PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Cite this article: Bastiaansen R, Ashwin P, von der Heydt AS. 2023 Climate response and sensitivity: time scales and late tipping points. *Proc. R. Soc. A* **479**: 20220483. https://doi.org/10.1098/rspa.2022.0483

Received: 15 July 2022
Accepted: 29 November 2022

Subject Areas:
climatology, chaos theory, differential equations

Keywords:
climate sensitivity, nonlinear dynamics, tipping points, energy balance model

Author for correspondence:
Peter Ashwin
e-mail: P.Ashwin@exeter.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6350222.

THE ROYAL SOCIETY PUBLISHING

Climate response and sensitivity: time scales and late tipping points

Robbin Bastiaansen[1,2], Peter Ashwin[3] and Anna S. von der Heydt[1]

[1]Department of Physics and IMAU, Utrecht University, Utrecht, The Netherlands
[2]Mathematical Institute, Utrecht University, Utrecht, The Netherlands
[3]Department of Mathematics and Statistics, University of Exeter, Exeter EX4 4QF, UK

RB, 0000-0002-5628-8360; PA, 0000-0001-7330-4951; ASvdH, 0000-0002-5557-3282

Climate response metrics are used to quantify the Earth's climate response to anthropogenic changes of atmospheric CO_2. Equilibrium climate sensitivity (ECS) is one such metric that measures the equilibrium response to CO_2 doubling. However, both in their estimation and their usage, such metrics make assumptions on the linearity of climate response, although it is known that, especially for larger forcing levels, response can be nonlinear. Such nonlinear responses may become visible immediately in response to a larger perturbation, or may only become apparent after a long transient period. In this paper, we illustrate some potential problems and caveats when estimating ECS from transient simulations. We highlight ways that very slow time scales may lead to poor estimation of ECS even if there is seemingly good fit to linear response over moderate time scales. Moreover, such slow processes might lead to late abrupt responses (late tipping points) associated with a system's nonlinearities. We illustrate these ideas using simulations on a global energy balance model with dynamic albedo. We also discuss the implications for estimating ECS for global climate models, highlighting that it is likely to remain difficult to make definitive statements about the simulation times needed to reach an equilibrium.

1. Introduction

The central question as to how the climate is likely to change as a function of anthropogenic CO_2 emissions can be posed as 'How does an observation of the climate system respond to changes in its radiative forcing induced by changes in atmospheric CO_2?'. This question has been studied in various ways for over a century [1,2], although efforts to answer it became more intense and in-depth over the last decades. Among early efforts was the pioneering work by Charney *et al.* in 1979, who made the first estimates of expected equilibrium warming after doubling of atmospheric CO_2 (while keeping vegetation and land ice fixed at present-day values) using a numerical global climate model (GCM) [3]. This metric has later been named the equilibrium climate sensitivity (ECS) and is still widely used. Since then, researchers have developed a number of different metrics that measure climate response to different scenarios of anthropogenic change in CO_2 and have incorporated information from other sources besides computer models, including historical observations and data from palaeoclimate records. Recently, these efforts were summarized in an assessment of the World Climate Research Programme [4] that synthesized different quantifications of climate response using these different lines of evidence and led to the headline that the Earth's ECS is likely between 2.6 K and 3.9 K.

One of the hurdles for this assessment was the variety of definitions of (the quantification of) climate sensitivity—and ECS especially—in the literature. The root of this problem can be attributed to the lack of data on equilibrium climate states or detailed long-term transient data. This can be due to low time resolutions in proxy data, lack of observational data or insufficient computing power to equilibrate modern GCMs. Consequently, equilibrium properties need to be estimated from incomplete datasets, leading to many slightly different ways to quantify climate sensitivity. Common to them all, however, is the need to extrapolate long-term dynamics from data on shorter time scales. In this paper, we discuss this extrapolation process, focusing on estimates of ECS using (idealized) experiments in climate models for the sake of mathematical simplicity. Of particular interest here is the exploration of linear and nonlinear, dynamics that can emerge in multi-scale dynamical systems, and their problematic effects on extrapolation.

The common way to obtain estimates of ECS in climate models involves the use of extrapolation and regression methods on non-equilibrated transient simulations—typically of 150-year long runs. Values for ECS obtained in this way are now often referred to as the *effective climate sensitivity* [5] signalling that it might not encompass all long-term climate change. Although there are many different ways to perform such extrapolation, they are usually based on linear concepts and frameworks. A recent review [6] of climate sensitivity highlighted that it is a key challenge to study the limits of such linear frameworks. Here, we will investigate these limits and in the process highlight the trade-offs that need to be made when designing experiments to quantify ECS: in order to measure a clear signal of warming in relation to the noise of natural variations, large perturbations are desirable but precisely in the case of larger perturbations the nonlinear behaviour becomes important and linear frameworks break down.

One of the most important tools to study past and future climate change are the GCMs used in the Coupled Model Intercomparison Projects (CMIP, e.g. [7]), because they provide a globally complete and detailed representation of the climate state while (approximately) satisfying the physical laws. However, specifically for these large models there is no way to determine whether a model really has arrived in the linear regime near an equilibrium, or even if such an equilibrium exists. In this paper we explore some simple conceptual examples of the potential nonlinear dynamics of the climate. We also make a number of observations that we hope illuminate some of the limitations of linear frameworks:

(i) We highlight cases where there may be strong dependence on the climate background state and the forcing levels.
(ii) We highlight examples where there may be a *good fit* to transient data but *poor extrapolation* preventing an accurate estimation of the ECS.

(iii) We show that nonlinear systems can have *slow tipping points*. When these are crossed the tipping dynamics play out on slow time scales, and it can take arbitrarily long times before nonlinear and/or asymptotic behaviour is observed.
(iv) We demonstrate how in the presence of multiple time scales with nonlinear feedbacks a *late tipping* can occur in which fast processes suddenly dominate after arbitrarily long slow transient behaviour. This highlights the potential for slow and/or late tipping points to be particular obstructions to estimating ECS.

The rest of this paper is organized as follows: in the remainder of this section we discuss in general the response of a nonlinear system to forcing. In §2, we consider the equilibrium response and ECS of the climate system in terms of limiting behaviour. Moreover, we point out the challenges that arise when estimating those from short time series, highlighting the trade-offs that emerge in terms of perturbation size and required simulation time. In §3, we examine the nonlinear effects that may appear as a result of climate dynamics on multiple time scales, including *slow* tipping which may in turn lead to *late but rapid* tipping. We illustrate these effects using multi-scale global energy balance models (GEBMs) with dynamic albedo and/or chaotic variability and an example from a LongRunMIP abrupt8xCO_2 run [8]. Finally, we briefly discuss these results, and the influence of time-varying forcing on estimation of climate response and sensitivity in §4.

(a) Response of nonlinear models to forcing

Consider a notional state of the climate system $y(t)$ for $t > t_0$ that evolves in response to various (unknown) forcings, with a partially known initial state y_0 and an input of atmospheric CO_2 generating a radiative forcing $\Delta F(t)$ that is specified for $t > t_0$. We write this climate state at time t as $y(t) = Y_t(y_0, t_0, \Delta F)$, where Y_t is an evolution operator that evolves forward the initial state y_0 (at time t_0) up to time t according to a climate model Y with (possibly time-dependent) radiative forcing ΔF.

Given a scalar observable $\mathcal{O}$ that maps the full climate state $y(t)$, the response clearly depends on the choice of observable $\mathcal{O}$, the choice of model Y, the forcing ΔF experienced by the system, the initial climate state y_0 at time t_0 and the time moment $t > t_0$ of interest. At the level of a single initial state y_0 starting at t_0 of which we have perfect knowledge and subject to deterministic forcing ΔF, the response in observable $\mathcal{O}$ at time $t > t_0$ is the difference in its value at times t_0 and t, i.e.

$$\mathcal{R}_{\mathcal{O},Y}(t; t_0, y_0; \Delta F) = \mathcal{O}(Y_t(y_0, t_0, \Delta F)) - \mathcal{O}(Y_{t_0}(y_0, t_0, 0)). \tag{1.1}$$

This corresponds to a *two-point* response in the terminology of Ashwin & von der Heydt [9]. As this is often the easiest response type to think about mathematically (and extensions to other types are possible albeit more technical), it is this response type we will be referring to throughout this paper. However, often we are interested not in specific trajectories but rather in the distribution of possible responses for a probability distribution μ_0 of initial states and forcing ΔF. In this case, we write the response as

$$\mathcal{R}_{\mathcal{O},Y}(t; t_0, \mu_0; \Delta F) = \mathcal{O}(Y_t(\mu_0, t_0, \Delta F)) - \mathcal{O}(Y_{t_0}(\mu_0, t_0, 0)). \tag{1.2}$$

This corresponds to a *distributional* response, namely it is a random variable with some distribution determined by the 'pushforward' of the initial probability distribution μ_0 by the dynamics. Furthermore, there are different interpretations of (1.2), depending on the choice of probability function. These include:

— A *physical measure* on a climate attractor [9,10]. This can be an observable measured in a long palaeoclimate time series, or an observable in a model, where the attractor is (partly) known from the underlying model equations.
— An *ensemble* of initial conditions that are thought to sample subgrid processes in a model (or observational data).

— An *empirical measure* for a finite segment of trajectory, i.e. a choice of states on $\{Y_t(y_0, t_0, 0) : t \in [t_0, t_1]\}$ over some finite interval with $t_0 < t_1$, with equal weight to any given time instant. Such a measure can be approximated from a finite length time series of a palaeoclimate record.

We note that in the case that μ_0 is a physical measure, the empirical measure as defined above will converge to that same distribution: for a more precise definition of a physical measure, see for example [11,12]. If there are multiple attractors then there can be several physical measures, and typical initial conditions converge to one of these depending which basin of attraction they are in.

2. Equilibrium response and ECS as limiting behaviour

While the response on any time scale can be relevant, the asymptotic or equilibrium, response as $t \to \infty$ is a particularly important way of characterizing any model. This response is typically easy to analyse and understand in simple models. Taking the limit $t \to \infty$ of (1.1), the equilibrium response is

$$\lim_{t\to\infty} \mathcal{R}_{O,Y}(t; t_0, y_0; \Delta F). \tag{2.1}$$

Of course, this begs the question of whether the limit exists. In particular, one cannot expect such a limit to hold for any forcing ΔF. For instance, if the forcing specifies uninhibited and constant emission of greenhouse gases, the climate system will not evolve to any equilibrium. Hence it makes sense to limit ourselves to forcing scenarios that have constant forcing levels as $t \to \infty$ (i.e. $\Delta F(t) \to \Delta F_*$ as $t \to \infty$). In practical model studies of ECS, often the forcing is just taken as a constant throughout the whole simulation.

Of particular interest is the equilibrium response to an instantaneous and abrupt doubling of atmospheric CO_2, which we indicate by the forcing $\Delta F_{\text{abrupt2xCO}_2}$. Then, the ECS is defined as the response of global mean surface temperature (GMST) to such forcing, i.e.

$$\text{ECS}(y_0) := \lim_{t\to\infty} \mathcal{R}_{\text{GMST},Y}(t; t_0, y_0; \Delta F_{\text{abrupt2xCO}_2}). \tag{2.2}$$

Even for such idealized forcing, such a limit may not be well-defined. In any but the simplest models, the asymptotic climate state will have stationary internal variability, for which the limit of the two-point response is not well-defined without first averaging for long enough that any internal variability is averaged out. In such cases, a distributional response may have a well-defined limit, although it can happen that even these do not converge in cases where there is non-ergodic behaviour [13].

