On the foundation of the α-β-γ approach to carbon-climate feedbacks

Christian H. Reick[1] and Guilherme L. Torres Mendonça[2,1]

[1]Max Planck Institute for Meteorology, Hamburg, Germany
[2]Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

Correspondence: Christian H. Reick (Christian.Reick@mpimet.mpg.de)

Abstract. The α-β-γ approach used to quantify the size of the feedbacks between climate and carbon cycle consists of two elements: the α-β-γ formalism expressing the feedback strength by the sensitivities α, β, and γ, and an experimental practice to determine these sensitivities from Earth system model simulations using a transient scenario where CO_2 is forced to rise far above its pre-industrial value. There are several reasons to be unsatisfied with this approach: the α, β, and γ sensitivities are introduced as linear expansion coefficients into the forcing and thus should be characteristics of the considered model as such, but they are known to be non-constant in time and to depend on the simulation scenario used to determine their values. Moreover, being linear, the whole approach should be valid only for sufficiently small forcing, so that the practice to calculate the sensitivities at maximum forcing reached in the simulations is rather questionable. Finally, the definition of the sensitivities as linear expansion coefficients into the forcing turns out to be inconsistent with the practice to apply the formalism to transient simulations: we demonstrate that, because of the internal memory of the Earth system, by such a definition all sensitivities are mathematically zero and thus not well defined. But as we show here, the whole approach can be justified when introducing the α, β, and γ sensitivities from the outset not as differential, but as difference quotients. In this way a linearization is not needed and one obtains a fully non-linear description of the feedbacks. Moreover, thereby the formalism can be extended to include also the synergy between the feedbacks so that it gets even exact. Nevertheless, the scenario and time dependence remain, being a necessary consequence of the application of the formalism to transient simulations. In this respect the α-β-γ approach to climate-carbon feedbacks differs from the well-known description of atmospheric feedbacks: in the latter case not transient, but equilibrium states are employed to quantify the feedbacks, a practice consistent with a linearization into the forcing; accordingly, the obtained sensitivities, as well as the feedback strengths calculated from them, are proper characteristics of the system, independently of how the equilibrium was reached. This would also be the case for the calculation of climate-carbon feedbacks by the α-β-γ formalism if one used equilibrium instead of transient simulations to compute the sensitivities. In the light of these results we discuss ~~in the outlook~~ the pros and cons of various options for future research ~~on the size~~ for quantifying climate-carbon feedbacks, including also the application of the generalized α-β-γ framework to obtain insight into the memory structure of the climate-carbon system.

1 Introduction

Without the huge amounts of anthropogenic CO_2 stored away by the oceans and the land biosphere, today's climate would be much less hospitable than it actually is. From estimates of the airborne fraction we know that currently ocean and land take up more than 50% of anthropogenic CO_2 emissions (Bennett et al., 2024), but in particular the ocean sink is expected to weaken in the future (Canadell et al., 2021). The value of the airborne fraction depends on various processes controlling the carbon uptake by land and oceans, in particular on the size of the feedbacks between climate and carbon cycle. To disentangle the contributions from the various feedbacks one needs model simulations of the Earth system in which the different feedback processes are switched on and off. Already the first such simulations revealed that the uptake of CO_2 by land and ocean is strongly weakened by global warming (Cox et al., 2000). By the invention of the α-β-γ formalism by Friedlingstein et al. (2003) the research on this topic was given a formal tool for a more objective quantification of the size of the different feedbacks. Since then this tool found application in numerous studies addressing the interactions between climate and carbon cycle (see e.g. the review (Friedlingstein, 2015)).

The quantification of climate-carbon feedbacks was in particular put forward in an international effort by the Coupled Climate Carbon Cycle Model Intercomparison Project (C^4MIP; see c4mip.net). In this project, the participating climate research centers performed Earth system simulations according to a common protocol, designed to obtain the necessary data for application of the α-β-γ formalism. The results from the various project phases were published in a series of papers (Friedlingstein et al., 2006; Arora et al., 2013, 2020) and prominently summarized in the carbon chapters of the then upcoming IPCC reports (Denman et al., 2007; Ciais et al., 2014; Canadell et al., 2021). The setups used for the simulations were from phase to phase a bit different, but all start out from pre-industrial conditions, in particular from pre-industrial atmospheric CO_2 concentration (about 280 ppm). While in the first phase of C^4MIP an 'emission-driven' setup was used, lateron the common experiment protocol was changed to a 'concentration-driven' experiment mode. In any case atmospheric CO_2 concentration is forced to gradually rise to several times its pre-industrial value towards the end of the 21^{st} century.

The usage of such transient simulations is *the* established way to quantify climate-carbon feedbacks by means of the α-β-γ formalism. Accordingly, all quantities of this formalism pertain to *transient* states of the Earth system. This is worth mentioning, because the α-β-γ formalism has according to (Friedlingstein et al., 2003, p. 694) been designed in the spirit of the description of atmospheric feedbacks by Hansen et al. (1984), but their formalism describes *equilibrium* feedbacks. This difference has an important consequence: following Hansen et al. (1984), also Friedlingstein et al. (2003) derived their formalism by employing a linearization into the forcing, but, as we show in the present study, while being consistent with the application to equilibrium states, this linearization turns out to be inconsistent with the usage of transient states.

That the linearity assumption may be problematic seems first recognized by Plattner et al. (2008): A major conclusion from the α-β-γ formalism is that the strength of the climate-carbon feedbacks is fully determined by the three sensitivities α, β, and γ (definitions follow below). Determining these in Earth system simulations, they found that in particular the values for the α and β sensitivities depend strongly on the employed transient scenario, even though, by the linearity assumption, they should be an invariant property of the invoked Earth system model alone. While Plattner et al. (2008, p. 2741) interpret this

scenario dependence by the presence of strong non-linearities in the system – noting as an example "system time lags" – that are ignored in the α-β-γ formalism, the linearity assumption is more directly questioned in the study by Gregory et al. (2009) appearing one year later: They also recognized the scenario dependence of the sensitivities, but point in addition to the C^{4}MIP simulation results presented in (Friedlingstein et al., 2006, Fig, 2), where one sees that the values of γ, and to a lesser extend also those of β, vary with the strength of the forcing. By these observations they conclude that the linear relations of the α-β-γ formalism, in which β and γ relate the land and ocean carbon uptake to CO_2 and temperature rise, are "inadequate" (Gregory et al., 2009, pp. 5242 and 5248). They argue that by the linearity of these relations effectively an instantaneous equilibration of the carbon system is assumed, whereby the memory of the system is ignored causing the sensitivities to get dependent on the forcing scenario and the strength of the forcing. Similar conclusions were drawn by Boer and Arora (2009, p. L02704), using flux-based analogues of the storage-based sensitivities β and γ. As Plattner et al. (2008), they attribute the scenario and forcing dependence to the presence of non-linearities that are not accounted for in a linear formalism.

But even though it thus seems to be known for long that the linearity assumption doesn't hold, the α-β-γ formalism continues to be used to quantify feedbacks. If at all this practice is defended, the employed arguments are rather unspecific: E.g. Arora et al. (2013, p. 5305) write that "however, the feedback parameters provide insight into the behaviour of feedbacks" and "provide a useful common framework for comparing models". And in (Gregory et al., 2009, p. 5248) we read that the formalism "remains useful as an interpretative tool and indicative of the sources of uncertainty." The present study is also a defense of the α-β-γ formalism, but in contrast to the quoted authors, we do not argue that the formalism can be used *despite* the invalidity of the linearity assumption, but because upon proper interpretation such a linearity assumption *is not needed.*

In the following we first recall the α-β-γ formalism (section 2) and how it is applied to determine the strengths of the climate-carbon feedbacks from transient Earth system simulations (section 3). The presentation of the formalism leaves out the question of linearization, which is then critically discussed in section 4. Here we show that when introducing the sensitivities as linear expansion coefficients into the forcing they are mathematically zero. By comparing the α-β-γ approach to climate-carbon feedbacks with the description of atmospheric equilibrium feedbacks (Hansen et al., 1984) it will get clear that the problem of linearization is not arising from the formalism itself, but from the practice to use transient instead of equilibrium simulations to determine the size of the feedbacks. Based on these considerations, in the subsequent main section 5 of the study we show that a linearization is not needed to justify the α-β-γ approach: It turns out that upon proper interpretation the approach accounts for the full non-linear response to the forcing. The whole approach is found to be justified as long as the synergy between the feedbacks is small, but, as we show in section 6, this latter assumption can be eliminated by accounting explicitly for the synergy, whereby the formalism gets even exact. In section 7, we discuss the consequences of this re-interpretation of the α-β-γ formalism, in particular to what extent published results obtained by application of the formalism are affected. Based on the foregoing considerations, we finally compare in section 8 the advantages and disadvantages of different options for future investigations of the size of climate-carbon feedbacks.

magnitude

unclear to me what this 'exactly' means. No pun intended.

I don't follow this as a non-mathematician.

2 The α-β-γ formalism

This section presents the α-β-γ formalism, originally introduced by Friedlingstein et al. (2003). For our purpose it will be sufficient to introduce it in a rather formal way, based on the feedback diagram shown in Fig. 1, without going into much detail of the underlying processes.

