. (2019). Comparison of Effective Radiative Forcing Calculations Using Multiple Methods, Drivers, and Models. *Journal of Geophysical Research: Atmospheres*, *124*(8), 4382-4394. doi: 10.1029/2018JD030188

Winton, M., Takahashi, K., & Held, I. M. (2010). Importance of Ocean Heat Uptake Efficacy to Transient Climate Change. *Journal of Climate*, *23*(9), 2333-2344. doi: 10.1175/2009JCLI3139.1

Yeager, S. G., Karspeck, A. R., & Danabasoglu, G. (2015). Predicted slowdown in the rate of Atlantic sea ice loss. *Geophysical Research Letters*, *42*, 10704 – 10713. doi: 10.1002/2015GL065364

Zhou, C., Wang, M., Zelinka, M. D., Liu, Y., Dong, Y., & Armour, K. C. (2023). Explaining Forcing Efficacy With Pattern Effect and State Dependence. *Geophysical Research Letters*, *50*(3), e2022GL101700. doi: 10.1029/2022GL101700

10 Open Research

Code is available in github (`https://github.com/Hegebf/Testing-Linear-Responses`), and will be deployed in zenodo to get a doi when the manuscript is accepted. The CMIP6 data are available through ESGF (`https://aims2.llnl.gov/search/?project=CMIP6/`), and the processed version used here is deployed in `https://doi.org/10.5281/zenodo.7687534`. LongRunMIP data can be accessed through `https://www.longrunmip.org/`.

Acknowledgments

We thank Jeff Ridley for discussions that helped us understand the behaviour of the model HadGEM-GC31-LL. We would also like to thank everyone who contributed to producing the LongRunMIP and CMIP6 model data used in this study. The work of author Hege-Beate Fredriksen was partly funded by the European Union as part of the EPOC project (Explaining and Predicting the Ocean Conveyor). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The work of Kai-Uwe Eiselt was part of the project UiT - Climate

Initiative, Ice-ocean-atmosphere interactions in the Arctic - from the past to the future, funded by the Faculty of Science and Technology, UiT the Arctic University of Norway. Peter Good was supported by the Met Office Hadley Centre Climate Programme funded by DSIT.

Appendix A Solution of generalized box model

Here we will derive the solution of a generalized box model, based on theory from Edwards and Penney (2007).

The general box model is given by the linear system:

$$\frac{d\mathbf{T}(t)}{dt} = \mathbf{C}^{-1}\mathbf{K}\mathbf{T}(t) + \mathbf{C}^{-1}\mathbf{F}(t) \tag{A1}$$

We consider first the homogeneous problem

$$\frac{d\mathbf{T}_h(t)}{dt} = \mathbf{A}\mathbf{T}_h(t)$$

where $\mathbf{A} = \mathbf{C}^{-1}\mathbf{K}$. We note that the matrix of possible solutions (the fundamental matrix) is:

$$\mathbf{\Phi}(t) = [\mathbf{v_1}e^{\gamma_1 t} \,|\, \mathbf{v_2}e^{\gamma_2 t} \,|\, \ldots \,|\, \mathbf{v_n}e^{\gamma_n t}].$$

where $\mathbf{v_n}$ are the eigenvectors corresponding to the eigenvalues γ_n of the matrix $\mathbf{A}$. If we also set an initial condition $\mathbf{T}(0) = \mathbf{T}_0$, the homogeneous solution takes the form:

$$\mathbf{T}_h(t) = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}\mathbf{T}_0 \tag{A2}$$

An alternative notation when $\mathbf{A}$ consists of constant coefficients is the matrix exponential $e^{\mathbf{A}t} = \mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}$, since

$$\frac{d\mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}}{dt} = \frac{de^{\mathbf{A}t}}{dt} = \mathbf{A}e^{\mathbf{A}t} = \mathbf{A}\mathbf{\Phi}(t)\mathbf{\Phi}(0)^{-1}.$$

We note that the elements of $e^{\mathbf{A}t}$ are a linear combination of elements of $\mathbf{\Phi}(t)$.

Consider the case where we have a pair of complex conjugate eigenvalues, $\gamma_1 = \overline{\gamma_2}$, $\mathbf{v_1} = \overline{\mathbf{v_2}}$. Let $\mathbf{v_2} = \mathbf{a} + i\mathbf{b}$ and $\gamma_2 = p + iq$, such that

$$\begin{aligned}
\mathbf{v_2}e^{\gamma_2 t} &= (\mathbf{a} + i\mathbf{b})e^{(p+iq)t} \\
&= (\mathbf{a} + i\mathbf{b})e^{pt}(\cos qt + i\sin qt) \\
&= e^{pt}(\mathbf{a}\cos qt - \mathbf{b}\sin qt) + ie^{pt}(\mathbf{b}\cos qt + \mathbf{a}\sin qt)
\end{aligned}$$

Then the pair of complex eigenvalue solutions can instead be given by the real and complex part of the expression above, such that:

$$\mathbf{\Phi}(t) = [e^{pt}(\mathbf{a}\cos qt - \mathbf{b}\sin qt) \,|\, e^{pt}(\mathbf{b}\cos qt + \mathbf{a}\sin qt) \,|\, \mathbf{v_3}e^{\gamma_3 t} \,|\, \ldots \,|\, \mathbf{v_n}e^{\gamma_n t}].$$