It is difficult to say anything definitive about the convergence of climate response in state-of-the-art GCMs. These models are numerical representations of the underlying physical equations, which have been developed to include many physical processes and ever-improving parametrizations of subgrid scale processes; they are very high-dimensional and complex. We do not have access to the attractors of these models and so cannot exclude the possibility of poor or no convergence. These models are roughly calibrated only by assessing how well they can reproduce the *present day* climate, including the historical period. However, in practice, reaching the true equilibrium may also be less relevant with such a model; the physical state of the climate system after a few centuries or even millennia could be difficult to predict anyway because of incomplete knowledge of the initial state y_0, model details and forcing. For these reasons, a pragmatic *effective climate sensitivity* [5,14] is often taken, in which response over a few centuries or millennia is taken, ignoring dynamics on longer time scales. However, we focus here on cases where the limit in (2.2) is well-defined.

(a) Background state, forcing scenario and ECS

In (2.2), it is clear that the ECS depends on the initial condition y_0 or *background state* where the latter refers to the initial climate attractor. However, often ECS is given without explicitly stating initial conditions. This can lead to ambiguity about what is meant by ECS when comparing simulations of current and palaeoclimates. Because of the possibility of multi-stability of the climate system, there may be multiple climate states possible for the same CO_2-level. In physical terms, the dependence on the background state originates from feedback processes that change as the forcing is applied [10], necessitating a proper communication of the background state considered when computing the ECS.

Further, in the definition of ECS (2.2) a doubling of atmospheric CO_2 is given as forcing scenario. However, in practice, ECS is often used as a measure of temperature increase *per* CO_2 doubling. So by assuming linearity of the climate response to forcing levels, ECS is employed to estimate warming for other CO_2 forcing levels. Specifically, for an abrupt 2^γxCO_2 forcing, an assumption of linear response would mean that warming of γ times the ECS is expected:

$$\lim_{t\to\infty} \mathcal{R}_{\mathrm{GMST},\gamma}(t; t_0, y_0; \Delta F_{\mathrm{abrupt}2^\gamma \mathrm{xCO}_2}) = \gamma\ \mathrm{ECS}(y_0). \tag{2.3}$$

However, this clearly ignores smooth nonlinear corrections that become significant for γ large enough. Moreover there may even be discontinuous corrections for large enough γ, if a tipping point is crossed.

We suggest it would be helpful to have greater clarity on when and how this linearity assumption breaks down in specific cases. For example, it has been verified that this linearity assumption does break down in GCMs for large enough perturbations: palaeoclimate simulations with a wide range of CO_2-concentrations suggest such linearity can be broken [15] and multi-millennial experiments in the model intercomparison project LongRunMIP [8] also show deviations from linearity; abrupt4xCO_2 experiments can lead to more than twice the warming of an abrupt2xCO_2 experiment in the same GCM. Further, abrupt8xCO_2 experiments led to less than twice the warming of an abrupt4xCO_2 experiment. Hence, the usage of ECS as a linear predictor for warming based on CO_2 levels can easily lead to over- or underestimations of warming.

(b) Challenges to estimating ECS from time series

It is computationally expensive to run state-of-the-art GCMs and, in principle, millennial length simulations may be needed to get close to equilibrium (e.g. the LongRunMIP [16]). Because there exists variability on many time scales and spatial feedback patterns in these models, there is no *a priori* method to determine when or indeed whether a nonlinear model has reached equilibrium. This means that the equilibrium response of a climate model cannot be directly found from time evolution of the model; instead, one needs to derive and extrapolate the equilibrium properties of the model from possibly relatively short transient data.

In general, estimation of ECS for a model (such as a GCM) involves four steps:

(i) *Design* of an experimental protocol (initial conditions, forcing levels, simulation time, ensemble of runs of the GCM);
(ii) *Selection* of a time period for fitting;
(iii) *Fitting* of transient observable data to a less complex model;
(iv) *Extrapolation* to derive equilibrium properties from the fitted model.

Many different protocols have been used—see e.g. [14, table 2] that lists 11 different methodologies. However, the most common standard for estimating ECS uses a technique by Gregory *et al.* [17]. Typically, a single abrupt CO_2-forcing experiment is run (starting from pre-industrial forcing levels, and it is standard to use an abrupt 4xCO_2 forcing) for some years (150 years is the benchmark for CMIP6 models). The transient data on change in the yearly and globally averaged observables near-surface-temperature ΔT and top-of-atmosphere radiative

imbalance ΔN is fitted to the linear model $\Delta N = \lambda \Delta T + f$. Then, equilibrium warming ΔT^* is estimated setting $\Delta N = 0$ in this linear model (since, in equilibrium, there should be radiative balance), yielding $\Delta T^*_{\text{est}} = -\lambda^{-1} f$. It is clear that GCMs are not well-approximated by this simple linear model over all time scales; because climate feedback processes operate at quite different time scales, ΔN and ΔT will have a nonlinear relationship that has non-zero curvature over the course of a long simulation, and the linear relationship only holds approximately for certain time intervals [8,18–20]. Better fits to the response over all the time scales can be found by considering a combination of several linearly decaying modes, i.e. by viewing the climate system as a combination of linear processes with quite different time scales [18,21–24].

Other protocols use results from the literature of linear response theory directly [25–33]. That is, in relative generality, the response (of an observable O) in the linear regime of a (nonlinear) system to a forcing can be characterized via a (causal linear observational) Green's function $G^{[O]}(t)$. Specifically, the yearly and globally (and ensemble) average near-surface-temperature increase ΔT at time t under a certain forcing scenario ΔF is given by the relation

$$\Delta T(t) = (G^{[T]} * \Delta F)(t) := \int_0^t G^{[T]}(s) \Delta F(t-s)\, \mathrm{d}s.$$

Using this relationship, transient data can be used to estimate the Green's function $G^{[T]}$ from which the equilibrium response can be extrapolated—which can be done through fitting to some prescribed function (typically a sum of decaying exponential functions) or through a discrete Fourier transform algorithm.

For all the fitting and extrapolation protocols, the optimal choices in the protocol are not always obvious as certain trade-offs need to be made:

(i) The simulation time needs to be as long as possible to ensure (a) we are in the linear response of the final equilibrium state and (b) fluctuations caused by natural variability can be averaged out. However, long simulations for GCMs are computationally expensive and even these will not be able to detect slow time scales beyond the length of simulation time.
(ii) A large ensemble and/or a long time period for fitting needs to be chosen to reduce noise caused by internal variability. However, each additional ensemble member increases the simulation effort and still the time period for fitting needs to start as late as possible to maximize the chance of being in a linear regime.
(iii) The perturbation needs to be as large as possible to maximize the signal-to-noise ratio for the fitting procedure. However, large perturbations may result in nonlinear effects, including tipping into different climate states.

Figure 1 illustrates two important trade-offs between perturbation size and integration time. In particular, the figure highlights the need to find a 'Goldilocks Zone' where the perturbation is neither too small nor too big. Examples of these trade-offs in a nonlinear setting using an conceptual energy balance model are discussed within §3.

(c) Slow linear responses and ECS

We start by illustrating some challenges that already arise in the linear response regime of a model. In such settings, extrapolation can be difficult if the time scale of the slowest response exceeds the length of time series available. To illustrate this, we now consider the evolution of a linear observable O of a finite M-dimensional linear system. In the absence of repeated eigenvalues, the Green's function will be a sum of exponential functions (with exponents being the eigenvalues) with the following functional form:

$$G^{[O]}(t) = \begin{cases} \sum_{j=1}^{M} \beta_j^{[O]} \mathrm{e}^{\lambda_j t} & \text{if } t \geq 0 \\ 0 & \text{if } t < 0 \end{cases}, \tag{2.4}$$

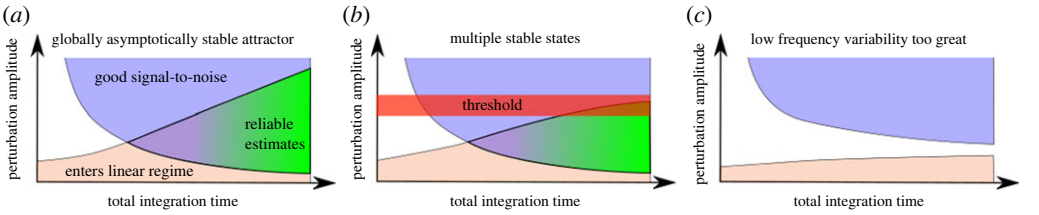

Figure 1. Schematic diagrams illustrating trade-offs between perturbation amplitude and integration time when computing ECS on perturbing a linearly stable state of a nonlinear climate model. The light blue regions illustrate the trade-off needed to give a good signal-to-noise ratio of the estimate of ECS. The pink region illustrates the trade-off needed to ensure the system has entered the linear regime. The green 'Goldilocks zone' shows points where accurate prediction of ECS is possible. (*a*) illustrates a case where the state is globally stable while (*b*) shows a case where a large enough perturbation (above the red bar) pushes the system out of the linear regime—perturbations above this may in principle give super-long transients and/or convergence to another stable state. Finally, (*c*) shows a case where accurate estimation of ECS is not possible—this may be because low frequency variability is too great and/or attraction to the stable state is too slow.

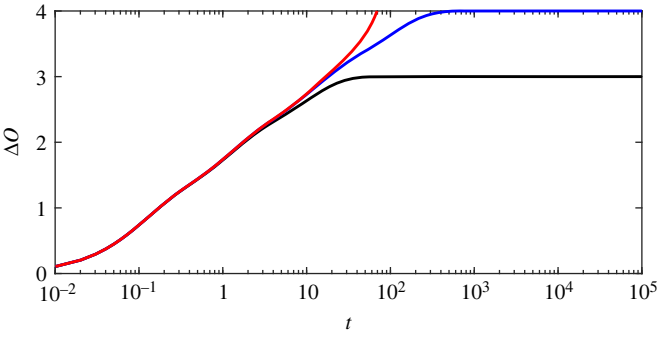

Figure 2. Examples of the response of observables ΔO_1 (black), ΔO_2 (blue) and ΔO_3 (red), sums of exponential functions (see text). Note that the *t*-axis is given in log-scale to highlight how long these different equations stay almost indistinguishable: the red case corresponds to a linearly unstable setting, i.e. a 'run-away' scenario.

where $\lambda_j \in \mathbb{C}$ represent eigenvalues of the linear system and $\beta_j^{[O]} \in \mathbb{R}$ depends on the corresponding eigenvector and observable; often λ_j is restricted to the negative reals but more generally they may be complex with oscillatory decay (e.g. [34]).

Estimating the Green's function for high (or infinite) dimensional systems can be extremely challenging—not least because linear operators in infinite dimensions may have a continuous (operator) spectrum. Nonetheless, one can assume a functional form for $G^{[O]}(t)$, and fit parameters from transient data. This approach has been applied successfully to many response problems in the climate system (e.g. [30,33–35]).

Let us now assume that (2.4) holds for the Green's function, and restrict to $\lambda \in \mathbb{R}$. Even then, the number of modes M needs to be determined, and that comes with its own problems as shown in figure 2 which compares responses of $\Delta O_1(t) = 3 - \mathrm{e}^{-10t} - \mathrm{e}^{-t} - \mathrm{e}^{-0.1t}$, $\Delta O_2(t) = \Delta O_1(t) + 1 - \mathrm{e}^{-0.01t}$ and $\Delta O_3(t) = \Delta O_1(t) + \mathrm{e}^{+0.01t} - 1$. The first two are bounded but with different asymptotic values; the third has a 'run-away' response. Nonetheless, figure 2 shows that all three observables are indistinguishable at first; only over longer time scales does the effect of the extra term become apparent. It is practically impossible to determine which of these functional forms is correct from short-time transient data only.