Starting point is a disturbance of pre-industrial climate and carbon cycle by anthropogenic CO_2 emissions into the atmosphere. Measuring time t since start of the perturbations – typically the preindustrial reference date 1850 AD – let $I(t)$ be the cumulated amount of anthropogenic CO_2 emissions that has been added until time t. Measuring $I(t)$ in carbon units, the resulting change in atmospheric CO_2, expressed as change in atmospheric carbon content ΔC_A, can by application of carbon conservation be written as

relative to pre-industrial atmos. C burden

$$\Delta C_A(t) = I(t) - \Delta C_{L+O}(t), \tag{1}$$

where ΔC_{L+O} is the change in the combined land and ocean carbon storage that happened in reaction to the emissions since pre-industrial times. In the published presentations of the α-β-γ formalism, land and ocean carbon are usually treated separately, but for our purpose of clarifying its foundations it suffices to treat them as a single quantity ΔC_{L+O}. Eq. (1) refers to the node depicted as Σ in Fig. 1.

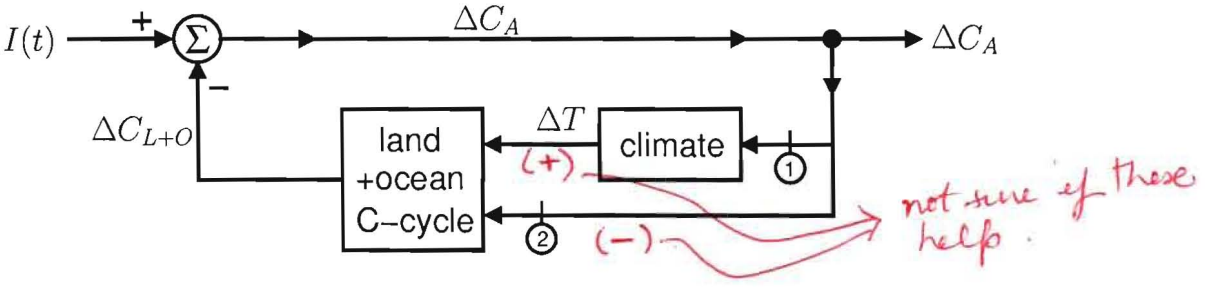

Figure 1. The conceptual model of the feedbacks between climate and carbon cycle underlying the α-β-γ formalism. All quantitities in this feedback diagram are differences to the pre-industrial state of the system. At the point Σ cumulated anthropogenic emissions $I(t)$ and the carbon ΔC_{L+O} taken up by land and ocean are summed to give the change in atmospheric carbon content ΔC_A – actually ΔC_{L+O} is subtracted because by convention land and ocean carbon take-up are considered to be positive. The symbols ① and ② mark the points where links between system components may be cut to disentangle the effects of the feedbacks from the behaviour of the fully coupled model: When cutting at point ① the model is called 'biogeochemically coupled', and when cutting at point ② 'radiatively coupled' (because a change in CO_2 modifies the radiative balance but has no biogeochemical effect). For more details see text.

or response which quantifies changes in response to

Feedbacks enter this picture by the reaction $\Delta C_{L+O}(t)$ of the stored land and ocean carbon to the changes in atmospheric CO_2. To describe this reaction, one introduces the three *sensitivities* α, β, and γ, after which the formalism was named. These characterize the behaviour of the two system components 'climate' and 'carbon cycle' individually. Considering first climate, this component is understood as a black box converting a change ΔC_A in atmospheric CO_2 concentration (input) into a change

in global temperature ΔT (output). Quantitatively its operation may be characterized by the conversion factor (sensitivity)

$$\alpha(t) := \frac{\Delta T(t)}{\Delta C_A(t)}. \tag{2}$$

Here – as for the other two sensitivities below – α is introduced as a time-dependent quantity. Such a time dependence arises because the reaction of the involved system components is not only a function of the input at time t, but also of the inputs at earlier times, i.e. those components have "memory".

In contrast to climate, the combined land and ocean carbon store is pictured with two inputs, namely the change $\Delta C_A(t)$ in atmospheric CO_2 and the temperature change $\Delta T(t)$. The reaction of ΔC_{L+O} to each of these inputs is considered separately by setting the respective other input to its pre-industrial value. These reactions are thus quantified by defining the sensitivities

$$\beta(t) := \frac{\Delta C_{L+O}(t)|_{\Delta T=0}}{\Delta C_A(t)} \tag{3}$$

$$\gamma(t) := \frac{\Delta C_{L+O}(t)|_{\Delta C_A=0}}{\Delta T(t)} \tag{4}$$

Note that we have introduced here the sensitivities α, β, and γ as *difference* quotients, not as *differential* quotients, a conceptual difference that will below get important for our re-interpretation of the formalism.

Next it is assumed that the change in land and ocean carbon induced by a *simultaneous* change in its inputs atmospheric CO_2 and temperature is well approximated by the sum of the individual carbon responses:

$$\begin{aligned}\Delta C_{L+O}(t) &= \Delta C_{L+O}(t)|_{\Delta T=0} + \Delta C_{L+O}(t)|_{\Delta C_A=0} \\ &= \beta(t)\Delta C_A(t) + \gamma(t)\Delta T(t).\end{aligned} \tag{5}$$

where for the second equality the definitions of β and γ have been employed. This relation – that we call 'additivity approximation' – looks rather innocent, but, as will get clear below, its justification needs quite some discussion. Using the definition (2) of α to replace ΔT in this equation, and inserting the result in (1) one obtains

$$\Delta C_A(t) = I(t) - \beta(t)\Delta C_A(t) - \alpha(t)\gamma(t)\Delta C_A(t). \tag{6}$$

Solving for ΔC_A finally gives

$$\Delta C_A(t) = \frac{1}{1+\beta(t)+\alpha(t)\gamma(t)} I(t). \tag{7}$$

This is the desired equation describing how atmospheric CO_2 changes in response to emissions when feedbacks are accounted for. By introducing the *feedback factor*

$$f(t) := -\left(\beta(t)+\alpha(t)\gamma(t)\right), \tag{8}$$

Eq. (7) assumes the standard form known from the general theory of feedbacks (see e.g. (Hansen et al., 1984, Eqs. (4) and (7)), (Peixoto and Oort, 1992, Eq. (2.10)), (Stull, 2017, Eq. (21.25)), (Roe, 2009, Eq. (8)))

$$\Delta C_A(t) = G(t) I(t) \quad \text{with} \quad G(t) := \frac{1}{1 - f(t)}. \tag{9}$$

For the interpretation of this equation it is useful to first consider what happens in the absence of feedbacks. In this case land and ocean carbon remain unchanged by the changing atmospheric CO_2, i.e. $\Delta C_{L+O}(t) = 0$ and all emissions $I(t)$ stay in the atmosphere. Hence the carbon conservation equation (1) reduces to $\Delta C_A(t) = I(t)$, leading by comparison with (9) to $f(t) = 0$; this is the reference case. If $f(t)$ is non-zero, atmospheric carbon is either enhanced above the reference case ($f(t) > 0 \Rightarrow \Delta C_A(t) > I(t)$, positive feedback), or it stays below ($f(t) < 0 \Rightarrow \Delta C_A(t) < I(t)$, negative feedback). The term $G(t)$ thus has the meaning of an amplification factor, commonly called *gain* (but there is a bit of confusion in naming, see (Stull, 2017, p. 807),(Roe, 2009, p. 97)). In the present context the gain has an even more specific meaning: from Eq. (9) it is seen that it may be interpreted as the fraction of emitted CO_2 that remained in the atmosphere, i.e. the gain is here what is commonly called the *airborne fraction* (Gregory et al., 2009).

One may understand Eq. (8) for the feedback factor as *the* major result of the α-β-γ formalism: By writing it as

$$f(t) = f_{bgc} + f_{rad} \quad \text{with} \quad \begin{aligned} f_{bgc}(t) &:= -\beta(t) \\ f_{rad}(t) &:= -\alpha(t)\gamma(t) \end{aligned} \tag{10}$$

one sees that the contributions from the two feedbacks are clearly separated: f_{bgc} depends only on the sensitivity β characterizing the response in what we call the the 'biogeochemical feedback path', while f_{rad} depends only on the sensitivities α and γ characterizing the 'radiative feedback path' (compare Fig. 1). In this way one has found expressions that quantify the contributions from the two feedbacks separately, by which one may e.g. judge their relative importance. Note that following (Gregory et al., 2009) in the literature the two feedbacks are often called 'carbon-concentration feedback' and 'climate-concentration feedback' while we use here, respectively, the terms 'biogeochemical feedback' and 'radiative feedback'.

Besides the feedback factor f and the airborne fraction also some other measures can be derived from the α-β-γ formalism to quantify the size of the climate carbon feedbacks. It is e.g. straightforward to define 'land-borne' and 'ocean-borne' fractions of emissions (Friedlingstein et al., 2006; Arora et al., 2013). And in the first publications of the subject (Friedlingstein et al., 2003, 2006) not the pre-industrial state was taken as a reference, but the biogeochemically coupled simulation, which leads to other definitions of gain and feedback factor. But for the purpose of the present rather formal considerations this doesn't make a difference.

3 Quantification of feedbacks by Earth system simulations

Structurally, the α-β-γ formalism is designed after the original feedback formalism brought into climate sciences by Hansen et al. (1984), being today a standard textbook topic (see e.g. (Peixoto and Oort, 1992; Stull, 2017)). But concerning the experi-

mental protocol used to quantify the feedbacks, there is an important difference: while for the Hansen formalism the simulation experiments are designed to bring the system to a new equilibrium state, for the α-β-γ formalism transient simulations are used. And this has – as we show in the present study – consequences for its interpretation. To prepare for a discussion of these consequences in the next section, we shortly describe here how the α-β-γ formalism is applied to quantify climate-carbon feedbacks by means of Earth system simulations. The experimental protocol has slightly changed between the different phases of C^{4}MIP (for the latest version see (Jones et al., 2016)), but for the present purpose it is sufficient to describe the general ideas underlying the experimental protocol, which remained the same through all phases of C^{4}MIP.