The fundamental matrix of the homogeneous problem is also used to describe the particular solution to the original nonhomogeneous system:

$$\mathbf{T}_p(t) = e^{\mathbf{A}t}\int e^{-\mathbf{A}t}\mathbf{C}^{-1}\mathbf{F}(t)dt = \int e^{\mathbf{A}(t-s)}\mathbf{C}^{-1}\mathbf{F}(s)ds.$$

We assume that the forcing vector $\mathbf{F}(t)$ is a vector of constants $\mathbf{w}$ multiplied by the global mean forcing $F(t)$. Further, we note that computing the matrix product $e^{\mathbf{A}(t-s)}\mathbf{C}^{-1}$ only results in extra constant factors to each entry of $e^{\mathbf{A}(t-s)}$, such that the resulting column vector obtained from $e^{\mathbf{A}(t-s)}\mathbf{C}^{-1}\mathbf{w}$ will therefore be a linear combination of the entries of $e^{\mathbf{A}(t-s)}$ (or $\mathbf{\Phi}(t)$).

Finally, the global mean surface temperature $T(t)$ can be described as a linear combination (area-weighted average) of the components of the vector $\mathbf{T}_p(t) + \mathbf{T}_h(t)$,

$$T(t) = G^*(t)T_0 + \int_0^t G(t-s)F(s)ds \tag{A3}$$

where

$$\begin{aligned} G(t) &= e^{pt}(c_1 \cos qt - c_2 \sin qt) + e^{pt}(c_3 \cos qt + c_4 \sin qt) + \sum_{n=3}^{K} k_n e^{\gamma_n t} & \text{(A4)} \\ &= k_1 e^{pt} \cos qt + k_2 e^{pt} \sin qt + \sum_{n=3}^{K} k_n e^{\gamma_n t} & \text{(A5)} \end{aligned}$$

and $G^*(t)$ takes the same form as $G(t)$, but has different coefficients k_n. In case of more pairs of complex solutions, we can replace more pairs from $\sum_{n=3}^{K} k_n e^{\gamma_n t}$ by oscillatory solutions of the same form as $k_1 e^{pt} \cos qt + k_2 e^{pt} \sin qt$. For the system to be stable we must require the real part of each eigenvalue to be negative. And in the case of only real negative eigenvalues, all terms including cosines and sines are dropped from $G(t)$.

If we know the full history of the system instead of setting an initial value, the solution is given by

$$T(t) = \int_{-\infty}^t G(t-s)F(s)ds \tag{A6}$$

Step-response
When studying the response to a unit-step forcing, we first decompose the response:

$$T(t) = \int_0^t G(t-s) \cdot 1 \, ds = \sum_{n=1}^{K} \int_0^t G_n(t-s)ds \tag{A7}$$

where $G_1(t) = k_1 e^{pt} \cos qt$ and $G_2(t) = k_2 e^{pt} \sin qt$ describe the damped oscillatory responses, and $G_n(t) = k_n e^{\gamma_n t}$ describe responses associated with real negative eigenvalues. For the latter, we have the temperature responses

$$T_n(t) = \int_0^t G_n(t-s)ds = \int_0^t k_n e^{\gamma_n (t-s)} ds = S_n(1 - e^{\gamma_n t}) \tag{A8}$$

where $S_n = -k_n/\gamma_n$. For $G_1(t)$, we find the step-response

$$\begin{aligned} T_1(t) = \int_0^t G_1(t-s)ds &= \int_0^t k_1 e^{p(t-s)} \cos q(t-s) \, ds \\ &= k_1 \left[\frac{e^{pt}\left(p \cos qt + q \sin qt\right) - p}{p^2 + q^2} \right] \\ &= S_{osc1} - S_{osc1} e^{pt} \cos qt + \frac{k_1 q}{p^2 + q^2} e^{pt} \sin qt \\ &= S_{osc1} \left[1 - e^{pt} \left(\cos qt - \frac{q}{p} \sin qt \right) \right] & \text{(A9)} \end{aligned}$$

where $S_{osc1} = -\frac{k_1 p}{p^2 + q^2}$, and similarly for $G_2(t)$, we find

$$\begin{aligned} T_2(t) = \int_0^t G_2(t-s)ds &= \int_0^t k_2 e^{p(t-s)} \sin q(t-s) \, ds \\ &= k_2 \left[\frac{e^{pt}\left(p \sin qt - q \cos qt\right) + q}{p^2 + q^2} \right] \\ &= S_{osc2} - S_{osc2} e^{pt} \cos qt + \frac{k_2 p}{p^2 + q^2} e^{pt} \sin qt \\ &= S_{osc2} \left[1 - e^{pt} \left(\cos qt + \frac{p}{q} \sin qt \right) \right] & \text{(A10)} \end{aligned}$$

where $S_{osc2} = \frac{k_2 q}{p^2+q^2}$. The total step-response is therefore,

$$T(t) = S_{osc1}\left[1 - e^{pt}\left(\cos qt - \frac{q}{p}\sin qt\right)\right] + S_{osc2}\left[1 - e^{pt}\left(\cos qt + \frac{p}{q}\sin qt\right)\right] + \sum_{n=3}^{K} S_n(1 - e^{\gamma_n t}) \tag{A11}$$

Finally, we note that if the forcing was stepped up to a different value than 1, this value will be a factor included in $S_{osc1}, S_{osc2}, \ldots, S_n$.

Using step-response to derive other responses

If we have estimates of the parameters $S_{osc1}, S_{osc2}, \ldots, S_n, p, q, \gamma_n$, we find that $k_1 = \frac{-S_{osc1}(p^2+q^2)}{p}$, $k_2 = \frac{S_{osc2}(p^2+q^2)}{q}$, $k_n = -S_n\gamma_n$, which we can plug into the expression for $G(t)$ and compute the response to other forcings.