In the climate system, the dynamics play out over many different time scales [36,37]. Hence, this should play an important role in understanding GCM experiments. In particular, it is

important to try to determine time scales on which the constructed estimations and extrapolations can be trusted, as there seems to be no way to completely rule out slow warming, or even slow tipping, on all slow time scales. GCMs very often do not include the very slow climate components such as land ice sheets dynamically, but still need very long spin-up times and almost never are integrated to full equilibrium. For example, palaeoclimate experiments with GCMs typically show considerable drifts in the globally averaged ocean temperature after several millennia of simulation, while already in good radiative balance (e.g. [38]).

3. Nonlinear response and ECS for climate models

In the previous section, we discussed potential problems associated with time scales that can affect estimation of ECS even for linear systems. In this section, we turn our attention to issues related to nonlinear response. Here, a particular challenge are tipping points where fast dynamics can suddenly take over even after long, very slowly evolving transient periods; this is impossible in a purely linear system.

To be explicit, we consider a GEBM with dynamics on two time scales and the possibility of tipping phenomena on a slow or a fast time scale. We introduce the model in §3a. Subsequently, we consider tipping-related effects due to slow time scales in §3b and due to internal variability in §3c.

(a) A fast-slow energy balance model

We consider a GEBM of Budyko–Sellers–Ghill type [39–41], which describes the evolution of GMST T according to the model

$$C\frac{\mathrm{d}T}{\mathrm{d}t} = Q_0(1-\alpha) - \varepsilon\sigma T^4 + \mu + \mu_{\mathrm{NV}}(t), \tag{3.1}$$

where C is the specific heat capacity, Q_0 is the incoming (predominantly short wave) solar radiation, α is the planetary albedo (so that $Q_0\alpha$ is the reflected solar radiation) and $\varepsilon\sigma T^4$ is the outgoing (predominantly long-wave) Planck radiation (with planetary emissivity ε and Boltzmann constant σ). Further, μ represents the mean radiative forcing due to increases in CO_2 and $\mu_{\mathrm{NV}}(t)$ models variability in radiative forcing, assumed to have zero mean. We take [42]

$$\mu = \mu_0 + A_0 \log\left[\frac{\rho(t)}{\rho(0)}\right] \tag{3.2}$$

with $A_0 = 5.35\,\mathrm{W\,m^{-2}}$, where $\rho(t)$ is the concentration of atmospheric CO_2 at time t, and μ_0 is a reference radiative forcing level for a CO_2 concentration of $\rho(0)$.

When albedo and/or emissivity are taken to be temperature-dependent, i.e. $\alpha = \alpha(T)$ and/or $\varepsilon = \varepsilon(T)$, the model can have multiple stable climate states each with different climate sensitivity. In this paper, we assume there is relaxation towards a temperature-dependent equilibrium albedo $\alpha_0(T)$ at a rate $\tau_\alpha \geq 0$, while we assume the emissivity responds immediately:

$$\left.\begin{aligned} \tau_\alpha\frac{\mathrm{d}\alpha}{\mathrm{d}t} &= [\alpha_0(T) - \alpha], \\ \alpha_0(T) &= \alpha_1 + (\alpha_2 - \alpha_1)\frac{1+\tanh(K_\alpha[T-T_\alpha])}{2} \\ \varepsilon(T) &= \varepsilon_1 + (\varepsilon_2 - \varepsilon_1)\frac{1+\tanh(K_\varepsilon[T-T_\varepsilon])}{2}. \end{aligned}\right\} \tag{3.3}$$

Note that α_0 and ε_0 change from one constant to another as T moves through a range of temperatures near $T_{\alpha,\varepsilon}$ [9]; $\alpha_0(T)$ models the (relatively slow) lowering of albedo in the presence of land ice sheets, while $\varepsilon_0(T)$ models a (relatively fast) transition from a clear to a cloudy planet with large quantities of low cloud. Each of them on their own can lead to a bistability between cold and warm climate states but we include both to allow the possibility of independent slow and fast tipping points. In fact, we believe that both the (slowly settling) temperature-dependent

Table 1. Values for the GEBM (3.1) with dynamic albedo (3.3) and chaotic variability (3.5) used in the numerical simulations. We take the standard choice for the Lorenz parameters: $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$. For the simulations, time t is rescaled to years. The equilibrium albedo and (equilibrium) emissivity are given by (3.3). The forcing μ is given by (3.2) and ν_{NV} represents the amplitude of a chaotic forcing via (3.4). The use of '· · ·' indicates that values of column *A* are also used in this case.

	A	*B*	*C*	*D*	units
C	5×10^8	· · ·	· · ·	· · ·	J m^{-2} K^{-1}
Q_0	341.3	· · ·	· · ·	· · ·	W m^{-2}
σ	5.67×10^{-8}	· · ·	· · ·	· · ·	W m^{-2} K^{-4}
α_1	0.7	· · ·	· · ·	· · ·	
α_2	0.289	· · ·	· · ·	· · ·	
T_α	274.5	· · ·	· · ·	· · ·	K
K_α	0.1	· · ·	· · ·	· · ·	K^{-1}
ε_1	0.5	· · ·	· · ·	· · ·	
ε_2	0.41	· · ·	· · ·	· · ·	
T_ε	288	· · ·	· · ·	· · ·	K
K_ε	0.5	· · ·	· · ·	0.1	K^{-1}
A_0	5.35	· · ·	· · ·	· · ·	W m^{-2}
τ_α	0	0	5×10^9	5×10^9	s
τ_{NV}	0	6×10^7	6×10^7	6×10^7	s
ν_{NV}	0	5	2×10^{-2}	2×10^{-2}	W m^{-2}

albedo and emissivity are required to have some of the later illustrated phenomena—late tipping in particular—that do not present themselves in models with constant albedo or emissivity.

We include natural variability of the energy input at the surface represented by chaotic forcing through a Lorenz-63 model, i.e. natural variability μ_{NV} is given by

$$\mu_{\mathrm{NV}} = \nu_{\mathrm{NV}} \sin\left(\frac{\pi x(t)}{20}\right), \tag{3.4}$$

where x adheres to the Lorenz-63 model, which conceptually represents the chaotic dynamics of weather processes [43]

$$\left.\begin{aligned} \tau_{\mathrm{NV}} \frac{\mathrm{d}x}{\mathrm{d}t} &= \sigma(y - x) \\ \tau_{\mathrm{NV}} \frac{\mathrm{d}y}{\mathrm{d}t} &= x(\rho - z) - y \\ \tau_{\mathrm{NV}} \frac{\mathrm{d}z}{\mathrm{d}t} &= xy - \beta z \end{aligned}\right\} \tag{3.5}$$

so that ν_{NV} is a measure for the strength of the variability and τ_{NV} is the characteristic time scale of chaotic variability. Parameter values used in the simulations in this paper are given in table 1, except where stated otherwise.

There are two special parameter settings that we distinguish. We say there is *dynamic albedo* if $\tau_\alpha > 0$; in the case $\tau_\alpha = 0$, albedo settles instantaneously so that we can eliminate (3.3) and set $\alpha = \alpha_0(T)$. We say there is *chaotic variability* if $\nu_{\mathrm{NV}} \neq 0$; in the case $\nu_{\mathrm{NV}} = 0$, there is no internal variability and we can eliminate the chaotic Lorenz-63 model (3.5).

It is well known that in the case of no internal variability, equations (3.1) can be bistable [39,41]. Due to the functional forms of temperature-dependent albedo and emissivity, the model (3.1) can have one, two or three stable equilibria depending on the parameter values. This is organized by a fifth order 'butterfly' singularity [44]: see the electronic supplementary material [45] for

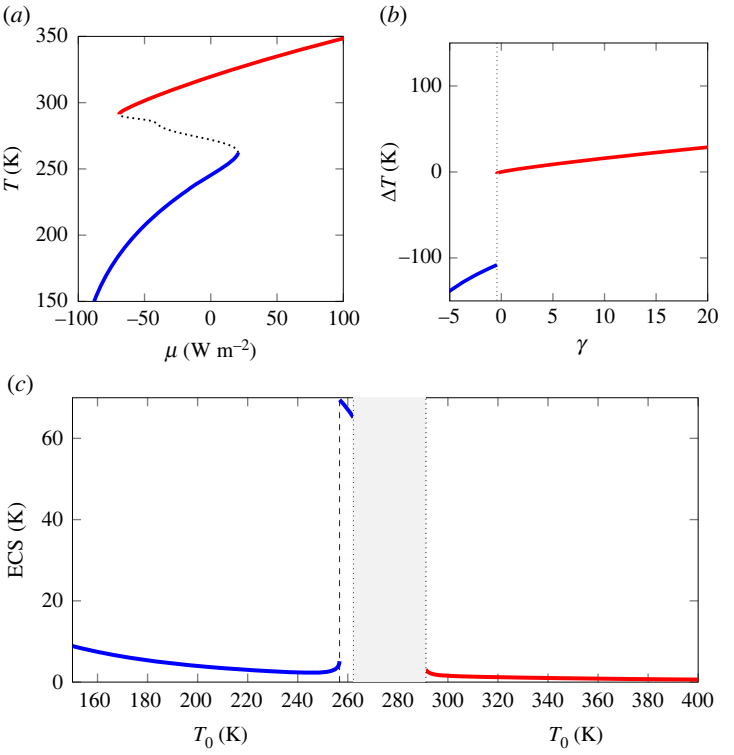

Figure 3. (*a*) Bifurcation diagram for (3.1) with parameters A from table 1. The bifurcation parameter μ represents radiative forcing due to atmospheric CO_2. Solid lines correspond to stable equilibria, and dashed lines to unstable equilibria. There are two different branches of stable equilibria: one that corresponds to a cold climate (blue) and one that corresponds to a warm climate (red). (*b*) Equilibrium 'two-point' response for different forcing levels corresponding to $2^{\gamma} \times CO_2$ starting from an initial state $T_0 = 293$ K corresponding to equilibrium temperature before perturbation. The red part of the figure corresponds to end states on the warm branch; the blue part to end states on the cold branch. The dashed line indicates the location of a tipping point. (*c*) ECS (i.e. equilibrium two-point response to CO_2 doubling) as function of the initial temperature T_0. Blue lines indicate starting points on the cold branch; red lines indicate starting points on the warm branch; the grey region corresponds to unfeasible initial temperatures (i.e. they lie on the unstable branch in (*a*)). The large peak in the blue line corresponds to tipping from the cold branch to the warm branch; the location of this tipping point is indicated with a dashed line.

a verification and in-depth analysis of the bifurcation structure of this model. Nonetheless, for the parameter values given in table 1, the model is bistable for a certain range of values of the parameter μ: in this bistable region, the model supports a stable cold 'icehouse' and a warm 'hothouse' climate state (see figure 3).

Note that the ECS of both type of states (for the same CO_2-level) differs between branches as albedo and emissivity are different between branches. However, the ECS within a branch is also not constant: figure 3*b* shows variation between initial points y_0 that lie on the same branch (intra-branch differences). In the climate literature, these variations are not well-quantified, mainly because they depend on a multitude of physical feedback processes, which are difficult to observe and model numerically in full [4,46]. Still, it is good to keep in mind that observed or estimated ECS might vary as the (initial) climate state changes.