One metric characterizing the strength of the feedbacks is the feedback factor f. To obtain f, one must determine the sensitivities α, β, and γ (see Eq. (8)). These cannot be observed directly (see however (Friedlingstein, 2015; Heinze et al., 2019)) so that one must determine them from simulations with Earth system models (ESMs), that describe the global dynamics of the atmosphere and the oceans together with the global carbon cycle. How suitable simulations must be set up can be read off from the feedback diagram in Fig. 1. First of all, one needs a forcing $I(t)$ that induces changes in the system; as described above, one uses simulations with rising CO_2 for this purpose. To diagnose the resulting changes one must have initialized the system at a reference state, obtained e.g. from a "control" simulation at zero forcing ($I(t) = 0$). For isolating the feedbacks, one needs two further simulations with feedbacks partially switched off by cutting the feedback branches either at point ① or at point ② in the feedback diagram. Cutting at point ①, the climate, represented by global temperature, remains – ideally – unchanged by the emissions ($\Delta T = 0$). Accordingly, in such a simulation, called 'biogeochemically coupled', a change in land and ocean carbon storage is caused only by the direct, non-radiative effect of rising CO_2, the very condition to obtain $\beta(t)$ from the simulation data of $\Delta C_A(t)$ and $\Delta C_{L+O}(t)$ (see Eq. (3)). Cutting instead at point ②, climate changes because of the radiative effect of CO_2 rise. In this simulation land and ocean carbon stores change because a temperature change affects the carbon chemistry of the upper ocean and the biological activities on land, but there is no direct effect of CO_2. Such a simulation, called 'radiatively coupled', realizes the conditions to obtain $\alpha(t)$ and $\gamma(t)$ from the simulation data for $\Delta C_A(t)$, $\Delta C_{L+O}(t)$, and $\Delta T(t)$ (see Eqs. (2) and (4)). Such simulations can be performed in an 'emission-driven' mode, where atmospheric CO_2 is calculated each time step from the balance of injected CO_2 and the carbon ecxchanges with land and ocean, and in a 'concentration-driven' mode, where the changing atmospheric CO_2 concentration is prescribed. In this latter case the cumulated emissions $I(t)$ compatible with the prescribed CO_2 can be diagnosed from the simulation results by the amounts of carbon additionally stored in the system, namely $I(t) = \Delta C_A + \Delta C_{L+O}$ (called 'compatible emissions' (Ciais et al., 2014, Box 6.4, p. 516)).

It should be noted that in practice the necessary separation between the feedbacks from the two branches cannot be fully realized: in the biogeochemically coupled simulation temperature typically changes slightly, among other reasons because of CO_2 induced changes in the transpiration of plants (CO_2-fertilization leads to evaporation cooling; see (Arora et al., 2013, p. 5294)). Note also that this is not the only way to obtain the sensitivities: Alternatively one of the two simulations can be replaced by a full simulation with all feedbacks active to calculate the sensitivities (see (Arora et al., 2013; Schwinger et al., 2014)) leading partially to quantitatively different results (Schwinger et al., 2014; Arora et al., 2020); to keep the following discussion focused, these alternatives will first be ignored before we come back to this issue in the final discussion section.

we clarify

For the present purpose these few remarks seem sufficient. As ~~will get clear~~ below, the important point here is that *transient* simulations are used to quantify the sensitivities so that they depend on time. In the published studies this is often not well visible, because not plots of their time dependence are shown, but only their values at the end of the simulation period (typically $t = 140$ years), usually tabulated for the different models; notable exceptions are (Zickfeld et al., 2011; Arora et al., 2013; Adloff et al., 2018; Williams et al., 2019). Here we show as an example in Fig. 2 time-dependent sensitivities, gain and feedback strengths obtained from the CMIP6 C^4MIP simulations performed with the MPI-ESM1-2-LR Earth system model.

4 Critique of the standard interpretation of the α-β-γ formalism

In the present essay we argue that the way the α-β-γ formalism is understood in the published literature involves a serious misconception. But before making this precise, it is useful to first clarify what assumptions are made to arrive at the gain equation (7) from which the dependence of the feedback factor on the sensitivities could be identified (Eq. (8)).

Surely, underlying the whole formalism is a certain process understanding of the feedback mechanisms in the coupled climate-carbon system. In particular, there is the strong assumption that in a quantitative description of these feedbacks it is sufficient to represent climate by its temperature, instead of accounting also for other aspects characterizing a specific climate (precipitation, momentum distribution, cloud cover, ...). This is indeed constitutive for setting up the α-β-γ formalism, but the misconception thematized here is related to the question whether the *mathematical* assumptions underlying the formalism are consistent with the experimental practice from which the sensitivities are determined, not whether the process understanding is correct. Accordingly, only those mathematical aspects will be considered in the following.

What are these mathematical assumptions? By the above presentation of the α-β-γ formalism it is seen that the only mathematical expression lacking proper justification is Eq. (5) for the combined response to temperature and CO_2 changes – we call this the *additivity approximation* of the responses; all other equations are either definitions, are justified by carbon conservation, or are derived algebraically. This is so obvious because of the particular way we have presented the formalism, differing from other presentations by introducing the sensitivities explicitly by definition, while they are usually introduced implicitly by writing (see e.g. (Friedlingstein et al., 2003, Eq. (6)),(Gregory et al., 2009, Eq. (9)), (Friedlingstein, 2015, Eqs. (2.2), (2.3)),(Arora et al., 2020, Appendix A))

$$\begin{aligned} \Delta T &= \alpha \Delta C_A \\ \Delta C_{L+O} &= \beta \Delta C_A + \gamma \Delta T. \end{aligned} \tag{11}$$

These equations look similar or even equivalent to Eqs. (2) and (5) above. But by omitting the time dependence – in particular the time dependence of the sensitivities – it is tacitly suggested that they may be understood as the lowest order terms of a Taylor expansion in ΔC_A and ΔT, which is only seldomly made explicit (see however (Boer and Arora, 2013, p. 3329), (Schwinger et al., 2014, p. 3871), (Williams et al., 2019, p. 284f)). Usually it is only noted that they are a 'linearization' (see e.g. (Friedlingstein et al., 2003, p. 694), (Arora et al., 2020, p. 4176)). Accordingly, throughout the literature the whole