Supporting Information for "Testing linearity and comparing linear response models for global surface temperatures"

Hege-Beate Fredriksen[1,2], Kai-Uwe Eiselt[1]and Peter Good[3]

[1]UiT the Arctic University of Norway, Tromsø, Norway

[2]Norwegian Polar Institute, Tromsø, Norway

[3]Met Office Hadley Centre, Exeter, United Kingdom

Contents of this file

1. Figures S1 to S4
2. Tables S1 to S10

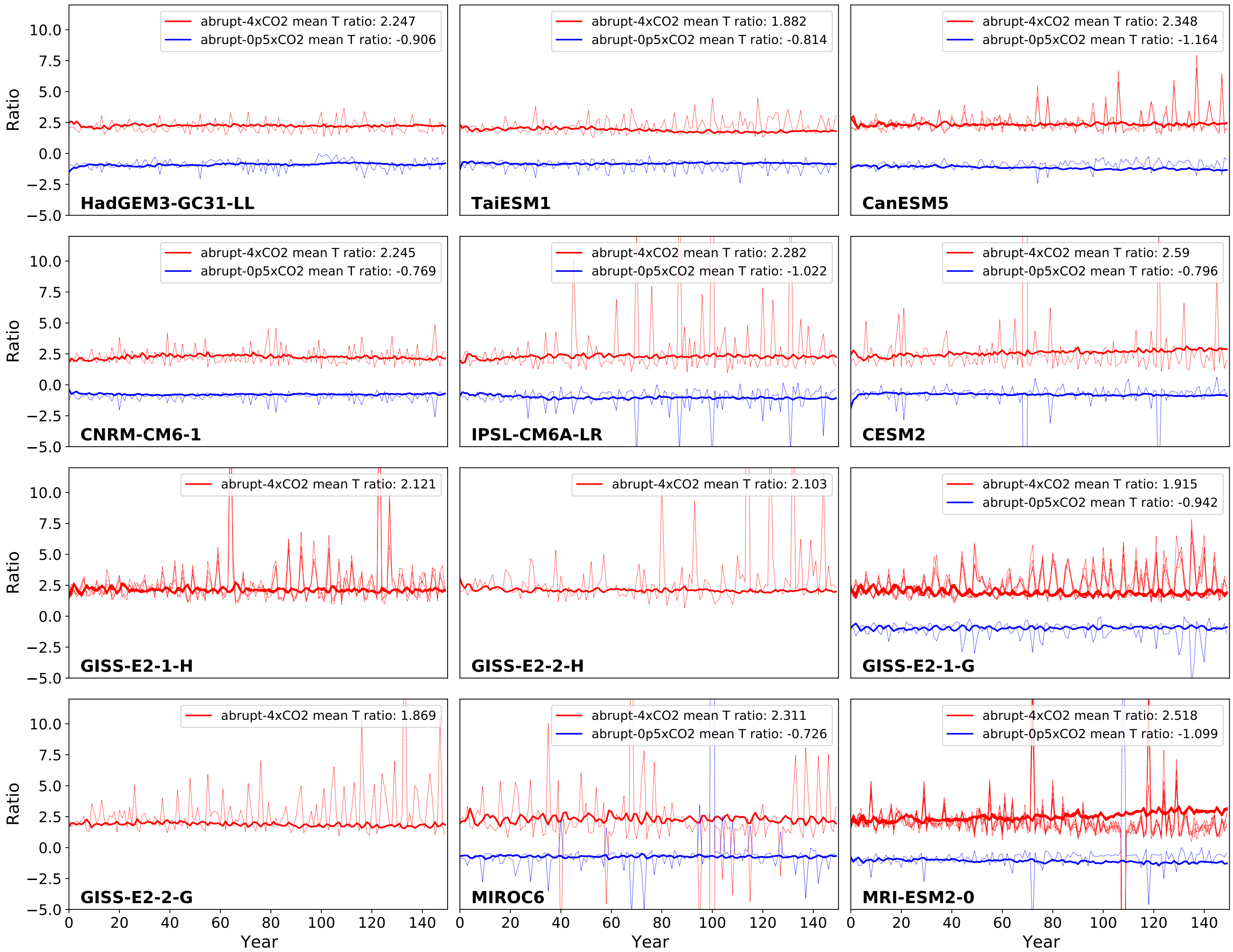

Figure S1. Ratios of T and N between abrupt-4xCO_2 (red)/abrupt-0p5xCO_2 (blue) experiments and abrupt-2xCO_2 experiments. Solid curves are T ratios and noisy thinner curves are N ratios.