(b) Nonlinear response: slow and/or late tipping and ECS

As discussed in §2, if the transient relaxation dynamics of a climate model is approximated well by a linear system this can be used to estimate ECS. This thus works for nonlinear systems with small enough forcings. For example, figure 4*a* shows a simulation of (3.1) with parameters C in

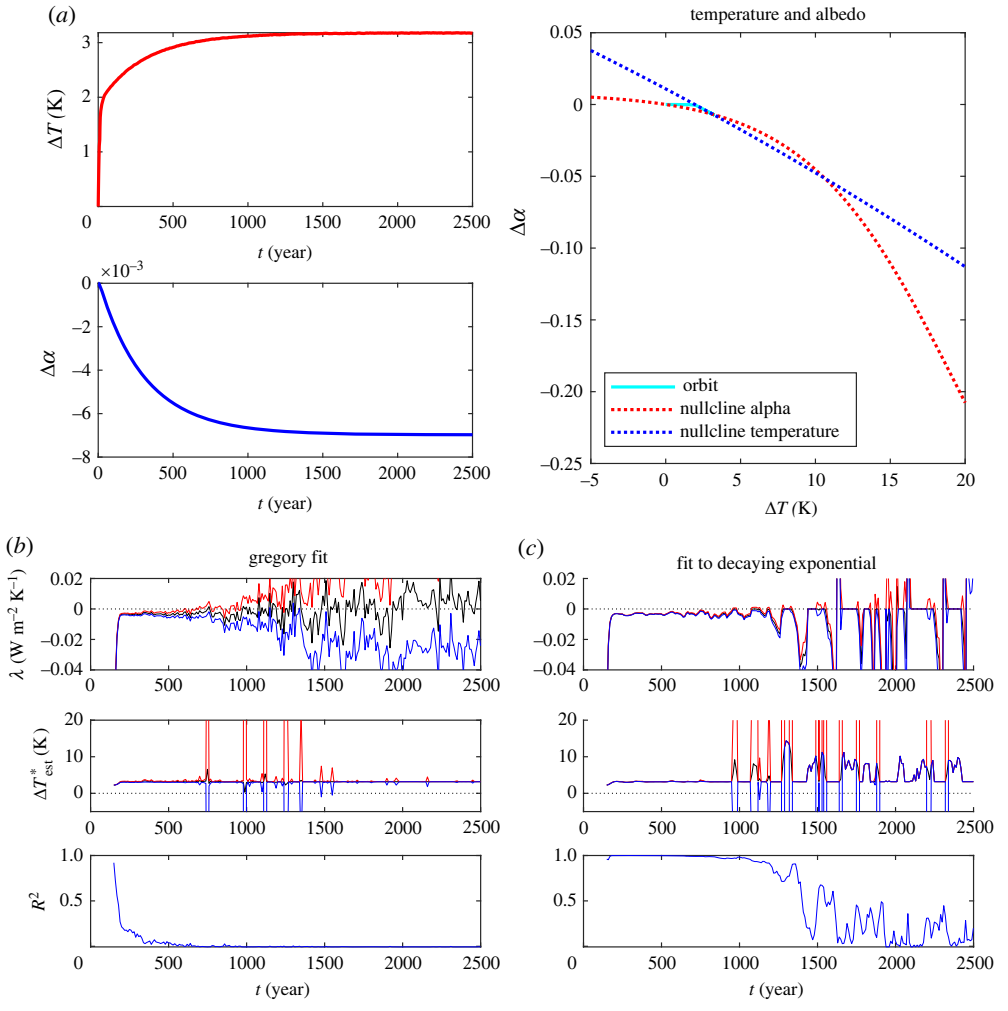

Figure 4. Warming in the fast-slow energy balance model (parameters C in table 1) subjected to an abrupt2xCO_2 forcing. The initial forcing μ_0 is chosen such that there is an initial equilibrium at $T_0 = 255$ K. (*a*) Time series for ΔT and $\Delta\alpha$ as well as the trajectory through (projected) phase space (cyan). The red dotted curve in the right panel denotes the nullcline on which $\mathrm{d}\alpha/\mathrm{d}t = 0$ and the blue dotted curve denotes the nullcline on which $\mathrm{d}T/\mathrm{d}t = 0$, which also acts as a slow manifold. (*b*) Gregory fits on time windows of 150 years, showing the thus estimated feedback parameter λ and expected equilibrium warming $\Delta T^*_{\mathrm{est}}$ over time, together with standard errors and R^2 statistic from the linear regression. (*c*) Fit to a decaying exponential on time windows of 150 years, showing the estimated feedback parameter λ and expected equilibrium warming $\Delta T^*_{\mathrm{est}}$ over time, similarly showing upper/lower standard error estimates and R^2 from the nonlinear regression.

table 1 subjected to an abrupt2xCO_2 forcing. The initial forcing μ_0 is chosen such that there is an equilibrium at $T_0 = 255$ K. There is a clear two-stage exponential decay to equilibrium with $\Delta T^* \approx 3.1$ K: the right panel of figure 4*a* shows that there is a nearby equilibrium attractor.

Figure 4*b* shows estimates using the Gregory method on rolling windows of 150 years. Within a time window, we regress the time series of ΔT and $\Delta N := C(\mathrm{d}\Delta T/\mathrm{d}t)$ to the linear model $\Delta N = f + \lambda\Delta T$, which gives estimates for the forcing f and the dominant feedback parameter λ. The regression is performed using the MATLAB fit to linear model `fitlm`; standard errors for best fit are shown, and the bottom panel shows the adjusted R^2-statistic for this window, where $R^2 = 1$

implies all variance in the signal is described by the model within the window ending at that time-point. Equilibrium warming is derived from these fits by extrapolation of the linear model, giving $\Delta T^*_{\text{est}} = -f/\lambda$. Note that the initial 150-year fit is already good. Indeed, one can see decreasing signal-to-noise ratio and R^2 for fits taken later in the time series, as noise dominates the dynamics of the state this late in the simulation.

A second equilibrium estimation protocol is shown in figure 4*c* in which blocks of 150 years are fitted to a decaying exponential function $T(t) = T_\infty + b\,e^{\lambda t}$ using the MATLAB fit to nonlinear model `fitnlm`. This gives an estimate for T_∞, b and λ with standard errors and is a direct approximation of a linear response to a Heaviside input; we show λ and $\Delta T^*_{\text{est}} = T_\infty - T_0$. Observe that, similarly to the Gregory fits, these fits also become degenerate for later time frames.

To contrast with figure 4, figure 5 shows the case for an abrupt4xCO_2 forcing but otherwise identical parameters and initial condition, in which the transient dynamics are not approximated well by a linear system, although a long transient period (due to the crossing of a slow tipping point) conceals the nonlinear dynamics. Figure 5*a* shows that the run seems to rapidly approach an equilibrium, but warming then continues slowly as albedo slowly decreases. Then, around $t = 1500$ yr, there is a surprising and rapid 'late tipping' followed by a relaxation to the final equilibrium. From the fits in figure 5*b*,*c*, approximately linear behaviour can be seen at first; however, we are near (but beyond) a fold bifurcation on the stable part of the slow manifold where the blue and red nullclines become tangent (i.e. a slow tipping point), and for this forcing the nullclines are barely detached. As the state passes this point (sometimes called a ghost attractor), the dynamics on the slow manifold speed up before tipping over a fold in the slow manifold, causing a rapid late tipping event to another stable branch of this slow manifold.

Figure 5*b* shows estimates using a Gregory fit. It can be seen that a fast decay is picked up initially, and slower decay dominates from about $t = 250$ yr. At around $t = 500$ yr, the fitted value for λ passes through zero, suggesting a linearly unstable climate, and the estimated warming becomes unreliable. Only after the late tipping event, from $t = 1750$ yr onwards, the fits make sense again, with negative λ and sensible warming estimates corresponding to the actual equilibrium warming of the simulation. Similarly, for the exponential fit shown in (c), $\lambda \approx -0.05$ corresponding to the initial fast decay but this quickly decays to pick up the slow decay with $\lambda \approx -0.001$ by about $t = 250$. The fit remains good up to $t \approx 500$ yr but after this the estimated errors on ΔT_{est} increase rapidly as the fit attempts to fit a decaying exponential to something that is actually growing slowly but exponentially. At $t \approx 1500$ yr the system passes through the late rapid tipping before settling to a fit to $\Delta T^*_{\text{est}} \approx 72$.

Clearly, in both of these fitting approaches the true equilibrium warming is not estimated accurately at all until after the late tipping event when the system is again approximately linear. From the fits up to about $t = 500$ yr there are no obvious hints that anticipate this late tipping and the fit results seem to indicate convergence to a noisy equilibrium state (hence for example the low R^2 score for the Gregory method as it is mostly noise at this point). Only after $t = 500$ yr there start to be some signs of the passing of a slow tipping point ($\lambda > 0$ in the Gregory method and large uncertainties in the exponential fit method) in this example, as the almost-equilibrium (ghost attractor) on the slow manifold is passed around this time.

Comparing figures 4 and 5, we see very similar fits and estimates up to $t = 500$ yr, further indicating the difficulty of distinguishing scenarios with and without late tipping. Moreover, the perturbation that exceeds the threshold shown in figure 1*b* lies somewhere between 2xCO_2 and 4xCO_2 for this model and parameters.

Figure 6 shows an analogous simulation of (3.1) under abrupt4xCO_2 forcing but parameters D of table 1. Again, the initial forcing μ_0 is such that there is an initial equilibrium at $T_0 = 255$ K. For these parameters there is no fold in the critical manifold meaning that there is not a rapid late tipping (in the bifurcation sense). However, similarly to figure 5, the initial (linear) warming is