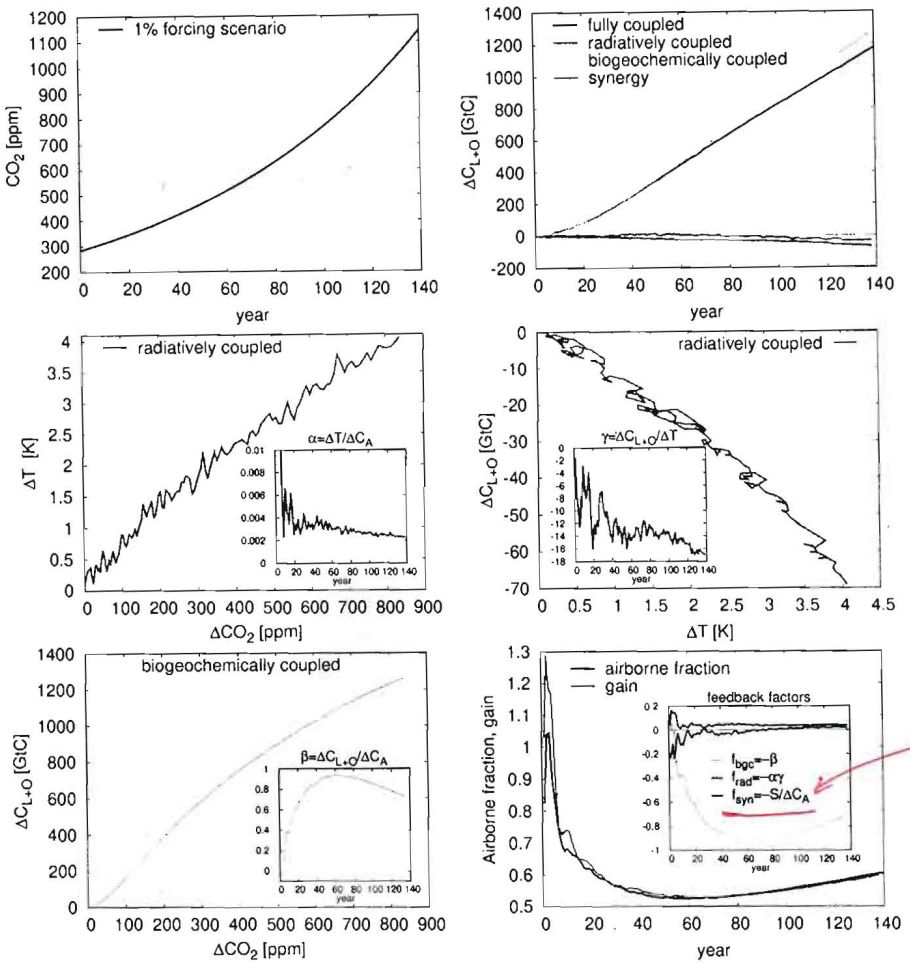

Figure 2. Example for a time dependent feedback analysis applied to the CMIP6 C^{4}MIP 1%-simulatons performed with MPI-ESM1-2-LR. **Top left:** The prescribed atmospheric CO_2 starting at a pre-industrial level of about 280 ppm, rising by 1% per year over 140 years of simulation. **Top right:** The change in land and ocean carbon ΔC_{L+O} found in the fully coupled ('full'), radiatively coupled ('rad'), and biogeochemically coupled ('bgc') simulations, as well as the synergy calculated as 'full - rad - bgc' (compare Eq. (26)); obviously the synergy is rather small compared to ΔC_{L+O} in the full simulation, indicating that the additivity approximation is well fulfilled for this set of simulations. **Middle row:** Data of the radiatively coupled 1%-simulation from which the sensitivities α and γ are derived (insets). **Bottom left:** Data of the biogeochemically coupled 1%-simulation from which the sensitivity β is derived (inset). **Bottom right:** Airborne fraction $\Delta C_A/I(t)$, where the implied emissions $I(t)$ have been calculated as $I(t) = \Delta C_{L+O} + \Delta C_A$ (see text) using the results from the fully coupled 1%-simulation, while the gain G has been calculated by Eqs. (9) and (8) from the sensitivities depicted in the other panels; apparently, the airborne fraction (obtained from the fully-coupled simulation) is well predicted from the gain (obtained from the other two simulations), justifying the applicability of the α-β-γ formalism for these simulations; note that if instead G had been calculated from Eq. (28) obtained by the completed formalism that by accounting for the synergy is exact, it would be indistinguishable from the airborne fraction. The inset shows the feedback factors f_{rad} and f_{bgc} from the two feedback paths (compare Eq. (10)) and f_{syn} characterizing the strength of the combined feedbacks (see Eq. (29)); obviously, the negative biogeochemical feedback strongly dominates the total climate-carbon feedback as is also obvious from the top right panel where in the biogeochemically coupled simulation the land and ocean carbon is much more affected than in the radiatively coupled simulation. – Origin of simulation data: 'full': (Wieners et al., 2019); 'bgc': (Brovkin et al., 2019a); 'rad': (Brovkin et al., 2019b).

Eqn 15 is correct, of course, but it would help to know how you got here. The following provides context.

$$\frac{dx}{dt} = b\,F(t) - \frac{x}{\tau}$$

$$X(t) = X_0 e^{-t/\tau} + \int_0^t b\,F(s) e^{-\frac{(t-s)}{\tau}}\,ds$$

$$\frac{dx}{dt} = b.F - \frac{x}{\tau}$$

$$X(t) = X_0 e^{-t/\tau} + bF\tau(1-e^{-t/\tau})$$

then from here you replace

$X(t)$ by $\Delta X(t)$

which mean $\Delta X_0(t) = 0$ for the pre-industrial case

and this gives you your eqn 15. Correct?

One helpful thing here is that as soon as

F becomes F(t) convolution (ie history/ memory) comes into play.

formalism is understood to be a linear formalism (see the quotes compiled in appendix A). In the following it will get clear that this is a fundamental misconception.

If the α, β, and γ sensitivities could be understood as the linear coefficients of a Taylor expansion then one could write – taking α as an example –

$$\Delta T(\Delta C_A) = \alpha \Delta C_A + \mathcal{O}((\Delta C_A)^2) \quad \text{with} \quad \alpha = \left.\frac{dT}{dC_A}\right|_{\Delta C_A = 0}. \tag{12}$$

We now show that by such a definition all sensitivities are zero so that the first non-zero term in such a Taylor expansion would be quadratic. To demonstrate our claim, we consider in the following the simplified situation of a system with a single time scale – a proof for general systems is given in appendix B.

Global changes in heat and carbon can be described by rate equations. The simplest such equation is

$$\frac{dX}{dt} = bF(t) - \frac{X}{\tau}, \tag{13}$$

where X is a reservoir filled by an input flux $bF(t)$ that is assumed to be proportional to the forcing $F(t)$ with proportionality constant b and with τ in the linear loss term denoting the memory time of the system. In the feedback diagram Fig. 1 this equation is meant either to describe the dynamics of the combined land and ocean carbon reservoir $X = C_{L+O}$, or the climate component understood as a heat reservoir with content cT, where c is its heat capacity so that when absorbing c in the right hand side constants b and τ the equation may as well be understood to describe the global temperature $X = T$. For the forcing we assume a function that mimics a CO_2 or temperature rise:

$$F(t) := F_0 + \Delta F(t), \quad \text{with} \quad \Delta F(t) := \begin{cases} a \cdot t & \text{for} \quad t \geq 0 \\ 0 & \text{for} \quad t < 0. \end{cases} \tag{14}$$

This is confusing me. It's easier to think that $F_0 = 0$. Is that okay?

This forcing has the constant "pre-industrial" value F_0, and starts rising at $t = 0$ with constant slope $a > 0$ – in particular the rise in CO_2 looks in reality more like an exponential, but assuming a linear rise is sufficient here: we want to investigate the behaviour of the sensitivities when introduced as linear expansion coefficients into the forcing, i.e. we are interested in the case of small forcing, ergo in the behaviour at small times, where any smooth forcing may be assumed to be linear.[1]

In agreement with the C^{4}MIP experiment protocol, the system is assumed to be initially in "equilibrium", i.e. $dX/dt = 0$ for "pre-industrial" times so that by Eq. (13) X has for $t \leq 0$ the constant value $X_0 = bF_0\tau$. Setting $\Delta X(t) := X(t) - X_0$, the rate equation takes the form $d\Delta X/dt = b\Delta F(t) - \Delta X/\tau$, whose solution is for $t \geq 0$

$$\Delta X(t) = b \int_0^t ds\, e^{-\frac{t-s}{\tau}} \Delta F(s). \tag{15}$$

unclear

[1] In case the forcing rises quadratically instead of linearly at small times, the definition of the sensitivities as linear coefficients of a Taylor expansion into the forcing fails from the outset and thus must not be considered.

Note also the following

if $F(t) = F(0)\,(1.01)^t = F(0).\,e^{\varepsilon t}$

i.e., our 1 pct CO_2 forcing $\qquad \varepsilon = \ln 1.01$

then the solution becomes

$$X(t) = \frac{b}{\varepsilon + 1/\tau} F(t)$$

i.e., $X(t)$ is again a linear function of $F(t)$

How do I interpret this?

Again, this all correct but please include few more steps here.

Entering the forcing $\Delta F(t) = a \cdot t$ and solving the integral one finds

$$\Delta X(t) = \tau a b [t - \tau(1 - \exp(-t/\tau)]. \tag{16}$$

Here ΔX is a function of t, but to calculate its sensitivity to the forcing, i.e. the analogue of the α, β and γ sensitivities defined as derivative as in Eq. (12), one needs ΔX as a function of the forcing. Such a change in the dependent variable is possible because $F(t)$ is in the range of interest $t > 0$ invertible, namely $t = \Delta F/a$. Using this in (16) one finally finds

$$\Delta X(\Delta F) = \tau b \Delta F - ab\tau^2 \left(1 - \exp\left(-\frac{\Delta F}{a\tau}\right)\right) = \frac{b}{2a}(\Delta F)^2 + \mathcal{O}((\Delta F)^3), \tag{17}$$

where the last result has been obtained by expanding the exponential into ΔF. For the simple system considered, this is the analogue of Eq. (12) from above, except that the Taylor expansion is made here explicit up to quadratic order and obviously this is the first non-zero term. Noting that $d\Delta X = dX$ and $d\Delta F = dF$ one thus finds that the linear expansion coefficient is zero:

$$\left.\frac{dX}{dF}\right|_{\Delta F=0} = 0. \tag{18}$$

Hence, for this simple system, it makes mathematically no sense to introduce its sensitivity to the forcing as linear expansion coefficient. And this holds also for general systems with memory as demonstrated in Appendix B.

This result explains even more clearly why in C^{4}MIP-type simulations the sensitivities are found to be non-constant: To lowest order the response to forcing is quadratic so that there is no linear regime over which the response scales with the forcing. Accordingly, there is even for small forcing no range for which the sensitivities calculated as the difference quotients (2)-(4) are constant, instead their value must vary with time from the outset starting from value zero. Such a behaviour, being incompatible with the assumption of a linear response, is visible for the example of the MPI-ESM1-2-LR simulations displayed in Fig. 2 and has been one of the reasons that motivated this study. In principle one should be able to demonstrate the quadratic dependence in Earth system simulation data, but this behaviour happens at small forcing where the response is hidden by natural variability. Nevertheless, for some model simulations at least indications of this behaviour are recognizable: For the MPI-ESM1-2-LR simulations displayed in Fig. 2 one sees that the sensitivity β is approximately zero at year zero, i.e. at forcing $\Delta CO_2 = 0$ (inset bottom left panel), as it should be for a vanishing linear response. The other carbon sensitivity γ is rapidly varying at small ΔT, so that the behaviour at small forcing is obscured, but with some fantasy one may imagine that the curve of the underlying data $\Delta C_{L+O}(\Delta T)$ (middle right panel) starts with zero slope as expected from Eq. (17). Such a quadratic dependence in the data underlying γ is more clearly visible in the original paper by Friedlingstein et al. (2003, Fig. 5) for the IPSL and Hadley ESMs, and in (Plattner et al., 2008, Fig. 12 e and f) and (Williams et al., 2019, Fig. 2 g) for a whole suite of Earth system models. Indications of the quadratic dependence of the response in land and ocean carbon are also seen in (Chimuka et al., 2023, Fig. 5; case "ramp-up"), (Asaadi et al., 2024, Figs. 3 and 4), and (Williams et al., 2019, Fig. 2 e,f).