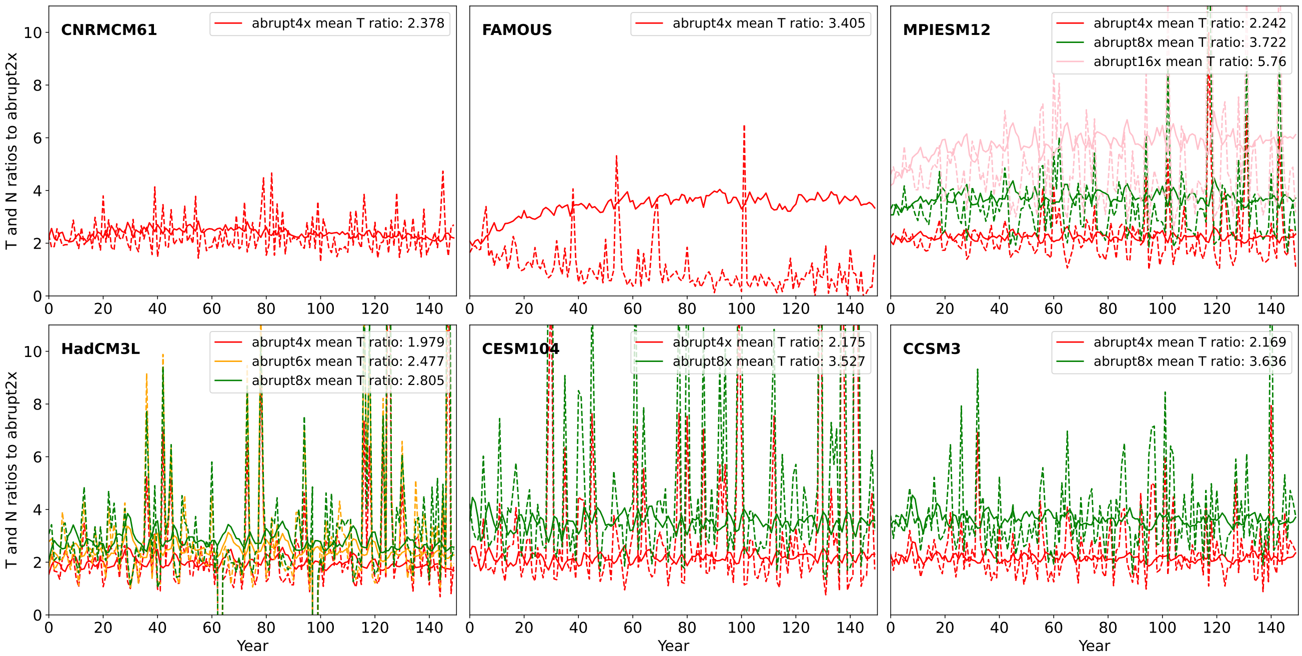

Figure S2. Ratios of T and N between abrupt4x (red)/abrupt6x (yellow)/abrupt8x (green)/abrupt16x (pink) experiments and abrupt2x experiments. Solid curves are T ratios and the dashed curves are N ratios. Only the first 150 years are used.

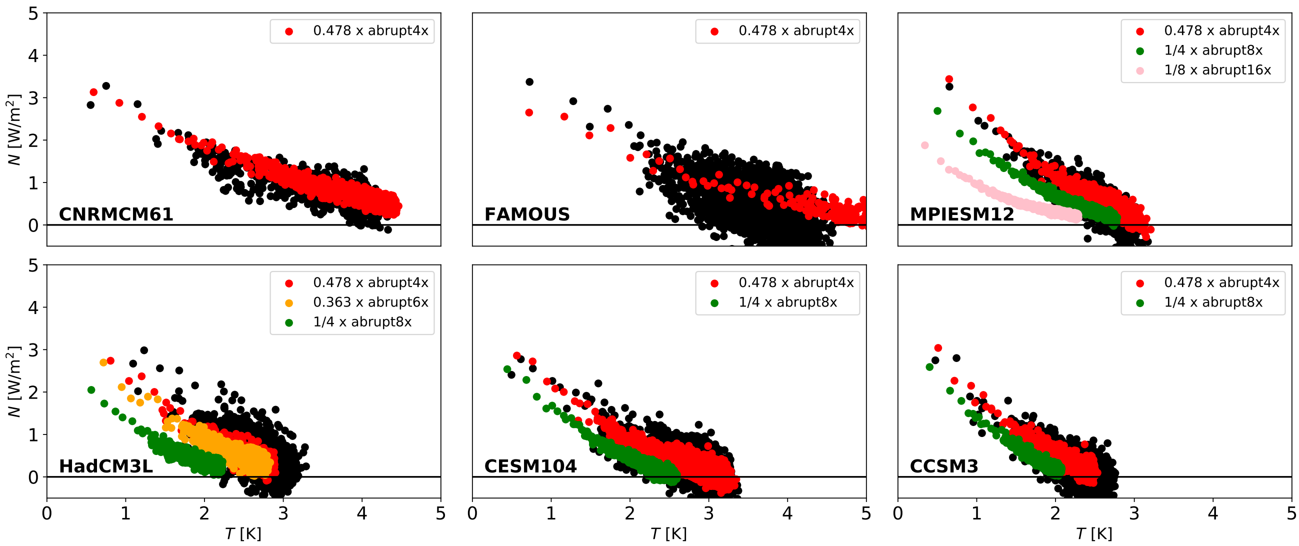

Figure S3. N and T both scaled to correspond to abrupt2x, using the scaling factors in the legends. Black dots are from the abrupt2x experiment, red is scaled abrupt4x, yellow is scaled abrupt6x, green is scaled abrupt8x, and pink is scaled abrupt16x.