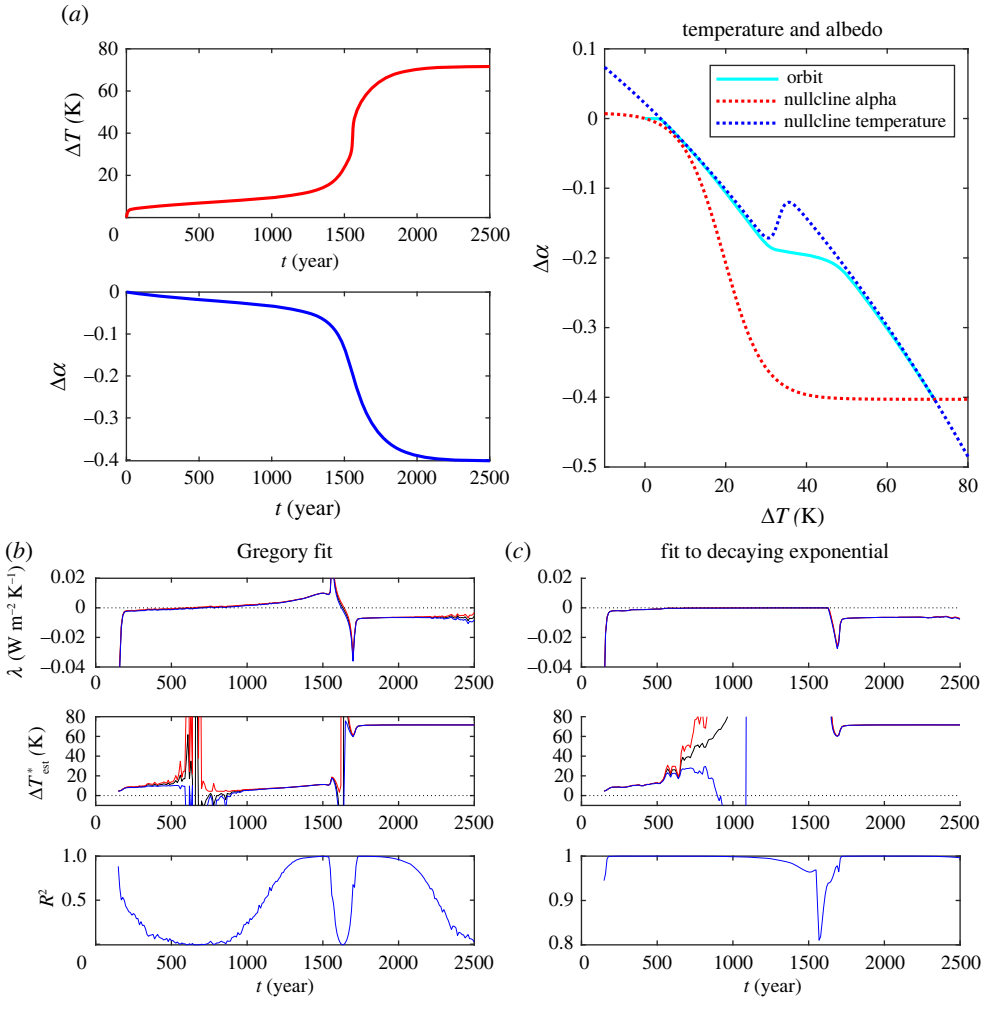

Figure 5. Warming in the fast-slow energy balance model (parameters C in table 1) subjected to an abrupt4xCO_2 forcing. The initial equilibrium is $T_0 = 255$ K. Here, a late tipping event happens as the dynamics drive the system over a fold point of the slow manifold. (*a*) Time series for ΔT and $\Delta\alpha$ as well as a the trajectory through phase space (cyan). The red dotted curve in the right panel denotes the nullcline on which $(\mathrm{d}\alpha/\mathrm{d}t) = 0$ and the blue dotted curve denotes the nullcline on which $\mathrm{d}T/\mathrm{d}t = 0$, which also acts as a slow manifold. Fits (*b*,*c*) as in figure 4.

not representative of the equilibrium warming and the transient means one can only see evidence of the final state after $t = 1500$ yr.

This indicates that even in the absence of (late) tipping points, an initial good fit cannot exclude a later rapid warming phase in systems that have dynamics on multiple time scales. For all three simulations presented in this section, extrapolations from fits to the initial few hundred years look very similar, although their long-term behaviour is very different, again highlighting that extrapolations may only be accurate after long transients that bring the system into a linear regime.

(c) Ensemble variability and ECS

When estimating ECS in models with internal variability, one of the ingredients is the precise choice of the initial conditions y_0. In §2a, we already discussed that the background climate

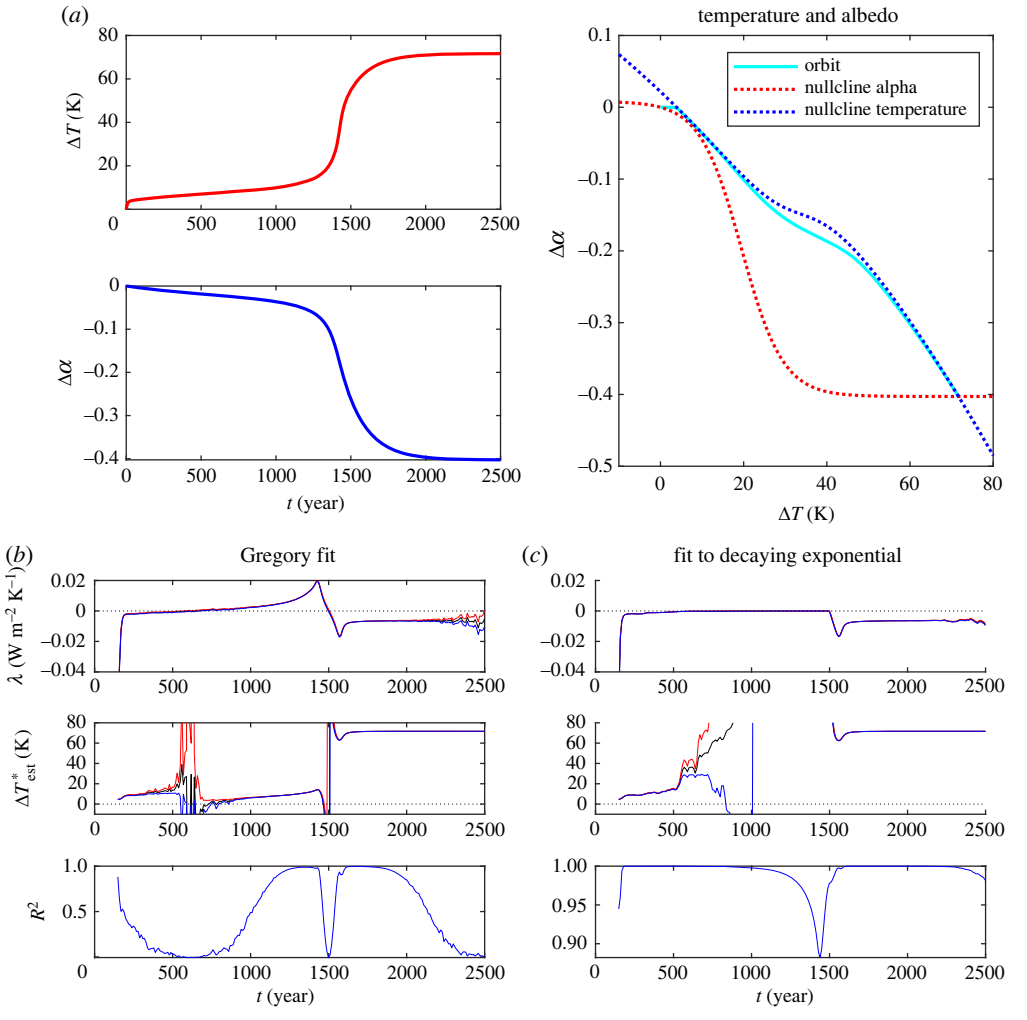

Figure 6. (*a*) Warming in the fast-slow energy balance model (parameters D in table 1) subjected to an abrupt4xCO_2 forcing. The initial equilibrium is $T_0 = 255$ K. The left panel shows time series for ΔT and $\Delta\alpha$. The right panel shows the trajectory through phase space (cyan). The red dotted curve in the right panel denotes the nullcline on which $\mathrm{d}\alpha/\mathrm{d}t = 0$ and the blue dotted curve denotes the nullcline on which $\mathrm{d}T/\mathrm{d}t = 0$, which also acts as a slow manifold. Note that for the parameters D there is no longer a late tipping in T of the speed as seen in figure 5, nonetheless there is a moderately rapid increase in T around 1500 years. Fits in (*b*,*c*) as for figure 4.

state (i.e. the initial attractor A_0) influences the transient and equilibrium response to forcings. However, the precise initial state y_0 on the initial attractor will also impact the observed transient dynamics and can potentially change the final equilibrium state. We illustrate such situations in this subsection.

Figure 7*a*,*b* shows an abrupt4xCO_2 experiment for an ensemble of different initial states on the same initial attractor (a warm climate state), for a simulation of (3.1) with parameters B of table 1—note the presence of chaotic variability. There is potential variation in the warming of the different ensemble members during the transient, which stems from different realizations of the natural variability, corresponding to the different initial states. Gregory fits over a time window starting at time 0 up to time t are shown in figure 8*a*. The associated regression to individual ensemble members (black) are poor, but the regression to the ensemble average (red) is much better as the noise (internal variability) is averaged out.

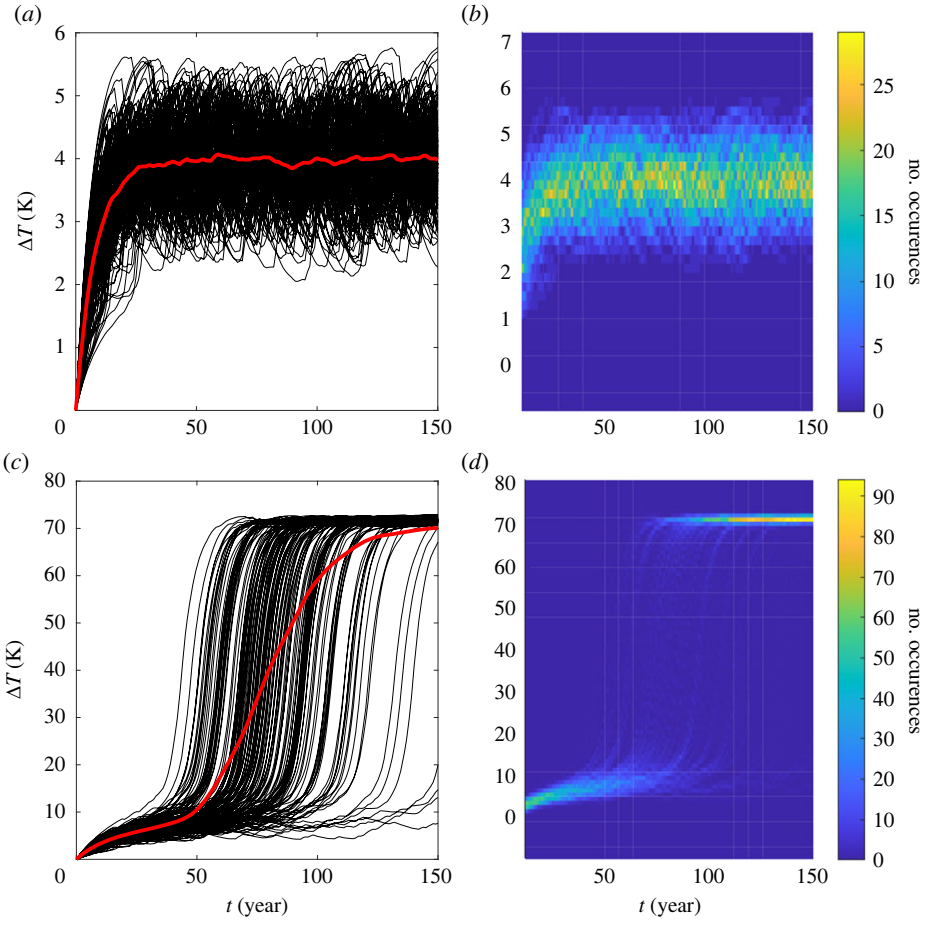

Figure 7. Results of a 150 ensemble run of abrupt4xCO_2 experiments for the energy balance model with chaotic forcing and instantaneous albedo relaxation (parameters B of table 1 with initial temperature (*a*,*b*) $T_0 = 293$ K (a warm climate) and (*c*,*d*) $T_0 = 255$ K (a cold climate)). Initial conditions for the Lorenz part of the model are randomly chosen for each ensemble member separately. (*a*,*c*) Time series of the warming ΔT over time for a random set of 100 of the ensemble members (black) and the ensemble average (red). (*b*,*d*) Heat maps indicating the number of times a certain warming has been observed per time step (note that temperature bins are differently sized between (*b*,*d*)).

Another example is given in figure 7*c*,*d* for a different initial attractor (a cold climate state), where natural variability pushes the state over a tipping point at different times during the simulation of each ensemble member. The simulations initially suggest relaxation towards a state close to the original colder state, but later they consistently exhibit tipping to a different (and much warmer state). In this example, the colder state is almost at equilibrium. As long as the natural variation in forcing is small enough, the system remains close to the colder state. For larger fluctuations the system tips into the warmer state. Figure 8*b* shows that even the ensemble average is not adequate to estimate ECS in this case; accurate estimates can only be made if the model has been run until (almost) all individual ensemble members have tipped. Nevertheless, the ensemble averaged response is still much better than the other approaches, because data from tipped and non-tipped ensemble members leads otherwise to very unreliable bimodal estimates with high variance.