The case of α gives an indication of the role of the scenario for seeing the effect of memory on the sensitivities: For α the situation is insofar a bit different as climate reacts much faster to a perturbation than the land and ocean carbon. Asaadi et al. (2024) analyzed the behaviour of several ESMs when after a standard 1% simulation the CO_2 rise is exactly reversed. While for all models the land and ocean carbon show clear hysteretic behaviour (Asaadi et al., 2024, Supplement Figs. S2 c, d, k, l) indicating that memory is of importance, temperature comes at the end of the reversal for two out of five models back to its value at simulation start (Asaadi et al., 2024, Fig. 1(d)). Accordingly, for these two models atmospheric CO_2 changes so slowly that temperature is essentially in equilibrium with the forcing so that one could see a hysteretic effect in α only if CO_2 would change much faster than in a 1% simulation. For the present discussion this means that the vanishing of the sensitivities is of practical relevance only if the forcing changes sufficiently rapidly compared to the relevant memory time scale.

To shed some more light on the origin of the vanishing of the sensitivities, it is instructive to ask, why then the sensitivities appearing in the structurally similar Hansen theory of atmospheric feedbacks (Hansen et al., 1984; Peixoto and Oort, 1992; Roe, 2009; Stull, 2017) are well defined as Taylor expansion coefficients? The main difference is that in Hansen's theory one considers *equilibrium* feedbacks, while the α-β-γ approach tries to specify the sensitivities for a *transient* state. This can formally be made precise as follows. Define the forcing as a function $F(t) = F_0 + q(t)\Delta F_{step}$, where ΔF_{step} is a constant called the *step size*, and the continuous function $q(t)$ determines how the forcing develops from the initial value F_0 to the value $F_0 + \Delta F_{step}$ at some later time T, i.e. formally $q(t \leq 0) = 0$ and $q(t > T) = 1$ and otherwise being arbitrary. For the simple system (13) the solution for this type of forcing – in the following called "step forcing" – is obtained by inserting $\Delta F(t) = q(t)\Delta F_{step}$ into the general solution (15). This gives for $t \geq 0$

$$\Delta X(t) = b\Delta F_{step} \int_0^t ds\, e^{-\frac{t-s}{\tau}} q(s), \tag{19}$$

In Hansen's theory one is interested only in the equilibrium response, i.e. in the response for $t \to \infty$. This gives

$$\Delta X_{equ} := \lim_{t\to\infty} \Delta X(t) \overset{(19)}{=} b\Delta F_{step} \left(\lim_{t\to\infty} e^{-\frac{t}{\tau}} \int_0^T ds\, e^{\frac{s}{\tau}} q(s) + \lim_{t\to\infty} e^{-\frac{t}{\tau}} \int_T^t ds\, e^{\frac{s}{\tau}} \right) = \Delta F_{step}\, b\tau, \tag{20}$$

where the integral from 0 to t was split at time T into two integrals (because of $t \to \infty$ $t > T$ can be assumed) and the final result was then obtained by noting that the first term in the large brackets vanishes, while the second term gives τ because $q(t > T) = 1$. For the sensitivity one thus finds

$$\left.\frac{dX_{equ}}{dF}\right|_{\Delta F_{step}=0} = b\tau, \tag{21}$$

and obviously it is non-zero and independent of the forcing history ($q(t)$ doesn't show up). That by Eq. (20) ΔX_{equ} depends linearly on ΔF_{step} is an artifact of the simplicity of the rate equation model considered here. For more general systems the

relation between ΔX_{equ} and ΔF_{step} contains also terms of higher order in ΔF_{step}, but it can be shown that the equilibrium sensitivity (21) stays to be non-zero and independent of the forcing history.

One may be tempted to compare (18) to (21) to try to understand why in the former the linear coefficient is zero while in the latter it is not; nevertheless, those two derivatives are not directly comparable. The former is the standard derivative of the response $X(t)$ with respect to the forcing $F(t)$ at time $t = 0$ when the forcing starts to increase. The second, however, is rather special: it is the derivative of the function that maps the step forcing at $t \to \infty$ into *equilibrium states* X_{equ}.

One learns from this exercise that there is no trouble with the sensitivities defined as linear coefficients of a Taylor expansion when they are derived from the equilibrium response to a step function forcing. Hence, the problem is not the α-β-γ formalism itself, but the way the sensitivities are determined in simulation experiments, namely by invoking transient simulations. In fact, the whole α-β-γ formalism can without change be applied to determine the strength of the feedbacks for the equilibrium response, as has been done in (Lade et al., 2018) for their analytical Earth system model – this would be one way to circumvent the problems discussed above. Also another conclusion can be drawn from this exercise: This case of an equilibrium response may more generally be understood as a situation where the change in forcing is much slower than the reaction of the system (as likely happens for the climate response defining α in 1% simulations; see above). Hence more precisely, the diagnosed problem arising in application of the α-β-γ formalism to transient simulations doesn't arise generally, but only if the formalism is applied to simulations where compared to internal time scales the forcing changes too rapidly – which is obviously the case for the global carbon cycle in 1% simulations (non-constancy of β and γ; scenario dependence of quantified feedbacks).

5 A re-interpretation of the α-β-γ formalism without linearity assumption

As discussed above, the reason why in the literature the α-β-γ formalism is denoted as linear, is that mistakenly the system is assumed to respond linearly to a sufficiently small forcing. With this assumption the α, β, and γ sensitivities are considered to be the linear coefficients of a Taylor expansion into the forcing. Our presentation of the formalism in section 2 doesn't contain such an assumption. Here the sensitivities (2)-(4) are introduced as difference quotients without any need to understand the involved differences as being small, as would be needed when understood as linear coefficients of a Taylor expansion. As a consequence, the additivity approximation (5) must be justified differently than done in the literature, where it is erroneously explained to emerge from the linear terms of a Taylor expansion of ΔC_{L+O} into ΔC_A and ΔT (see the discussion around Eqs. (11) above).

Such an alternative justification follows from a glance at the feedback diagram 1: The change in land and ocean carbon ΔC_{L+O} is a function of the combined changes of atmospheric CO_2 and temperature, i.e. $\Delta C_{L+O} = \Delta C_{L+O}(\Delta C_A, \Delta T)$; for brevity the time dependence of the involved quantities is not made explicit in this notation. Assuming that the two feedbacks are completely independent, carbon conservation implies that their individual contributions to the change in land and ocean carbon must add, i.e.

$$\Delta C_{L+O}(\Delta C_A, \Delta T) = \Delta C_{L+O}(\Delta C_A, 0) + \Delta C_{L+O}(0, \Delta T), \tag{22}$$

where the first right-hand term arises from the biogeochemical feedback path ($\Delta T = 0$), while the other arises from the radiative feedback path ($\Delta C_A = 0$). Noting that the definitions (3) and (4) of β and γ characterize exactly those individual responses to the forcing, it follows that

$$\begin{aligned}\Delta C_{L+O}(\Delta C_A, \Delta T) &= \frac{\Delta C_{L+O}(\Delta C_A, 0)}{\Delta C_A}\Delta C_A + \frac{\Delta C_{L+O}(0, \Delta T)}{\Delta T}\Delta T \\ &= \beta(t)\Delta C_A(t) + \gamma(t)\Delta T(t).\end{aligned} \tag{23}$$

Here, no linearization has been involved, i.e. none of the changes ΔC_{L+O}, ΔC_A or ΔT must be small, to the consequence that each entry in (23) has the value it has attained at time t of the simulation; more precisely the additivity approximation should thus be written as shown in Eq. (5) above.

To make even more clear that in this way the α-β-γ formalism accounts for the feedbacks to all orders in the perturbation, it is instructive to justify the additivity approximation also in a more formal way. Following the factor separation technique by Stein and Alpert (1993), one may Taylor expand ΔC_{L+O} to *all orders* into its "factors" ΔC_A and ΔT. This gives

$$\Delta C_{L+O}(\Delta C_A, \Delta T) = \sum_{n=1}^{\infty}\sum_{m=0}^{n} a_{n,m}(\Delta C_A)^m(\Delta T)^{n-m}, \tag{24}$$

$$\text{with} \quad a_{n,m} = \frac{1}{m!(n-m)!}\frac{\partial^m}{\partial C_A^m}\frac{\partial^{n-m}}{\partial T^{n-m}}C_{L+O}\Big|_{\Delta C_A=0, \Delta T=0}.$$

After a bit of algebra this can be re-written as

$$\begin{aligned}\Delta C_{L+O}(\Delta C_A, \Delta T) = &\sum_{n=1}^{\infty} a_{n,n}(\Delta C_A)^n + \sum_{n=1}^{\infty} a_{n,0}(\Delta T)^n \\ &+ \Delta C_A \Delta T \sum_{n=0}^{\infty}\sum_{m=0}^{n} a_{n+2,m+1}(\Delta C_A)^m(\Delta T)^{n-m}.\end{aligned} \tag{25}$$

Inspection of the explicit forms of $a_{n,n}$ and $a_{n,0}$ reveals that the first and second right-hand side terms are the Taylor expansions of $\Delta C_{L+O}(\Delta C_A, 0)$ and $\Delta C_{L+O}(0, \Delta T)$.[2] Hence

$$\Delta C_{L+O}(\Delta C_A, \Delta T) = \Delta C_{L+O}(\Delta C_A, 0) + \Delta C_{L+O}(0, \Delta T) + S, \tag{26}$$

where S is introduced as a short hand for the last right-hand side term of (25) denoted as 'synergy' by Stein and Alpert (1993). Except for the synergy S, this is exactly the additivity approximation (22). The synergy term contains only mixed contributions in ΔC_A and ΔT so that the additivity approximation is equivalent to the assumption that the synergy is small compared to the contributions from the individual factors ΔC_A and ΔT. Remembering that the validity of the additivity approximation is mathematically the only assumption entering the above presentation of the α-β-γ formalism, these last considerations clearly

[2]For completeness it may be noted that from the discussion in section 4 it is known that applied to transient states the expansion coefficients of the linear terms vanish, i.e. $a_{1,1} = 0$ and $a_{1,0} = 0$.

show not only that this framework can be justified without any linearity assumption, but that whenever this approximation is appropriate, the framework is valid to all orders in the forcing.