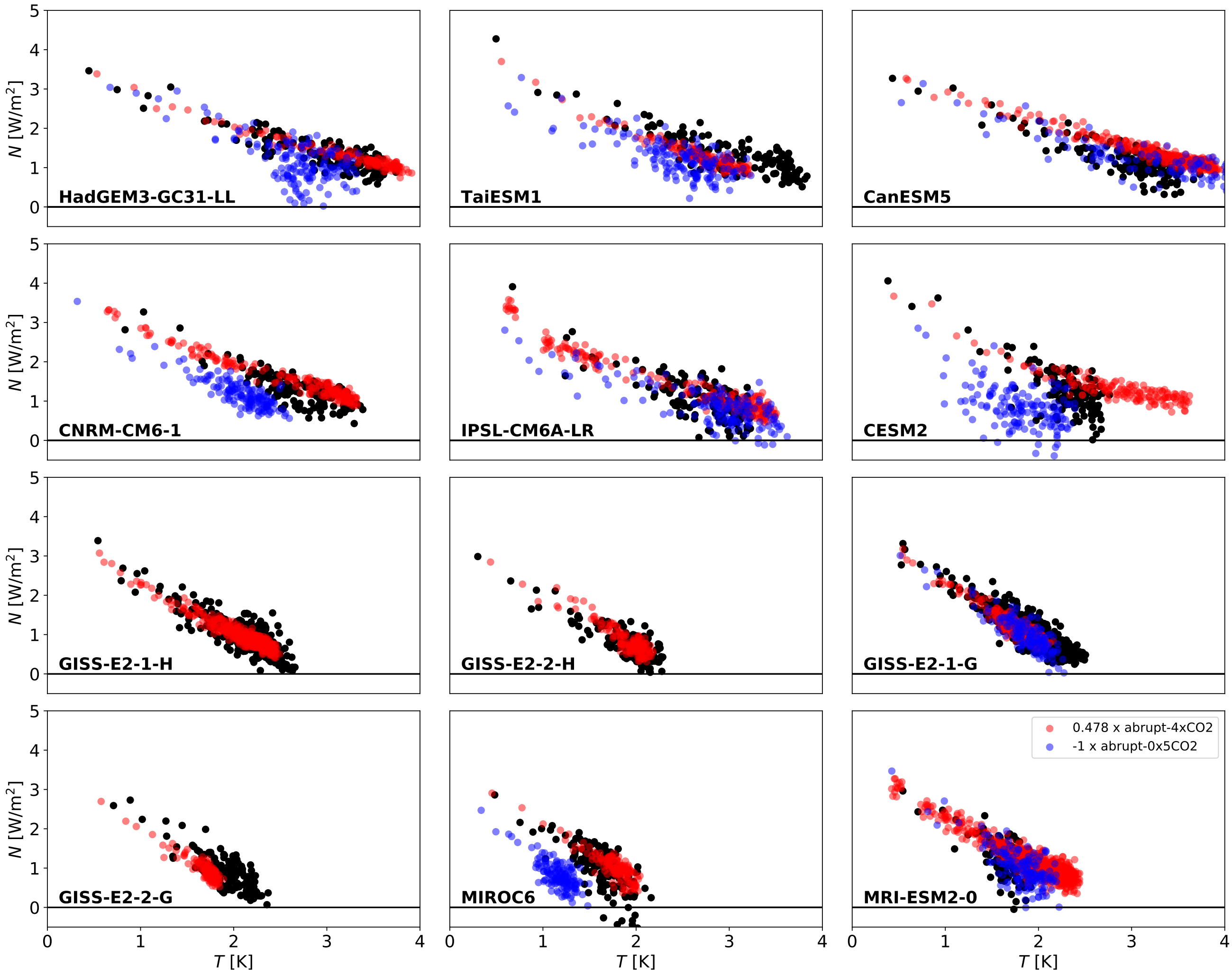

Figure S4. N and T both scaled to correspond to abrupt-2xCO2, using the same scaling factors for all models (see legend in the bottom right). The black circles are from the abrupt-2xCO2 experiment, red is the scaled abrupt-4xCO2 experiment and blue the scaled abrupt-0p5xCO2 experiment.

Table S1. Forcing ratios of abrupt-2xCO2 to abrupt-4xCO2 experiments, estimated from Gregory regressions of the first 5, 10, 20 and 30 years. The ensemble mean is the result of first averaging all model data for each year, and then perform regressions.

	5	10	20	30	Mean
CESM2	0.50	0.54	0.54	0.55	0.53
CNRM-CM6-1	0.51	0.50	0.50	0.54	0.51
CanESM5	0.48	0.48	0.49	0.49	0.49
GISS-E2-1-G	0.49	0.49	0.48	0.49	0.49
GISS-E2-1-H	0.49	0.51	0.49	0.52	0.50
GISS-E2-2-G	0.53	0.53	0.56	0.57	0.55
GISS-E2-2-H	0.48	0.51	0.49	0.45	0.48
IPSL-CM6A-LR	0.64	0.54	0.52	0.53	0.56
MIROC6	0.53	0.44	0.42	0.45	0.46
MRI-ESM2-0	0.50	0.49	0.46	0.47	0.48
TaiESM1	0.50	0.49	0.51	0.52	0.51
HadGEM3-GC31-LL	0.43	0.48	0.48	0.49	0.47
Ensemble mean	0.50	0.50	0.49	0.50	0.50
Mean of model results	0.51	0.50	0.50	0.51	0.50

Table S2. Forcing ratios of abrupt-2xCO2 to abrupt-0p5xCO2 experiments, estimated from Gregory regressions of the first 5, 10, 20 and 30 years. The ensemble mean is the result of first averaging all model data for each year, and then perform regressions.