Even worse, the equilibrium response may depend drastically on the precise initial state y_0; some part of the initial attractor A_0 can be attracted to a final attractor A_1, while the rest is attracted to a different final attractor $\tilde{A}_1$. This effect has been called a *partial tipping* [9,47,48]. The relative

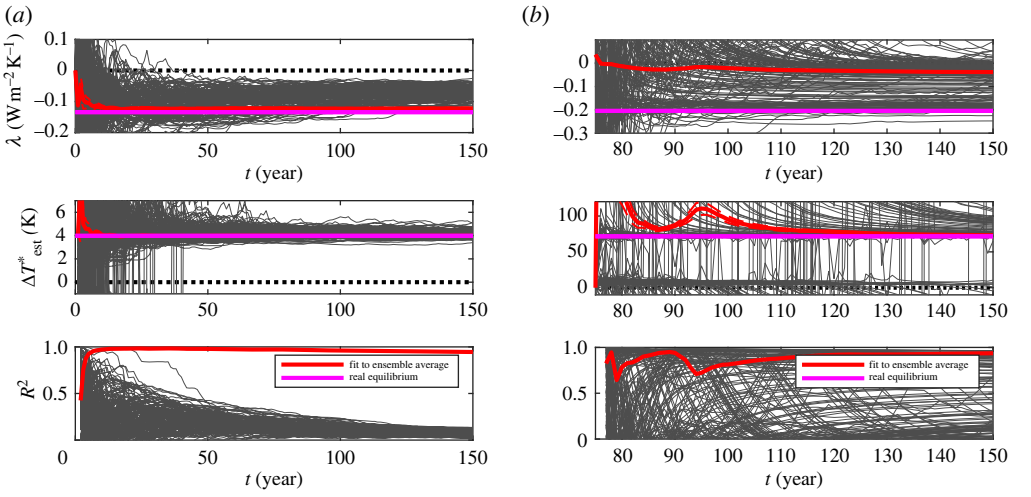

Figure 8. (*a*) Gregory fits for ensemble data from figure 7*a*,*b* up to year *t*, showing the feedback parameter λ, the estimated equilibrium warming $\Delta T^*_{\mathrm{est}}$ and the R^2 statistic. (*b*) Similar for figure 7*c*,*d*. Grey lines indicate results on individual ensemble members; red lines indicate results of regression on the ensemble mean; magenta lines indicate the theoretical real values. Thesolid lines indicate the expected values and the dashed lines the standard errors (almost imperceptible as standard errors are typically very low). The black dotted line denotes the zero-line.

simplicity of the chaotic GEBM means we have not been able to find partial tipping behaviour for constant forcing, but we can illustrate this phenomenon by forcing the model temporarily with an abrupt4xCO_2 forcing, after which the initial CO_2-levels are restored at time $t = 75$ yr. Figure 9 shows the results of this experiment. One can clearly see that some ensemble members experience tipping but others do not. In this situation (details not shown), partial tipping means that none of the ECS estimation techniques will paint a full picture. The ensemble average does contain some information on the number of tipping and non-tipped states but we suggest more meaningful estimates would need to be made for the attractors separately, first by categorizing each individual ensemble member as tipped or not, and using estimation techniques on these categories separately.

(d) Evidence of late tipping within GCM runs

For GCM runs with conditions corresponding to the relatively stable conditions of the Holocene pre-industrial climate, the accepted wisdom is that we do not expect to find any major global tipping effects as extreme as the icehouse to hothouse transitions explored above. Nonetheless there are hints that we may be close to a variety of regional tipping points [49] such as changes in the Atlantic meridional overturning circulation (AMOC) or West Antarctic icesheet collapse, and some emissions scenarios are likely to take us over these tipping points. Crossings of these regional tipping points can result in a global signal, such as changes in the AMOC leading to global climatic changes [50,51]. Further, as emission reduction scenarios may take us over tipping points only temporarily [52], also the possibility of a partial tipping of an attractor may be very relevant to study in GCMs.

Initial conditions for GCM runs are notoriously difficult to set—they are typically taken as the end of a spin-up simulation, or as a state at some time during a control experiment (in both of which atmospheric CO_2 is kept fixed at the starting levels). In ensemble runs, variation of initial states on the initial attractor are sometimes explored either by sightly perturbing an initial state (called 'micro-perturbations'), or by taking several states of a control run, typically separated by a few months up to a few years, depending on the time scale of the internal variability that is

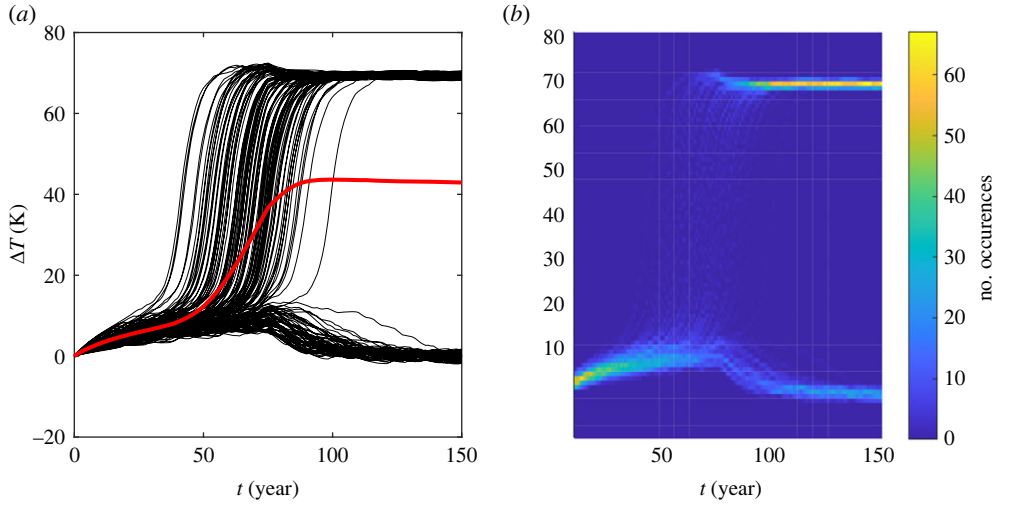

Figure 9. Partial tipping in the results of a 150 ensemble run of an experiment for the energy balance model with chaotic forcing and instantaneous albedo relaxation (parameters B from table 1) in which there is a temporary CO_2 quadrupling forcing for 75 years after which the initial CO_2-levels are restored. The initial temperature is $T_0 = 255$ K (a cold climate). (*a*) Time series of the warming ΔT over time for a random set of 100 of the ensemble members (black) and the ensemble average (red). (*b*) Heat map indicating the number of times a certain warming has been observed per time step (temperature bins of size 1 K have been used).

being considered (called 'macro-perturbations') [53,54]. Nonetheless, even after substantial spin-up there may be continued variability that can cause extrapolations such as effective climate sensitivity to continue varying over centennial time scales [5]. For example, [55,56] find multi-century changes in an atmosphere-ocean GCM, mostly to do with the strength of the AMOC, depending on the magnitude of the CO_2 perturbation.

The response of GCMs can also include late rapid changes. An example of such a late warming event is visible around year 2300 of the abrupt8xCO_2 run in the model CESM 1.0.4 within LongRunMIP [8]. Figure 10 shows features of this run, along with associated abrupt2xCO_2 and abrupt4xCO_2 runs of the same model for comparison. In figure 10*a*, the time series for the increase in (yearly averaged) global mean near-surface temperature is shown. For the abrupt8xCO_2 experiment a late and sudden increase can be seen around $t = 2300$ yr (highlighted in red in the figure), which is not present in the other experiments. We have analysed this data using the Gregory method on millennia-long rolling windows (to suppress the natural variability on shorter time scales) in figure 10*c*,*d*. We found an increase in the feedback parameter λ around the same time, and also an underestimation of the equilibrium warming for $t < 2400$ yr. This is similar to our findings in a conceptual energy balance model (figure 5) albeit less distinct. Hence, we suggest that this late warming event in the abrupt8xCO_2 run could be an example of a late tipping event in a GCM. The electronic supplementary material [45] shows this tipping behaviour is probably due to a qualitative (regional) transition of the AMOC which appears for the abrupt8xCO_2 run, but is not present in the abrupt2xCO_2 or abrupt4xCO_2 runs. Unlike in figure 5, the tipping for the abrupt8xCO_2 run is of transient nature: the final state is an 'AMOC on' state in all cases. Morover, AMOC tipping generally has less impact on the global mean surface temperature than ice-albedo feedback processes that generate the tipping in the GEBM.

4. Conclusion and discussion

Although many authors have pointed out deficiencies with estimating and using ECS, it clearly remains an important metric for understanding the response of climate models to changes in

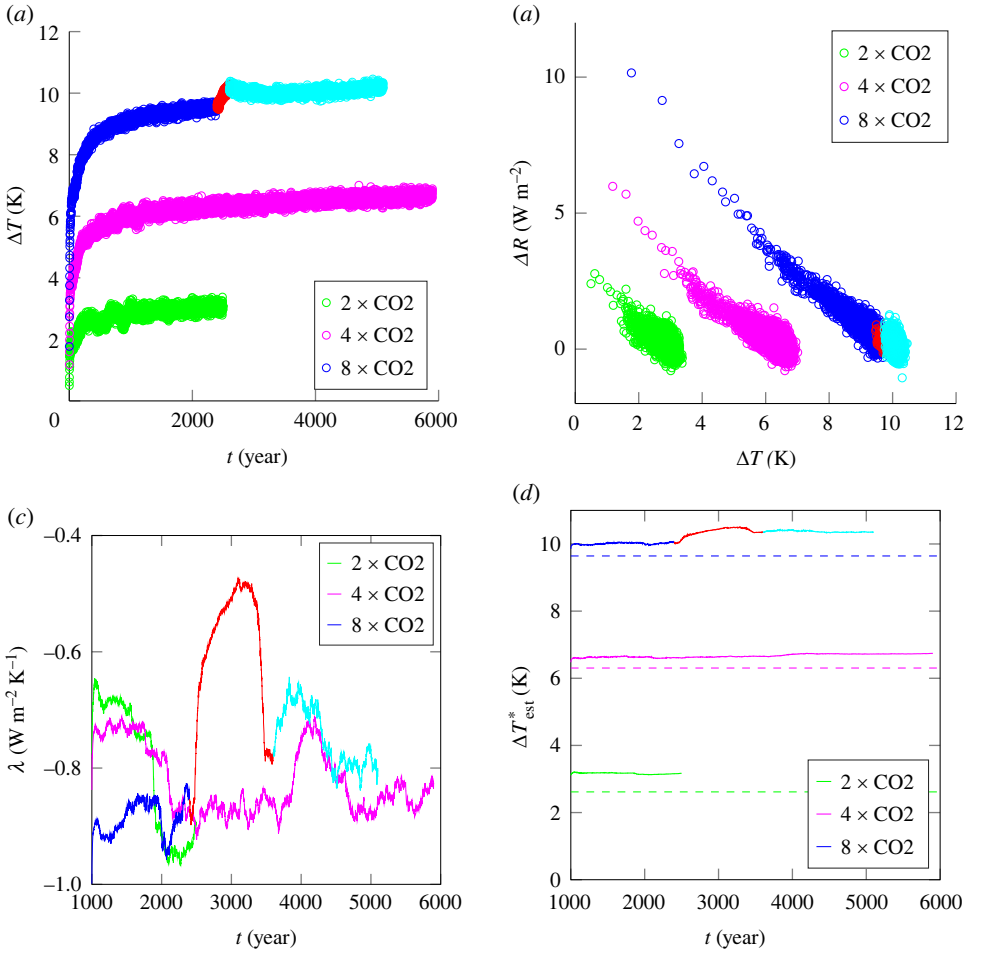

Figure 10. Outcomes of multi-millennial experiments in the GCM 'CESM 1.0.4.' for an abrupt2xCO_2 (green), an abrupt4xCO_2 (magenta) and an abrupt8xCO_2 (blue and red) experiment (data from longrunmip [8]). For the abrupt8xCO_2 experiment, a sudden late increase in temperature can be seen around year 2500. (*a*) Time series of global mean surface temperature. (*b*)Gregory plot. (*c*) Results for estimated climate feedback parameter λ obtained via Gregory fits on time windows of 1000 years. (*d*) Results of estimated equilibrium temperature $\Delta T^*_{\mathrm{est}}$ via Gregory fits on time windows of 1000 years, and estimated warming from a Gregory fit on years 20–150 (dashed lines). For the abrupt8xCO_2 experiment the dark, respectively, light blue indicates years before, respectively, after the late fast warming. The red data points in (*a*,*b*) indicate the time period of the late warming; in (*c*,*d*) the red data points indicate that the regression has used some of the data of this late warming period.