6 Completion of the α-β-γ formalism by accounting for the synergy between feedbacks

Depending on the system considered, the additivity approximation may be more or less appropriate. But independent of its quantitative usefulness, there is a structural reason to introduce it, namely as a necessary element of the α-β-γ formalism to disentangle the two considered feedbacks. This structural function of the additivity approximation is particularly well visible when completing the formalism by explicit inclusion of the synergy S between the feedbacks.

Eq. (26) reads when using the notation of section 2

$$\Delta C_{L+O}(t) = \Delta C_{L+O}(t)|_{\Delta T=0} + \Delta C_{L+O}(t)|_{\Delta C_A=0} + S(t). \tag{27}$$

Starting from this exact equation instead from the approximate equation (5), which was based on the additivity approximation, one reaches by the same reasonings the following modified expression for the gain (airborne fraction):

$$G(t) = \frac{1}{1+\beta(t)+\alpha(t)\gamma(t)+\frac{S(t)}{\Delta C_A(t)}}. \tag{28}$$

Obviously, the synergy shows up here as another contribution to the feedback factor $f(t)$ (compare Eqs. (8) and (9)) so that Eq. (10) generalizes to

$$f(t) = f_{bgc} + f_{rad} + f_{syn} \quad \text{with} \quad f_{syn}(t) := -\frac{S(t)}{\Delta C_A(t)}. \tag{29}$$

Concerning the additivity approximation there is a subtle difference between the previous formulation of the α-β-γ formalism and the present, completed one: while above additivity of the contributions from the two feedbacks was introduced as an approximation, it shows up here as a necessary way to disentangle the individual contributions of the two feedbacks from their combined and not further separable contributions (synergy). Thereby this completed formalism is free of any assumptions that needed to be justified, while the interpretation of the gain G as airborne fraction $\Delta C_A(t)/I(t)$ remains unchanged (compare section 2). Whether the two-feedbacks picture of the climate-carbon system underlying the α-β-γ formalism is appropriate would be seen from the values of the three components of the feedback factor $f(t)$ in Eq. (29): for the picture to be valid f_{syn} must be much smaller than $f_{bgc} + f_{rad}$.

The application of this completed formalism is straightforward when the triple of fully-coupled, biogeochemically-coupled, and radiatively-coupled simulatons is available. While f_{bgc} and f_{rad} can as usual be calculated via α, β, and γ from the last two simulations, to obtain f_{syn} it needs in addition the fully-coupled simulation to determine $S(t)$ from Eq. (27) where data from all three simulations are involved. Otherwise the same cautionary remark concerning a clear separation of the feedbacks in simulation experiments applies also here: The assumption of $\Delta T(t) = 0$ in the biogeochemically coupled simulation is

in practice not exactly realized (see the discussion in section 3). For the MPI-ESM-LR2 simulations depicted in Fig. 2, the calculations from the two variants of the α-β-γ formalism make almost no difference, as seen in the bottom-right panel from the almost zero value of f_{syn} (inset) and the almost identical curves of the airborne fraction calculated directly from the fully-coupled simulation and the gain calculated from the original feedback factor (8), thus giving only an approximation to the airborne fraction because the synergy is not accounted for. In other Earth system simulations the synergy can be rather large (Zickfeld et al., 2011; Schwinger et al., 2014) (but see the remarks on these studies in the next section).

As a bit of a side remark it may be added that the exact gain equation (28) may as well be written as

$$G(t) = \frac{1}{1+\beta(t)+\alpha(t)\widehat{\gamma}(t)}, \tag{30}$$

which looks like the original expression for the gain (see Eq. (7)) except that γ is replaced by[3]

$$\widehat{\gamma} := \gamma + \frac{S}{\alpha \Delta C_A} \overset{(2),(4)}{=} \frac{\Delta C_{L+O}|_{\Delta C_A=0} + S}{\Delta T} \overset{(27)}{=} \frac{\Delta C_{L+O} - \Delta C_{L+O}|_{\Delta T=0}}{\Delta T}. \tag{31}$$

By the last equality one has recovered an expression for $\widehat{\gamma}$ introduced previously by Arora et al. (2020, p. 4188). They advocate to use $\widehat{\gamma}$ instead of γ to characterize the radiative feedback.[4] The idea behind this suggestion is based on the observations by Zickfeld et al. (2011) and Schwinger et al. (2014) that in the radiatively coupled simulation (from which γ is calculated) part of the radiative feedback operating in the fully-coupled simulation is missing, mainly due to a reduced mixing of carbon between upper and deep ocean upon warming. So Arora et al. (2020, p. 4188) suggest to include this missing part into the newly defined $\widehat{\gamma}$ by calculating it from the carbon change in the fully coupled simulation (containing the missing part) corrected by the carbon change in the biogeochemically-coupled simulation (compare last equality in (31)) to keep only the radiative effect. In the light of the present study they thus suggest to include in $\widehat{\gamma}$ the synergy between the two feedbacks, which is well visible by the first equality in (31) involving an explicit dependence on the synergy. Accordingly, when using $\widehat{\gamma}$ instead of γ, the synergy is counted as part of the radiative feedback, even though with the same right one might count the synergy to be part of the biogeochemical feedback (define $\widehat{\beta} := \beta + S/(\alpha \Delta C_A)$). So in this way one is giving up the idea of a clear separation between the contributions from the two feedbacks underlying the α-β-γ formalism. But as shown in the present section, this clear separation may be retained when accounting explicitly for the synergy by calculating all three feedback factors f_{rad}, f_{bgc}, and f_{syn}.

7 Discussion

Even though the α-β-γ formalism for climate-carbon feedbacks has been built after the Hansen et al. (1984) formalism for atmospheric feedbacks, the above considerations revealed that its application to transient instead of equilibrium states of the Earth system requires a different interpretation of its elements. In particular – and in contrast to common understanding (com-

[3] For notational economy the time dependence has been omitted.

[4] This is actually the way γ was calculated in the first C^{4}MIP study (Friedlingstein et al., 2006) using the biogeochemically-coupled simulation as a reference (see the discussion in (Gregory et al., 2009, p. 5244)).

I'm confused. Is linearity the issue here or the memory?

Your conclusion that linear term goes away is only true when

$$F = at \quad \text{or}$$

$$F = at^2 \text{ (as I understand it)}$$

but (as I understand it) when

$$F(t) = F(0) \cdot (1.01)^t$$

then

$X(t)$ is a linear function of $F(t)$

Okay I take back what I wrote above after realizing the following. In the 1pctCO2 run

$$F(t) = F(0) \cdot (1.01)^t - F(0)$$

$$= 285\,(1.01)^t - 285$$

$$= 285\left[\left(1 + \varepsilon t + \frac{1}{2}\varepsilon^2 t^2 + \cdots\right) - 1\right]$$

$$= 285\,\varepsilon t + O(t^2)$$

$\varepsilon = \ln(1.01)$

t and this is still linear

pare appendix A) –, it cannot be justified as the Hansen formalism by a linearization of a Taylor expansion into the forcing. This got evident by demonstrating that because of the memory of the Earth system the sensitivities vanish when understood as linear expansion coefficients. Accordingly, for the considered transient system states, the linear contributions to the feedbacks are strictly zero and a non-zero feedback strength arises only because of nonlinear contributions. From a practical point of view this mathematical statement is only relevant when the internal memory of the system is of the order of or even longer than the characteristic time scale at which the forcing changes. As discussed, for the standard 1% CO_2 scenario this is the case for the response of the land and ocean carbon cycle while the temperature response defining α may follow that 1% forcing without noticeable delay. Accordingly, in practice the problem of vanishing sensitivities arises only for β and γ, but, being indispensable elements of the α-β-γ formalism, this is sufficient to question the common understanding of the whole approach.

That there is a problem with linearity is known from simulation results at least since the paper by Gregory et al. (2009) who clearly stated that the "inconstancy of β and γ indicates that the linear formulas $\Delta C_{L+O}|_{\Delta T=0} = \beta \Delta C_A$ and $\Delta C_{L+O}|_{\Delta C_A=0} = \gamma \Delta T$ are inadequate" (notation adapted). Nevertheless one can make sense of these formulas when, as discussed above, one understands the sensitivities as *difference* quotients (see Eqs. (2) – (4)) that are not intended to approximate a *differential* quotient in the limit of small forcing, which is the usual understanding of the sensitivities when introduced as linear expansion coefficients. With this new understanding, an 'inconstancy' of the sensitivities is rather natural because they are not any more understood as linear expansion coefficients, which was the reason to expect their 'constancy'.