	5	10	20	30	Mean
CESM2	-0.75	-1.11	-1.18	-1.28	-1.08
CNRM-CM6-1	-1.11	-1.16	-1.13	-1.22	-1.15
CanESM5	-1.06	-1.16	-1.11	-1.08	-1.10
GISS-E2-1-G	-1.03	-0.99	-1.00	-1.02	-1.01
IPSL-CM6A-LR	-1.53	-1.32	-1.40	-1.37	-1.41
MIROC6	-1.33	-1.14	-1.07	-1.14	-1.17
MRI-ESM2-0	-0.94	-0.95	-0.87	-0.86	-0.90
TaiESM1	-1.26	-1.28	-1.34	-1.36	-1.31
HadGEM3-GC31-LL	-1.05	-1.02	-0.98	-0.99	-1.01
Ensemble mean	-1.16	-1.15	-1.12	-1.15	-1.15
Mean of model results	-1.12	-1.12	-1.12	-1.15	-1.13

Table S3. Forcing ratios of longrunmip abrupt-2x to abrupt-Nx experiments, estimated from Gregory regressions of the first 5, 10, 20 and 30 years. The ensemble mean is the result of first averaging all model data for each year, and then perform regressions. If excluding FAMOUS for N=4, the model mean result is reduced to 0.46.

N = 4	5	10	20	30	Mean
MPIESM12	0.44	0.45	0.45	0.46	0.45
HadCM3L	0.31	0.54	0.55	0.52	0.48
FAMOUS	0.60	0.65	0.66	0.67	0.64
CNRMCM61	0.49	0.48	0.48	0.52	0.49
CESM104	0.38	0.41	0.45	0.45	0.42
CCSM3	0.48	0.49	0.41	0.43	0.45
Ensemble mean	0.46	0.50	0.51	0.53	0.50
Mean of model results	0.45	0.50	0.50	0.51	0.49

N = 6	5	10	20	30	Mean
HadCM3L	0.22	0.41	0.40	0.38	0.35

N = 8	5	10	20	30	Mean
MPIESM12	0.30	0.32	0.33	0.33	0.32
HadCM3L	0.22	0.41	0.40	0.38	0.35
CESM104	0.23	0.26	0.27	0.27	0.26
CCSM3	0.29	0.30	0.26	0.26	0.28
Ensemble mean	0.26	0.32	0.31	0.32	0.30
Mean of model results	0.26	0.32	0.31	0.31	0.30

N = 16	5	10	20	30	Mean
MPIESM12	0.22	0.24	0.24	0.25	0.24

	T_{4x}/T_{2x}	(T_{4x}/T_{2x}) / (F_{4x}/F_{2x})	T_{0p5x}/T_{2x}	(T_{0p5x}/T_{2x}) / (F_{0p5x}/F_{2x})
CESM2	2.57	1.37	-0.78	0.85
CNRM-CM6-1	2.23	1.15	-0.76	0.88
CanESM5	2.34	1.14	-1.15	1.27
GISS-E2-1-G	2.10	1.02	-0.95	0.96
GISS-E2-1-H	2.08	1.05	nan	nan
GISS-E2-2-G	1.86	1.01	nan	nan
GISS-E2-2-H	2.09	1.01	nan	nan
IPSL-CM6A-LR	2.27	1.26	-1.00	1.41
MIROC6	2.28	1.05	-0.72	0.84
MRI-ESM2-0	2.48	1.18	-1.08	0.98
TaiESM1	1.87	0.95	-0.81	1.06
HadGEM3-GC31-LL	2.24	1.05	-0.90	0.90
Mean	2.20	1.10	-0.91	1.02

Table S4. Mean ratios for CMIP6 models. The mean over 150 years are used, and the forcing ratios used are taken from the Mean columns in Tables S1 and S2.

	T_{4x}/T_{2x}	(T_{4x}/T_{2x}) / (F_{4x}/F_{2x})
MPIESM12	2.23	1.00
HadCM3L	1.96	0.94
FAMOUS	3.33	2.14
CNRMCM61	2.37	1.17
CESM104	2.16	0.91
CCSM3	2.16	0.98
Mean	2.18	1.00

Table S5. Mean ratios for LongRunMIP, using the first 150 years for estimation. The anomalous values for FAMOUS are omitted when computing the mean values. The forcing ratios are taken from the Mean column in Table S3.

Table S6. RMSE values for CMIP6 abrupt-4xCO2 experiments, part I.