forcing CO_2. In particular, ECS and variants of it are key metrics that find their way into decision making about climate change and its likely impact on human activities. This happens for instance via integrated assessment models of the socioeconomic impact of an emissions pathway such as used in [57–59]. In this paper, we have illustrated how such linear concepts could break down in many different ways, even after long transient periods in which they seem valid, when nonlinear dynamics start to play a role. Although we have focused in this paper on climate response to idealized abruptNxCO_2 forcing scenarios, we also want to stress that in multi-stable nonlinear systems the precise outcome can also depend on the pathway taken—that is, not only the amount but also the timing of emissions can be important. This further complicates and challenges simplistic linear frameworks. See for instance [52,60] for examples in conceptual settings and [61] for a discussion on how this can strongly influence integrated assessments.

Climate is a multi-scale process that takes place on many fast and slow time scales, so it is unrealistic to assume that all dynamics can be modelled by a univariate linear model. Moreover, we also cannot expect to estimate processes that take place over substantially longer time scales than the simulated duration. As we have illustrated in this paper, even in the case of pure linear response ECS cannot be accurately estimated unless the simulation times are long enough to resolve the slow time scales such as those common in large scale ocean dynamics or land ice sheets. On top of that, with the examples in §3 we have illustrated how nonlinear effects in a multi-scale climate model can lead to additional warming effects—such as slow and late tipping, with long transients that give no obvious early hints of these late events. Even if a fit is very good for a long period of time, there can still be large and abrupt late tipping points. In such cases it is not sufficient to include a nonlinear correction term (such as a quadratic correction to the Gregory method [62]) to otherwise linear regression models. This is because a late tipping can lead to a discontinuous response in the long-term behaviour.

In §2 and in figure 1, we have introduced several trade-offs that need to be made when estimating ECS for a climate model. It would be of great interest to locate the 'Goldilocks Zone' in which reliable and accurate estimates ECS are possible, in order to give suitable protocols for experiments with GCMs. In particular, it would be good to understand (i) the minimum times and ensemble sizes needed to reliably estimate ECS and (ii) the thresholds in perturbation size for general GCMs that lead to tipping behaviour. This will depend not just on the current climate state but also on the processes that are included in the model and the form of the forcing. We suggest there is a need to find criteria that imply that an estimation protocol will work—and on which time scales. For instance, the Gregory method when applied on data from one decade can typically predict a few decades but is unlikely to be predictive on the scale of centuries; similarly, if only 150 years of data is available, it is unlikely to obtain an accurate estimation on millennial time scales.

Although we cannot make general quantitative statements about the 'Goldilocks Zone' for GCMs, we can give some more qualitative suggestions for model experiments included in the ever-growing CMIP activity, which could help reduce estimation errors for climate sensitivity and eventually help to better quantify the necessary trade-offs.

First, important practical issues for future climate projections include understanding the trade-offs made when estimating ECS, as well as determining which tipping points may be crossed on time scales of centuries. Given the increasing evidence that several climate tipping elements, in particular those related to ice sheet dynamics, could be triggered already at relatively low warming [49], we suggest to use abrupt2xCO_2 experiments as a standard to estimate ECS, rather than the currently used abrupt4xCO_2 experiments, to prevent the crossing of any tipping point. The presence of natural variability with this choice may lead to decreased signal-to-noise ratio, which could be remedied by somewhat longer integration times (see figure 1), or additional estimates of ECS using exponential fitting techniques involving multiple observables or time scales [18,33]. The latter can also help to understand the different time scales involved in the response. Moreover, using GCMs to model and understand past tipping elements can further help to quantify the necessary trade-offs to be made.

Second, regarding the background state-dependence of ECS, the increasing number of Paleo model intercomparison projects for periods in the past (e.g. Last Glacial Maximum or Pliocene) can provide valuable insights. These MIPs typically perform one or more experiments with CO_2-concentrations and boundary conditions of that respective time period in addition to pre-industrial control simulations. We suggest to add a standard experiment, where at the end of the palaeo-simulation, the CO_2-concentration is abruptly doubled (from its palaeo-value) and the simulation continued for another 150–200 years. This will allow to estimate ECS in the same way as is done for the models in present-day or pre-industrial mode, and can be used to map out the background state-dependence of ECS.

Third, it is known that the response to CO_2-forcing in any climate model can depend on the initial conditions both in the atmosphere and the ocean [53,54]. It is, therefore, desirable to use

(large) ensembles of simulations for one model with different initial conditions to estimate ECS in addition to multi-model ensembles with only one to a few realizations each. For example, the large spread in AMOC response to CO_2-increase as observed in the CMIP6 ensemble may be related to different ocean initial states [63].

The most drastic examples of nonlinear response given in this paper concern tipping phenomena. This begs the question of how relevant this is for future projections with GCMs. After all, in these models the GMST response is typically fairly linear to changes in forcing levels, and the transient response seems linear over quite long time scales. This might suggest that tipping points for GMST are not very relevant. However, the parameter space of such models has arguably not been sufficiently explored to capture even known cases of Palaeoclimate tipping such as [64]. As GCMs for future prediction are usually optimized for stable Holocene pre-industrial climates, they may operate in a too stable manner [65]. Local or regional tipping has been observed more frequently in GCMs [66] and can be observed in past climate records [36]. Tipping effects at regional levels may give only a small signal in the global average (although e.g. the AMOC restoration in the abrupt8xCO_2 experiment in figure 10 is visible in GMST). Often the global redistribution averages out—similar to what is described in [67,68]. However, regional tipping can cause dramatic local impacts, and *cascading* effects between regional tipping elements [69,70] may lead eventually to a global response. Hence, for useful impact projection it also will be important to go beyond classifying climate response purely via GMST to quantifying spatial responses of various observables to anthropogenic forcing scenarios.

Data accessibility. Simulation data from models in LongRunMIP data.iac.ethz.ch/longrunmip/, including the here used model CESM 1.0.4, requests for access can be made to the coordinators of LongRunMIP. More information and details of the simulations can be found on longrunmip.org and in [8]. The numerical code to simulate and subsequently analyse the conceptual energy balance model introduced in equations (3.1), (3.3), (3.4) is available from https://github.com/peterashwin/late-tipping-2022. Further details of the model are provided in the electronic supplementary material [45].

Authors' contributions. R.B.: formal analysis, investigation, methodology, writing—original draft, writing—review and editing; P.A.: formal analysis, investigation, methodology, writing—original draft, writing—review and editing; A.S.v.H.: formal analysis, investigation, methodology, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This project is TiPES contribution no. 168: This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement 820970.

Acknowledgements. We thank Richard Wood for discussions related to this work, and the reviewers for their insightful comments.

References

1. Arrhenius S. 1896 On the influence of carbonic acid in the air upon the temperature of the ground. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **41**, 237–276. (doi:10.1080/14786449608620846)
2. Lapenis AG. 1998 Arrhenius and the intergovernmental panel on climate change. *Eos Trans. AGU* **79**, 271–271. (doi:10.1029/98EO00206)
3. Charney JG. 1979 *Carbon dioxide and climate: a scientific assessment*. Washington, DC: National Academy of Science.
4. Sherwood S *et al.* 2020 An assessment of Earth's climate sensitivity using multiple lines of evidence. *Rev. Geophys.* **58**, e2019RG000678. (doi:10.1029/2019RG000678)
5. Senior CA, Mitchell JF. 2000 The time-dependence of climate sensitivity. *Geophys. Res. Lett.* **27**, 2685–2688. (doi:10.1029/2000GL011373)
6. Knutti R, Rugenstein MA. 2015 Feedbacks, climate sensitivity and the limits of linear models. *Phil. Trans. R. Soc. A* **373**, 20150146. (doi:10.1098/rsta.2015.0146)
7. Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, Taylor KE. 2016 Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958. (doi:10.5194/gmd-9-1937-2016)

8. Rugenstein M*et al.* 2019 LongRunMIP: motivation and design for a large collection of millennial-length AOGCM simulations. *Bull. Am. Meteorol. Soc.* **100**, 2551–2570. (doi:10.1175/BAMS-D-19-0068.1)
9. Ashwin P, von der Heydt AS. 2020 Extreme sensitivity and climate tipping points. *J. Stat. Phys.* **179**, 1531–1552. (doi:10.1007/s10955-019-02425-x)
10. von der Heydt AS, Ashwin P. 2016 State dependence of climate sensitivity: attractor constraints and palaeoclimate regimes. *Dyn. Stat. Clim. Syst.* **1**, dzx001. (doi:10.1093/climsys/dzx001)
11. Eckmann JP, Ruelle D. 1985 Ergodic theory of chaos and strange attractors. In *The theory of chaotic attractors* (eds BR Hunt, TY Li, JA Kennedy, HE Nusse), pp. 273–312. New York, NY: Springer.
12. Young LS. 2017 Generalizations of SRB measures to nonautonomous, random, and infinite dimensional systems. *J. Stat. Phys.* **166**, 494–515. (doi:10.1007/s10955-016-1639-0)
13. Young LS. 2002 What are SRB measures, and which dynamical systems have them? *J. Stat. Phys.* **108**, 733–754. (doi:10.1023/A:1019762724717)
14. Rugenstein MAA, Armour KC. 2021 Three flavors of radiative feedbacks and their implications for estimating equilibrium climate sensitivity. *Geophys. Res. Lett.* **48**, e2021GL092983. (doi:10.1029/2021GL092983)
15. Caballero R, Huber M. 2013 State-dependent climate sensitivity in past warm climates and its implications for future climate projections. *Proc. Natl Acad. Sci. USA* **110**, 14 162–14 167. (doi:10.1073/pnas.1303365110)
16. Rugenstein M *et al.* 2020 Equilibrium climate sensitivity estimated by equilibrating climate models. *Geophys. Res. Lett.* **47**, L083898. (doi:10.1029/2019GL083898)
17. Gregory J, Ingram W, Palmer M, Jones G, Stott P, Thorpe R, Lowe J, Johns T, Williams K. 2004 A new method for diagnosing radiative forcing and climate sensitivity. *Geophys. Res. Lett.* **31**, L03205.
18. Bastiaansen R, Dijkstra HA, von der Heydt AS. 2021 Multivariate estimations of equilibrium climate sensitivity from short transient warming simulations. *Geophys. Res. Lett.* **48**, e2020GL091090. (doi:10.1029/2020GL091090)
19. Andrews T, Gregory JM, Webb MJ. 2015 The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *J. Clim.* **28**, 1630–1648. (doi:10.1175/JCLI-D-14-00545.1)
20. Knutti R, Rugenstein MA, Hegerl GC. 2017 Beyond equilibrium climate sensitivity. *Nat. Geosci.* **10**, 727–736. (doi:10.1038/ngeo3017)
21. Lutsko NJ, Popp M. 2019 Probing the sources of uncertainty in transient warming on different timescales. *Geophys. Res. Lett.* **46**, 11 367–11 377. (doi:10.1029/2019GL084018)
22. Cummins DP, Stephenson DB, Stott PA. 2020 Optimal estimation of stochastic energy balance model parameters. *J. Clim.* **33**, 7909–7926. (doi:10.1175/JCLI-D-19-0589.1)
23. Dai A, Huang D, Rose BE, Zhu J, Tian X. 2020 Improved methods for estimating equilibrium climate sensitivity from transient warming simulations. *Clim. Dyn.* **54**, 4515–4543. (doi:10.1007/s00382-020-05242-1)
24. 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. *J. Clim.* **26**, 1859–1876. (doi:10.1175/JCLI-D-12-00196.1)
25. Lucarini V. 2018 Revising and extending the linear response theory for statistical mechanical systems: evaluating observables as predictors and predictands. *J. Stat. Phys.* **173**, 1698–1721. (doi:10.1007/s10955-018-2151-5)
26. Ruelle D. 2009 A review of linear response theory for general differentiable dynamical systems. *Nonlinearity* **22**, 855–870. (doi:10.1088/0951-7715/22/4/009)
27. Ragone F, Lucarini V, Dynamics FLC. 2016 A new framework for climate sensitivity and prediction: a modelling perspective. *Clim. Dyn.* **46**, 1459–1471. (doi:10.1007/s00382-015-2657-3)
28. Lucarini V, Sarno S. 2011 A statistical mechanical approach for the computation of the climatic response to general forcings. *Nonlinear Process. Geophys.* **18**, 7–28. (doi:10.5194/npg-18-7-2011)
29. Proistosescu C, Huybers PJ. 2017 Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Sci. Adv.* **3**, e1602821. (doi:10.1126/sciadv.1602821)