With this modified interpretation of the sensitivities, the only assumption needed to be justified when applying the α-β-γ formalism to transient simulations is the validity of the additivity approximation (5). Usually it is justified by expanding the change in land carbon $\Delta C_{L+O}(\Delta C_A, \Delta T)$ up to linear order into the changes in CO_2 and temperature (see section 4), i.e. the additivity of the individual contributions from CO_2 and temperature is thought to follow from linearization, not recognizing that these linear contributions are zero. But, as shown above, linearity is not needed for its justification, it is sufficient to assume that the synergy arising from the combined action of changing CO_2 and temperature is much smaller than the contributions from the individual changes (see the discussion following Eq. (24)). That this additivity is essentially an independence of the climate and CO_2 effects on the carbon cycle is well known, but has so far not been considered to be a property inherent to the climate-carbon feedback system independent of the response being linear or not. This is particularly evident in the studies (Zickfeld et al., 2011) and (Schwinger et al., 2014) whose declared aim was – see their paper titles – to quantify *nonlinear* contributions to the climate-carbon feedbacks. In view of the present study these "nonlinear contributions" consist of the full size of the feedbacks because, as demonstrated, linear contributions to the feedback vanish in the considered transient simulations. Properly re-interpreted these studies instead investigated the origin and size of the synergy (compare Eq. (26)). This misunderstanding of the synergy as a measure of nonlinear corrections to the otherwise linear feedbacks goes at least back to (Gregory et al., 2009, p. 5244), and arises because linearity is thought to be a *necessary* condition for additivity, while in fact linearity is only a *sufficient* condition for the additivity approximation to be valid as got clear above by demonstrating that it may be justified without assuming linearity. Accordingly, one cannot conclude from a good match between the airborne fraction calculated by the gain equation (9) (that is based on the additivity assumption) and the airborne fraction obtained directly from the simulations "that the linear perturbation assumption ... holds ... i.e. that the changes are small enough to

Note that this is true when $F(t) = at$ when $F(t) = at^2$ but not when $F(t) = F(0) \cdot (1.01)^t$ (okay my bad, I agree)

please consider rewording. Unclear.

Say it even more clearly. 1% per year CO_2 increase << time scales of land/ocean

only true for $F = at$ or $F = at^2$ if I am right

my bad I agree

ignore higher-order terms" (Friedlingstein et al., 2003, p. 697). In fact, as we have seen here, its exclusively the "higher-order terms" that cause the feedbacks in transient simulations and the good match only indicates that the additivity assumption is well fulfilled.

These remarks are not meant to disparage in any way the merits of those studies, for most studies applying the α-β-γ formalism the invalidity of the linearity assumption has no consequence – when ignoring in those studies all remarks relating to linearity and nonlinearity everything is fine. Scientifically affected are studies where the sensitivities are employed as scaling constants. An example is found in (Koch et al., 2019, p. 28), where the authors used the value of γ obtained from C^{4}MIP simulations to estimate the land and ocean carbon uptake for a much smaller temperature change than that from which γ had been determined. Affected might also be studies that draw consequences from the complementarity between linear and nonlinear behaviour. An example for this is found in (Jones and Friedlingstein, 2020, Supplement p. 2) where a loss of linearity emerging at larger forcing is used to explain systematic differences seen between the values of airborne fraction calculated for doubled and quadrupled CO_2 rise from the sensitivities via the gain by Eqs. (9) and (8) – such an argument fails because as shown above the response to the forcing is already to lowest non-vanishing order nonlinear.

While in these examples only side aspects of the studies are affected, the attempt to obtain a more accurate value of β and γ by accounting for the small temperature rise found in the biogeochemically-coupled simulations poses a more serious problem. Arora et al. (2013) discussed how to obtain the β and γ sensitivities from different pairs of the radiatively, biogeochemically, and fully coupled simulations. Starting point is (following (Schwinger et al., 2014, Appendix A)) the additivity approximation Eq. (5), written as

$$\Delta C^E_{L+O} = \beta \Delta C^E_A + \gamma \Delta T^E, \tag{32}$$

where the upper index $E \in \{\text{rad,bgc,full}\}$ specifies to which of the three experiments the respective quantity refers. As discussed in section 2, the two right-hand-side terms emerge from the two feedbacks, the first one from the biogeochemical coupling, the second one from the radiative coupling. Both terms are present for the fully-coupled experiment ('full'), while for the radiatively-coupled experiment ('rad') the first term is zero. In the ideal picture of a complete separation of the two feedbacks (compare Fig. 1) for the biogeochemically-coupled experiment ('bgc') climate (represented by temperature) is unaffected by the CO_2 rise, so that in this case the term involving γ is zero and can be dropped. This is the standard framing underlying the α-β-γ formalism described in section 2. But – as already noted above – in real bgc-experiments temperature doesn't stay constant but undergoes a small change. Accordingly, Arora et al. (2013) keep in Eq. (32) the term involving γ for this experiment even though it doesn't emerge from the radiative feedback of CO_2 but by other processes (e.g. a reduction of transpiration accompanying CO_2 fertilization (Arora et al., 2013, p. 5294)). For each pair of experiments one may solve the respective two equations for β and γ. With a non-zero γ term for the biogeochemically-coupled experiment one obtains for the two experiment pairs that involve the bgc-experiment modified expressions for the calculation of β and γ (see (Schwinger et al., 2014, Appendix A), (Arora et al., 2020, Table 1)). The resulting β and γ values differ from the standard ones by only a few percent (Arora et al., 2020, p. 4178 and Figs. 5 and 6). Nevertheless, this modified approach should not be used because it

485 involves the invalid linearity assumption: The pair of equations can only be solved if in particular γ represents the same quantity in the two experiments. In view of Eq. (5) this means that γ should have the same time dependence in both experiments, which is not the case: In the fully- and radiatively-coupled experiments the temperature rise is very different in pace and size from that in the bgc experiment so that the γ in the bgc-equation characterizes a very different scenario than in the two other experiments – a consequence of the fact that γ is not a linear Taylor expansion coefficient characterizing the underlying system
490 independent of the scenario. To nevertheless rescue the idea to account for the small temperature rise in the bgc-experiment one might be tempted to use in the respective equation the value of γ obtained in the rad-simulation at the very temperature obtained in the bgc-simulation. But this is equally questionable because in the rad-simulation that temperature is reached much earlier than in the bgc-simulation so that the land and ocean carbon has because of the memory adapted differently to the temperature rise. Hence the only way out is to assume that the temperature rise in the bgc-simulation is much smaller than in the
495 other simulations so that it can be ignored, meaning that only the standard framing of no temperature rise in the bgc-simulation is conceptually sound.

Note that the whole problem arises only because of application of the α-β-γ formalism to transient simulations. If instead applied in the spirit of Hansen et al. (1984) to equilibrium simulations, all sensitivities would be scenario independent and one could indeed account for the small temperature rise in the bgc-simulation as suggested by Arora et al. (2013).

500 A similar problem has earlier been recognized by Plattner et al. (2008, p. 2741) for the application of the α-β-γ formalism to emission-driven simulations (see also the discussion in (Zickfeld et al., 2011, p. 16f)). They noted that the way γ has been calculated from such simulations in (Friedlingstein et al., 2003, 2006; Denman et al., 2007) assumes that β is independent of the size of the CO_2 rise – more precisely, β obtained at one CO_2 concentration is used to infer the change in carbon uptake of land and ocean at another CO_2 concentration by linear scaling. By correcting for this invalid application of a single β, Plattner
505 et al. (2008) obtained much smaller γ values, more consistent with values found in concentration-driven simulations.

Overall, the present re-interpretation of the α-β-γ formalism rescues the common practice to apply it at large forcing, that would not be justified when understanding the formalism as being linear. Moreover, it explains the for long recognised scenario dependence of its results. The present considerations have also shown that in future applications it may be useful to estimate in addition to the individual contributions f_{rad} and f_{bgc} of the two feedbacks to the overall feedback also their synergistic
510 contribution f_{syn} because in relation to the individual contributions it is a measure to what extent the idealized picture of the two separate feedbacks underlying the α-β-γ formalism is appropriate.

Another consequence of the present study is that the memory of the system cannot be ignored when quantifying the size of the feedbacks in transient systems, a topic so far not well investigated (see however (Chimuka et al., 2023)). That the memory is important is indeed not surprising as the 'committed change' in climate and carbon cycle (see e.g. (Arias et al., 2021, Box
515 TS.1, p. 39)), happening because of the delayed response even after anthropogenic perturbations have come to a halt, is known to be not ignorable when thinking e.g. about emission pathways suitable to stabilize future climate (Corner et al., 2023). That the memory must be understood as part of the feedback problem is also obvious from a general systems perspective because generally the response of a dynamical system should depend on the whole history of the forcing (compare appendix B). Such a perspective is underlying the 'generalized α-β-γ formalism', where the sensitivities are replaced by linear response functions

A reader who hasn't read this paper will not follow what this means.

(Torres-Mendonça et al., 2024) that characterize the system as such, independent of the perturbation scenario. A major result of this more general approach to climate-carbon feedbacks is that because of the memory the feedback gain is different at different time scales. In this respect the equilibrium response addressed by Hansen et al. (1984) is very different: they quantify the strengths of the feedbacks after all delayed responses from the initial perturbation have faded away, i.e. the effect of the memory is excluded by construction of this approach to feedbacks.

↑ Umm! Perhaps a different word than "faded".

In closing this discussion, it may be remarked that the present study provides a counter example to the claim by Roe (2009, p. 99) that "feedbacks are just Taylor series in disguise". In fact, our re-interpretation of the α-β-γ formalism manages without any expansion into the perturbation, because it is valid to any non-linear order.

8 Outlook

For future investigations of the size of climate-carbon feedbacks we see three options. First, one may continue as in previous phases of C^{4}MIP with the quantification of feedbacks in *transient* simulations by means of the standard α-β-γ formalism. In this case, the present re-interpretation of the formalism suggests abandoning the practice of computing feedbacks only over a fixed time span in favor of investigating instead how feedbacks develop in time, in particular also in scenarios that lead to a stabilization of climate, a topic of public interest (see also (MacDougall, 2019)). While it is a clear advantage of the standard α-β-γ formalism that – as shown – can be applied to any scenario, a strong disadvantage is that its results apply only to the very scenario from which they are derived. Following (Asaadi et al., 2024, p. 414) one can thus conclude that this approach "... should be seen as a technique for assessing the relative sensitivities of models and understanding their differences ... rather than as absolute measures of invariant system properties".