model	member	two-exp	three-exp	two-exp + osc	% change1	% change2
ACCESS-CM2	r1i1p1f1	0.096	0.096	0.089	0.000	-7.261
ACCESS-ESM1-5	r1i1p1f1	0.127	0.114	0.111	-10.392	-2.389
ACCESS-ESM1-5	r2i1p1f1	0.104	0.101	0.102	-3.036	0.872
AWI-CM-1-1-MR	r1i1p1f1	0.125	0.118	0.118	-5.490	0.188
BCC-CSM2-MR	r1i1p1f1	0.092	0.076	0.078	-17.366	1.891
BCC-ESM1	r1i1p1f1	0.075	0.064	0.067	-13.916	4.082
CAMS-CSM1-0	r1i1p1f1	0.083	0.071	0.071	-13.784	0.015
CAMS-CSM1-0	r2i1p1f1	0.087	0.084	0.084	-3.756	0.620
CAS-ESM2-0	r1i1p1f1	0.097	0.088	0.085	-9.554	-3.548
CESM2	r1i1p1f1	0.088	0.075	0.078	-14.594	4.346
CESM2-FV2	r1i1p1f1	0.131	0.122	0.116	-7.310	-4.958
CESM2-WACCM	r1i1p1f1	0.086	0.081	0.079	-6.122	-3.109
CESM2-WACCM-FV2	r1i1p1f1	0.118	0.115	0.108	-2.287	-6.181
CIESM	r1i1p1f1	0.111	0.096	0.091	-13.337	-5.750
CMCC-CM2-SR5	r1i1p1f1	0.153	0.152	0.153	-0.812	0.661
CMCC-ESM2	r1i1p1f1	0.167	0.162	0.165	-3.219	2.117
CNRM-CM6-1	r1i1p1f2	0.111	0.097	0.097	-13.008	-0.048
CNRM-CM6-1-HR	r1i1p1f2	0.111	0.079	0.076	-28.670	-3.629
CNRM-ESM2-1	r1i1p1f2	0.120	0.120	0.115	0.000	-4.169
CNRM-ESM2-1	r2i1p1f2	0.101	0.101	0.096	0.000	-4.404
CNRM-ESM2-1	r3i1p1f2	0.096	0.096	0.094	0.000	-2.530
CanESM5	r1i1p1f1	0.113	0.093	0.096	-17.727	4.128
CanESM5	r1i1p2f1	0.117	0.092	0.092	-21.178	-0.593
E3SM-1-0	r1i1p1f1	0.144	0.125	0.140	-13.432	12.680
EC-Earth3	r3i1p1f1	0.153	0.147	0.141	-4.366	-3.906
EC-Earth3	r8i1p1f1	0.134	0.134	0.133	-0.136	-1.099
EC-Earth3-AerChem	r1i1p1f1	0.138	0.137	0.134	-0.844	-2.366
EC-Earth3-CC	r1i1p1f1	0.142	0.139	0.142	-2.506	2.150
EC-Earth3-Veg	r1i1p1f1	0.138	0.134	0.136	-2.425	1.091
FGOALS-f3-L	r1i1p1f1	0.129	0.121	0.125	-6.581	3.522
FGOALS-f3-L	r2i1p1f1	0.128	0.122	0.126	-4.244	3.469
FGOALS-f3-L	r3i1p1f1	0.115	0.108	0.109	-6.213	0.413
FGOALS-g3	r1i1p1f1	0.073	0.072	0.072	-1.265	0.290
GFDL-CM4	r1i1p1f1	0.113	0.108	0.107	-4.819	-0.520
GFDL-ESM4	r1i1p1f1	0.090	0.084	0.090	-5.993	6.326

Table S7. RMSE values for CMIP6 abrupt-4xCO2 experiments, part II.

model	member	two-exp	three-exp	two-exp + osc	% change1	% change2
GISS-E2-1-G	r102i1p1f1	0.147	0.146	0.134	-0.275	-8.424
GISS-E2-1-G	r1i1p1f1	0.129	0.129	0.119	-0.239	-7.836
GISS-E2-1-G	r1i1p3f1	0.158	0.157	0.150	-0.306	-4.785
GISS-E2-1-G	r1i1p5f1	0.185	0.179	0.171	-3.199	-4.465
GISS-E2-1-H	r1i1p1f1	0.122	0.112	0.112	-7.558	-0.109
GISS-E2-1-H	r1i1p3f1	0.123	0.121	0.123	-1.764	1.795
GISS-E2-1-H	r1i1p5f1	0.141	0.129	0.131	-8.242	0.850
GISS-E2-2-G	r1i1p1f1	0.103	0.101	0.101	-2.055	-0.028
GISS-E2-2-H	r1i1p1f1	0.094	0.087	0.086	-7.596	-0.242
HadGEM3-GC31-LL	r1i1p1f3	0.109	0.098	0.099	-9.806	0.502
HadGEM3-GC31-MM	r1i1p1f3	0.143	0.092	0.089	-35.752	-3.257
ICON-ESM-LR	r1i1p1f1	0.158	0.140	0.130	-11.601	-6.992
IITM-ESM	r1i1p1f1	0.106	0.099	0.102	-5.885	2.634
INM-CM4-8	r1i1p1f1	0.068	0.057	0.063	-15.632	10.321
INM-CM5-0	r1i1p1f1	0.087	0.077	0.079	-11.543	1.974
IPSL-CM5A2-INCA	r1i1p1f1	0.123	0.114	0.114	-7.165	-0.060
IPSL-CM6A-LR	r1i1p1f1	0.150	0.122	0.119	-18.672	-2.691
KIOST-ESM	r1i1p1f1	0.115	0.108	0.092	-6.742	-14.876
MIROC-ES2L	r1i1p1f2	0.159	0.155	0.156	-2.856	0.730
MIROC6	r1i1p1f1	0.167	0.164	0.163	-1.915	-0.269
MPI-ESM-1-2-HAM	r1i1p1f1	0.108	0.089	0.089	-17.801	0.455
MPI-ESM1-2-HR	r1i1p1f1	0.079	0.076	0.078	-3.200	2.185
MPI-ESM1-2-LR	r1i1p1f1	0.129	0.119	0.118	-7.906	-1.435
MRI-ESM2-0	r10i1p1f1	0.118	0.116	0.099	-1.781	-14.644
MRI-ESM2-0	r13i1p1f1	0.101	0.099	0.088	-2.800	-10.852
MRI-ESM2-0	r1i1p1f1	0.103	0.102	0.085	-0.614	-16.470
MRI-ESM2-0	r1i2p1f1	0.111	0.109	0.083	-2.222	-23.718
MRI-ESM2-0	r4i1p1f1	0.104	0.101	0.097	-2.958	-4.137
MRI-ESM2-0	r7i1p1f1	0.111	0.101	0.094	-9.111	-7.172
NESM3	r1i1p1f1	0.104	0.088	0.088	-14.984	0.006
NorCPM1	r1i1p1f1	0.091	0.091	0.090	0.000	-0.935
NorESM2-LM	r1i1p1f1	0.175	0.175	0.162	0.000	-7.727
NorESM2-MM	r1i1p1f1	0.172	0.172	0.172	-0.000	-0.197
SAM0-UNICON	r1i1p1f1	0.127	0.127	0.111	0.000	-13.109
TaiESM1	r1i1p1f1	0.145	0.117	0.103	-19.762	-11.485
UKESM1-0-LL	r1i1p1f2	0.111	0.102	0.108	-8.126	5.738