30. Hasselmann K, Sausen R, Maier-Reimer E, Voss R. 1993 On the cold start problem in transient simulations with coupled atmosphere-ocean models. *Clim. Dyn.* **9**, 53–61. (doi:10.1007/BF00210008)
31. Lembo V, Lucarini V, Ragone F. 2020 Beyond forcing scenarios: predicting climate change through response operators in a coupled general circulation model. *Sci. Rep.* **10**, 1–13. (doi:10.1038/s41598-020-65297-2)
32. Aengenheyster M, Feng QY, Van Der Ploeg F, Dijkstra HA. 2018 The point of no return for climate action: effects of climate uncertainty and risk tolerance. *Earth Syst. Dyn.* **9**, 1085–1095. (doi:10.5194/esd-9-1085-2018)
33. Bastiaansen R, Dijkstra HA, Heydt ASVD. 2021 Projections of the transient state-dependency of climate feedbacks. *Geophys. Res. Lett.* **48**, e2021GL09467.
34. Torres Mendonça GL, Pongratz J, Reick CH. 2021 Identification of linear response functions from arbitrary perturbation experiments in the presence of noise–Part 1: method development and toy model demonstration. *Nonlinear Process. Geophys.* **28**, 501–532. (doi:10.5194/npg-28-501-2021)
35. Maier-Reimer E, Hasselmann K. 1987 Transport and storage of CO_2 in the ocean—an inorganic ocean-circulation carbon cycle model. *Clim. Dyn.* **2**, 63–90. (doi:10.1007/BF01054491)
36. von der Heydt AS, Ashwin P, Camp CD, Crucifix M, Dijkstra HA, Ditlevsen P, Lenton TM. 2020 Quantification and interpretation of the climate variability record. *Glob. Planet. Change* **197**, 103399. (doi:10.1016/j.gloplacha.2020.103399)
37. Mitchell JM. 1976 An overview of climatic variability and its causal mechanisms. *Quat. Res.* **6**, 481–493. (doi:10.1016/0033-5894(76)90021-1)
38. Baatsen MLJ, von der Heydt AS, Huber M, Kliphuis MA, Bijl PK, Sluijs A, Dijkstra HA. 2020 The middle-to-late Eocene greenhouse climate, modelled using the CESM 1.0.5. *Clim. Past* **16**, 2573–2597. (doi:10.5194/cp-16-2573-2020)
39. Budyko MI. 1969 The effect of solar radiation variations on the climate of the Earth. *Tellus* **21**, 611–619. (doi:10.3402/tellusa.v21i5.10109)
40. Sellers WD. 1969 A global climatic model based on the energy balance of the Earth-atmosphere system. *J. Appl. Meteorol. Climatol.* **8**, 392–400. (doi:10.1175/1520-0450(1969)008<0392:AGCMBO>2.0.CO;2)
41. Ghil M. 1976 Climate stability for a Sellers-type model. *J. Atmos. Sci.* **33**, 3–20. (doi:10.1175/1520-0469(1976)033<0003:CSFAST>2.0.CO;2)
42. Myhre G *et al.* 2013 Antopogenic and natural radiative forcing. In *Climate change 2013: the Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report on the Intergovernmental Panel on Climate Change* (eds Stocker TF, *et al.*), pp. 659–740. Cambridge, UK: Cambridge University Press.
43. Lorenz EN. 1963 Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141. (doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2)
44. Montaldi J. 2021 *Singularities, bifurcations and catastrophes*. Cambridge, UK: Cambridge University Press.
45. Bastiaansen R, Ashwin P, von der Heydt AS. 2023 Climate response and sensitivity: timescales and late tipping points. Figshare. (doi:10.6084/m9.figshare.c.6350222)
46. von der Heydt AS *et al.* 2016 Lessons on climate sensitivity from past climate changes. *Curr. Clim. Change Rep.* **2**, 148–158. (doi:10.1007/s40641-016-0049-3)
47. Alkhayuon HM, Ashwin P. 2018 Rate-induced tipping from periodic attractors: partial tipping and connecting orbits. *Chaos* **28**, 033608. (doi:10.1063/1.5000418)
48. Ashwin P, Newman J. 2021 Physical invariant measures and tipping probabilities for chaotic attractors of asymptotically autonomous systems. *Eur. Phys. J. Spec. Top.* **230**, 3235–3248. (doi:10.1140/epjs/s11734-021-00114-z)
49. Armstrong McKay DI*et al.* 2022 Exceeding 1.5°C global warming could trigger multiple climate tipping points. *Science* **377**, eabn7950. (doi:10.1126/science.abn7950)
50. Stouffer RJ *et al.* 2006 Investigating the causes of the response of the thermohaline circulation to past and future climate changes. *J. Clim.* **19**, 1365–1387. (doi:10.1175/JCLI3689.1)
51. Jackson L, Kahana R, Graham T, Ringer M, Woollings T, Mecking J, Wood R. 2015 Global and European climate impacts of a slowdown of the AMOC in a high resolution GCM. *Clim. Dyn.* **45**, 3299–3316. (doi:10.1007/s00382-015-2540-2)
52. Ritchie PDL, Clarke JJ, Cox PM, Huntingford C. 2021 Overshooting tipping point thresholds in a changing climate. *Nature* **592**, 517–523. (doi:10.1038/s41586-021-03263-2)

53. Deser C *et al.* 2020 Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Clim. Change* **10**, 277–286. (doi:10.1038/s41558-020-0731-2)
54. Hawkins E, Smith RS, Gregory JM, Stainforth DA. 2016 Irreducible uncertainty in near-term climate projections. *Clim. Dyn.* **46**, 3807–3819. (doi:10.1007/s00382-015-2806-8)
55. Manabe S, Stouffer RJ. 1993 Century-scale effects of increased atmospheric CO_2 on the ocean–atmosphere system. *Nature* **364**, 215–218. (doi:10.1038/364215a0)
56. Manabe S, Stouffer RJ. 1994 Multiple-century response of a coupled ocean-atmosphere model to an increase of atmospheric carbon dioxide. *J. Clim.* **7**, 5–23. (doi:10.1175/1520-0442(1994)007<0005:MCROAC>2.0.CO;2)
57. Pörtner HO*et al. IPCC, 2022: Climate change 2022: impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press (In Press).
58. van Vuuren DP, van der Wijst KI, Marsman S, Hof AF, Jones CD. 2020 The costs of achieving climate targets and the sources of uncertainty. *Nat. Clim. Change* **10**, 329–334. (doi:10.1038/s41558-020-0732-1)
59. Nordhaus W. 1992 The 'DICE' model: background and structure of a dynamic integrated climate-economy model of the economics of global warming. Cowles Foundation Discussion Papers 1009 Cowles Foundation for Research in Economics, Yale University.
60. Ashwin P, Wieczorek S, Vitolo R, Cox P. 2012 Tipping points in open systems: bifurcation, noise-induced and rate-dependent examples in the climate system. *Phil. Trans. R. Soc. A* **370**, 1166–1184. (doi:10.1098/rsta.2011.0306)
61. Grubb M, Wieners C, Yang P. 2021 Modeling myths: on DICE and dynamic realism in integrated assessment models of climate change mitigation. *Wiley Interdiscip. Rev. Clim. Change* **12**, e698. (doi:10.1002/wcc.v12.3)
62. Bloch-Johnson J, Pierrehumbert RT, Abbot DS. 2015 Feedback temperature dependence determines the risk of high warming. *Geophys. Res. Lett.* **42**, 4973–4980. (doi:10.1002/2015GL064240)
63. Bellomo K, Angeloni M, Corti S, Hardenberg J. 2021 Future climate change shaped by inter-model differences in Atlantic meridional overturning circulation response. *Nat. Commun.* **12**, 3659. (doi:10.1038/s41467-021-24015-w)
64. Hopcroft PO, Valdes PJ. 2021 Paleoclimate-conditioning reveals a North Africa land–atmosphere tipping point. *Proc. Natl Acad. Sci. USA* **118**, e2108783118. (doi:10.1073/pnas.2108783118)
65. Valdes P. 2011 Built for stability. *Nat. Geosci.* **4**, 414–416. (doi:10.1038/ngeo1200)
66. Drijfhout S, Bathiany S, Beaulieu C, Brovkin V, Claussen M, Huntingford C, Scheffer M, Sgubin G, Swingedouw D. 2015 Catalogue of abrupt shifts in intergovernmental panel on climate change climate models. *Proc. Natl Acad. Sci. USA* **112**, E5777–E5786. (doi:10.1073/pnas.1511451112)
67. Rietkerk M, Bastiaansen R, Banerjee S, van de Koppel J, Baudena M, Doelman A. 2021 Evasion of tipping in complex systems through spatial pattern formation. *Science* **374**, eabj0359. (doi:10.1126/science.abj0359)
68. Bastiaansen R, Dijkstra HA, von der Heydt AS. 2022 Fragmented tipping in a spatially heterogeneous world. *Environ. Res. Lett.* **17**, 045006. (doi:10.1088/1748-9326/ac59a8)
69. Dekker MM, von der Heydt AS, Dijkstra HA. 2018 Cascading transitions in the climate system. *Earth Syst. Dyn.* **9**, 1243–1260. (doi:10.5194/esd-9-1243-2018)
70. Wunderling N, Donges JF, Kurths J, Winkelmann R. 2021 Interacting tipping elements increase risk of climate domino effects under global warming. *Earth Syst. Dyn.* **12**, 601–619. (doi:10.5194/esd-12-601-2021)