A second option without this disadvantage would be to switch to the quantification of *equilibrium* feedbacks. Such an approach would be consistent with the calculation of the various physical feedbacks: In the last IPCC report, results for the radiative forcings of physical feedbacks were shown separately from those obtained for climate-carbon feedbacks (see (IPCC, 2021, Figures TS.17(a) and TS.17(b) on p. 96)), because of their incomparability despite common units – the former were, following (Hansen et al., 1984), scenario-independent equilibrium feedbacks, while the latter were the scenario-dependent values from transient C^{4}MIP simulations. As the present study has shown, the α-β-γ formalism is for equilibrium feedbacks well justified by a Taylor expansion up to linear order in the perturbation. Accordingly, equilibrium results on the size of radiative forcing obtained from the α-β-γ formalism could be well integrated into a common picture with the radiative forcings for purely physical feedbacks. A drawback is that such results are limited to the considered equilibrium states, thereby ignoring any memory of the system that is of eminent importance for any realistic scenario of climate change.

This limitation is overcome by the third option, the application of the generalized α-β-γ formalism (Rubino et al., 2016; Enting and Clisby, 2019; Enting, 2022; Torres-Mendonça et al., 2024): here the α, β, and γ sensitivities are generalized to linear response functions that include the whole information on the memory of the system. As the equilibrium sensitivities, these characterize the system as such, i.e. independent of the scenarios from which these functions are derived. But in contrast to the equilibrium approach, with the generalized α-β-γ formalism one can predict the behaviour of the coupled climate-carbon

system not only in *equilibrium* but also in *transient* scenarios, as long as the underlying linearity assumption concerning the perturbation strength is fulfilled. And unlike the other two approaches, this framework permits investigating how the memory of the system shapes the feedback at different time scales and how models differ in this respect. As demonstrated in Torres-Mendonça et al. (2024) standard C^4MIP simulations can be used to determine the α, β, and γ response functions needed to calculate the time scale dependent feedback sizes, but in order to reduce the technical burden in determining them one should employ more appropriate types of simulations, e.g. step-type simulations as suggested in a somewhat different context by (Lembo et al., 2020).

In deciding among these three options, it may help to consider generally the purpose of a forcing-feedback framework in the study of complex systems with feedback loops. In our view such a framework should provide insight into the way internal processes shape the response to a forcing, thereby disentangling what is internal (feedbacks) from what is external (forcing) to the system. This is indeed achieved both by the equilibrium approach to feedbacks and by the generalized α-β-γ formalism, because both frameworks quantify the feedback strength as an inherent system property. The same cannot be said of the standard α-β-γ approach: here, even under our re-interpretation, the feedback strength depends on the forcing, so that internal and external components are not clearly separated.

Appendix A: Quotes from the literature claiming the 'linearity' of the α-β-γ formalism

In this appendix we compiled a number of quotes from the existing literature on climate-carbon feedbacks to demonstrate that the (erroneous) understanding of the α-β-γ formalism to be based on linearity is rather pervasive.

quote	reference
"linearised feedback framework"	C^4MIP website c4mip.net/background (accessed 2024-05-14)
"The coupling between the carbon cycle and the climate system can be linearized by the following set of equations ..."	(Friedlingstein et al., 2003, p. 694)
"linear sensitivity parameters α, β, and γ"	(Plattner et al., 2008, p. 2738)
"assuming linearity"	(Gregory et al., 2009, p. 5244)
"assuming that the concentration-carbon response is proportional to C, the climate-carbon response is proportional to T, and that they can be combined linearly"	(Gregory et al., 2009, p. 5248)
"The carbon sensitivities reflect the linearity in the response of the coupled climate-carbon cycle system to elevated CO_2 and climate change."	(Zickfeld et al., 2011, p. 4271)
"if the higher-order terms in the expansion ... are small"	(Boer and Arora, 2013, p. 3329)
"expanding in Taylor series"	(Boer and Arora, 2013, p. 3329)
"assume an approximately linear response of the globally integrated surface-atmosphere CO_2 flux in terms of global mean temperature and CO_2 concentration change"	(Arora et al., 2013, p. 5291)
"The carbon cycle feedback ... can be derived based on a Taylor series expansion"	(Schwinger et al., 2014, p. 3871)
"The linear feedback analysis with the β and γ metrics of Friedlingstein et al. (2006)"	(Ciais et al., 2014, p. 519)
"assuming that the carbon cycle responds linearly to atmospheric CO_2 and climate change"	(Friedlingstein, 2015, p. 4)
"the linearity assumption implicit to the C^4MIP feedback analysis"	(Adloff et al., 2018, p. 415)
"$\beta \equiv \partial C_{L+O}/\partial C_A|_0$ and $\gamma \equiv \partial C_{L+O}/\partial T|_0$" (notation adapted)	(Williams et al., 2019, p. 285)
"the feedback framework makes assumptions of linearity"	(Jones and Friedlingstein, 2020, in Supplement)
"These equations assume linearization ..."	(Arora et al., 2020, p. 4176)
(concerning failure of exact additivity) "... this is not the case because of the nonlinearities involved"	(Arora et al., 2020, p. 4176)
"Linear Feedback Analysis" (section title); "the traditional linear feedback approach"	(Canadell et al., 2021, p. 735)

Appendix B: Proof that generally for a system with memory the sensitivity vanishes

Considering a general system with memory, it is shown here that its sensitivity, understood as linear expansion coefficient into the forcing, vanishes.

Proposition. *Consider a general physical system with memory. Formally, such a system may be understood as a functional, mapping a real valued function $F(t)$ (input) of time t to a real valued function $X(t)$ (output). By 'physical' it is meant that the system is causal and time-invariant (i.e. the system response depends only on the time elapsed). Let the sensitivity κ of the*

system be defined by

$$\kappa := \lim_{t\to 0} \frac{\Delta X(t)}{\Delta F(t)} = \left.\frac{dX(t)}{dF(t)}\right|_{F=F(0)}, \tag{B1}$$

where $\Delta X(t) := X(t) - X(0)$ *and* $\Delta F(t) := F(t) - F(0)$. *Under the additional – rather weak – assumptions that* $F(t)$ *is (multiply) differentiable and that the system output varies continuously when its input varies continuously, it follows that* $\kappa = 0$.

Comment: The definition of κ is analogous to the definitions of α, β, and γ in Eqs. (2)-(4), except that by these equations they have been defined as difference quotients, while by taking the limit $t \to 0$, κ is a differential quotient. Thereby the definition of κ mimics that of α, β, and γ in the standard interpretation of the α-β-γ formalism, namely as the linear coefficient of a Taylor expansion (here the expansion of X into F).

Proof. Given the assumed continuity, it is known that the functional mapping $F()$ to $X()$ may be represented by a Volterra expansion (Fréchet, 1910; Schetzen, 2010). Assuming for simplicity that $X(t=0) = 0$ and $F(t=0) = 0$, this Volterra series representation of a physical system is given by

$$X(t) = \sum_{n=1}^{\infty} \int_0^t ds_1 \int_0^t ds_2 \ldots \int_0^t ds_n \, K_n(s_1, s_2, \ldots, s_n) F(t-s_1) F(t-s_2) \ldots F(t-s_n). \tag{B2}$$

For a system with memory the "Volterra kernels" $K_n()$ are continuous functions, so that, weighted by the kernels, the integrals accumulate the forcing from past times $t - s_i$. Here only the behaviour for $F \to 0$ is of interest and this behaviour is determined only by the term of the Volterra expansion linear in F because the higher order terms vanish faster than linear in that limit. Accordingly, it is sufficient to consider the truncated expansion

$$X(t) = \int_0^t ds\, K(s) F(t-s) + \mathcal{O}(F^2), \tag{B3}$$

where the index "1" of K_1 and s_1 has been dropped. By the chain rule one obtains

$$\kappa = \lim_{t\to 0} \frac{dX(t)}{dt} \left(\frac{dF(t)}{dt}\right)^{-1}. \tag{B4}$$

Plugging in (B3) and performing the derivative gives as claimed

$$\kappa = \lim_{t\to 0} \left(K(t)F(0) + \int_0^t ds\, K(s) \frac{dF(t-s)}{dt} \right) \left(\frac{dF(t)}{dt}\right)^{-1} = 0. \tag{B5}$$

This conclusion is obvious when $\lim_{t\to 0} dF/dt \neq 0$. In case this derivative vanishes for $t \to 0$, one can apply repeatedly L'Hôpital's rule to the quotient of integral and F-derivative until some higher derivative of F in the numerator is non-zero – and one of them must be non-zero because otherwise $F(t) = 0$ for all t. □

Comment: This proof is given here to demonstrate that by the memory of the system the sensitivities get zero. One may wonder where exactly in this proof the memory comes in: For a system without memory the kernels $K()$ in (B2) would be products of δ-functions so that the Volterra expansion reduces to a Taylor expansion of $X(t)$ into $F(t)$ *at the same time* t, meaning that the system follows the perturbation immediately. Formally, the difference between $K()$ being a δ-function or a continuous function comes in in Eq. (B5), where the integral would not vanish for $K()$ being a δ-function.

Author contributions. The idea for the study was jointly developed by both authors. CHR developed the methodology, did the formal and numerical analysis, the plotting and wrote the first draft. This was then content- and textwise refined and partly rewritten by both authors.

Competing interests. The authors declare that they have no con- flict of interest.

Acknowledgements. CHR thanks Martin Claussen for hinting to a connection between the α-β-γ formalism and the factor separation technique. The authors thank Thomas Kleinen for his concise internal review at MPI. GLTM was supported by research grant no. 2023/04579-5 from the São Paulo Research Foundation (FAPESP)