Table S8. RMSE values for CMIP6 abrupt-2xCO2 experiments

model	member	two-exp	three-exp	two-exp + osc	% change1	% change2
CESM2	r1i1p1f1	0.096	0.096	0.096	0.000	-0.029
CNRM-CM6-1	r1i1p1f2	0.106	0.106	0.104	-0.046	-1.814
CanESM5	r1i1p2f1	0.117	0.115	0.113	-1.786	-1.919
GISS-E2-1-G	r102i1p1f1	0.144	0.144	0.143	0.000	-0.376
GISS-E2-1-G	r1i1p1f1	0.140	0.140	0.136	0.000	-3.105
GISS-E2-1-G	r1i1p3f1	0.164	0.158	0.153	-3.483	-3.061
GISS-E2-1-G	r1i1p5f1	0.180	0.180	0.179	-0.167	-0.606
GISS-E2-1-H	r1i1p1f1	0.121	0.120	0.119	-0.310	-0.914
GISS-E2-1-H	r1i1p5f1	0.143	0.139	0.139	-2.329	0.022
GISS-E2-2-G	r1i1p1f1	0.116	0.116	0.112	-0.219	-3.268
GISS-E2-2-H	r1i1p1f1	0.085	0.081	0.080	-4.737	-1.617
HadGEM3-GC31-LL	r1i1p1f3	0.094	0.094	0.094	-0.000	-0.095
IPSL-CM6A-LR	r1i1p1f1	0.132	0.127	0.132	-3.902	3.989
MIROC6	r1i1p1f1	0.158	0.158	0.158	-0.049	-0.151
MRI-ESM2-0	r1i1p1f1	0.105	0.105	0.103	0.000	-1.220
TaiESM1	r1i1p1f1	0.111	0.111	0.097	-0.000	-12.556

Table S9. RMSE values for CMIP6 abrupt-0p5xCO2 experiments

model	member	two-exp	three-exp	two-exp + osc	% change1	% change2
CESM2	r1i1p1f1	0.108	0.107	0.107	-1.232	-0.015
CNRM-CM6-1	r1i1p1f2	0.099	0.098	0.092	-1.013	-6.314
CanESM5	r1i1p2f1	0.104	0.104	0.099	-0.085	-4.829
GISS-E2-1-G	r1i1p1f1	0.120	0.119	0.119	-0.775	-0.067
HadGEM3-GC31-LL	r1i1p1f3	0.174	0.166	0.103	-4.880	-37.868
IPSL-CM6A-LR	r1i1p1f1	0.137	0.119	0.109	-13.440	-7.981
MIROC6	r1i1p1f1	0.074	0.074	0.070	-0.012	-4.546
MRI-ESM2-0	r1i1p1f1	0.100	0.100	0.098	-0.000	-1.767
TaiESM1	r1i1p1f1	0.100	0.094	0.098	-5.397	3.457

Table S10. RMSE values for LongRunMIP experiments

model	exp	two-exp	three-exp	two-exp + osc	% change1	% change2
MPIESM12	abrupt2x	0.124	0.119	0.119	-4.066	0.012
MPIESM12	abrupt4x	0.143	0.132	0.132	-8.095	0.026
MPIESM12	abrupt8x	0.146	0.114	0.114	-22.206	0.188
MPIESM12	abrupt16x	0.171	0.097	0.123	-43.441	27.638
HadCM3L	abrupt2x	0.179	0.175	0.174	-2.113	-0.403
HadCM3L	abrupt4x	0.125	0.117	0.118	-6.782	0.811
HadCM3L	abrupt6x	0.123	0.117	0.116	-5.587	-0.104
HadCM3L	abrupt8x	0.128	0.124	0.125	-3.127	1.440
FAMOUS	abrupt2x	0.180	0.177	0.177	-1.652	-0.171
FAMOUS	abrupt4x	0.215	0.142	0.143	-33.919	0.778
CNRMCM61	abrupt2x	0.111	0.107	0.106	-3.359	-1.105
CNRMCM61	abrupt4x	0.117	0.100	0.100	-14.394	0.002
CESM104	abrupt2x	0.153	0.145	0.134	-4.755	-7.499
CESM104	abrupt4x	0.168	0.133	0.132	-20.924	-0.396
CESM104	abrupt8x	0.222	0.168	0.156	-24.219	-7.707
CCSM3	abrupt2x	0.092	0.091	0.091	-1.229	-0.452
CCSM3	abrupt4x	0.102	0.096	0.094	-5.082	-2.096
CCSM3	abrupt8x	0.111	0.086	0.086	-22.644	0.028
IPSLCM5A	abrupt4x	0.132	0.107	0.107	-18.925	0.007
HadGEM2	abrupt4x	0.133	0.104	0.104	-21.529	0.357
GISSE2R	abrupt4x	0.093	0.080	0.079	-13.923	-0.800
ECHAM5MPIOM	abrupt4x	0.195	0.180	0.178	-7.719	-1.045