Reviewer Responses

Reviewer comments are in red. Our responses are in black. Quoted text that are changes to the manuscript are *italicised*.

In this paper the authors use a conceptual model and a more complex model to investigate possible climate carbon cycle instabilities. They find that whether such an instability can occur depends on the equilibrium climate sensitivity of the model and the strength of CO2 fertilization. This paper is a very interesting read, using some really cool methods and producing some interesting results. I believe the paper will make a valuable contribution to the field and is very suitable for the journal. I believe major revisions are in order before it can be published though. I've listed some main comments/concerns regarding the paper, followed by a few more specific comments.

We thank Reviewer One for a thoughtful and thorough review of the paper. We give our responses below.

Main Comments

Framing

Currently, part of the framing is that some Earth System Models (ESMs) have a high Equilibrium Climate Sensitivity (ECS), which might result in climate-carbon cycle instabilities in these models. I think how it is framed in the introduction is fine. However, I'm not sure whether how it is treated in the penultimate paragraph of Section 4 (lines 294 – 298) works that well, since if ESMs would suffer from these type of instabilities, they would be tuned such that the instabilities wouldn't occur. From a steady state perspective, I do not expect ESMs to have these instabilities. From a transient perspective then it might be relevant. I suggest reframing the conclusions in Section 4 to take this into account.

Whilst it is possible that the instabilities may be deliberately tuned away, ESMs are not typically run in a configuration where this instability could appear. This is because, in general, CMIP runs are performed with a prescribed CO2 concentration. This is instead of an interactive global carbon cycle, where the concentration is dynamically updated due to respiration and other fluxes. Hence, prescribing CO2 means that these feedback loops are cut. To reflect this, we add to the discussion: *"It is natural to ask if any CMIP6 models are unstable. Whilst it is possible that modelling groups may `tune away' this instability, it may persist in other models. One reason that the instability may be that this instability occurs slowly, so it may not be detectable over the typical duration over which ESMs are run for. More fundamentally, the most widely used CMIP experiments (such as piControl, 1pctCO2, abrupt-4xCO2, historical and the SSP scenarios) are forced with prescribed concentrations of CO2, rather than dynamically updating CO2 concentrations due to the carbon cycle. This cuts the feedback loop between warming and*

respiration. Therefore, even if ESMs were unstable, this would not be apparent in typical CMIP simulations."

Conceptual Model

I think the use of a conceptual model can be very powerful, however, I think in this paper the description of the model and the assumptions should be extended.

Upon skimming the Cummins et al. (2020) paper, I realized that T1 and T2 are actually supposed to be temperature anomalies. Is this correct? This is not mentioned in the text I think. Obviously, this is extremely important to be able to properly interpret the results.

T1 and T2 are indeed anomalies. To clarify this, we write: *"One box represents the globally averaged temperature anomaly of the atmosphere and upper ocean, T_1, and the other represents the globally averaged temperature anomaly of the deep ocean, T_2."*

It is unclear to me to what extent the model is based on previous work. To what extent is the model original to this paper? From what I understand now, (part of) the model is indeed based on some earlier work. I suggest making this connection clearer.

That is correct, the model we use is a combination of existing models of different Earth System components. To clarify we add *"We do this [construct the model] by combining existing models of components of the climate-carbon cycle system".*

I suggest explicitly stating that the temperature state variables and the carbon state variables are not representing the same 'box'.

This is a good point as the upper temperature box also accounts for the atmosphere. We add *"[ocean carbon] boxes do not partition the ocean in the same way the temperature boxes do."*

I suggest explicitly stating the units of the state variables.

We add units (of Kelvin and PgC) in brackets after the state variables are introduced.

How much carbon is there in the system, i.e. what is the sum of equation 2? Are the results sensitive to this quantity? I'd say this is especially important when put in the context of anthropogenic emissions that would raise this quantity on the timescales assessed in this paper.

The stores of atmospheric and land carbon are free parameters of the model; they are set to pre-industrial values. However the total carbon is dominated by the ocean carbon which is a function of the ocean parameters. Our ocean model parameters were chosen to get the historical ocean carbon sink within observational estimates. As our ocean carbon model has only two boxes, it is not suitable to simulate the size of the total ocean carbon store on multi-millennial timescales. However, as it is the ratio of the carbon stored in the atmosphere to the

carbon stored on the land that appears in equation (13), the stability of the system is not dependent on the absolute size of the total ocean carbon store. To clarify this we add: *"We fit to the ocean uptake rather than the total ocean carbon as the response to elevated CO2 plays an important role in the system's stability. As our ocean carbon model has only two boxes, it is not suitable to accurately simulate the size of the total ocean carbon store on multi-millennial timescales."* and *"As it is the ratio, rather than the sum, of carbon stored in the atmosphere to carbon stored in the land that appears in equation (13), the stability of the system does not depend straightforwardly on the total carbon store."*

The assumption for the no temperature sensitivity in soil carbon stems from the Varney et al. (2023) paper. Skimming through this paper, I suspect the main motivation for this assumption is found in Fig. 10c where it shows that more than 50% of the changes can be explained by changes in CO2. Do I understand it correctly that this means that the rest of these changes are related to climate change? Looking at Fig. 10 these can still be relatively large for some models. Is this all temperature or also other changes in the climate system?

We do not assume 'no temperature sensitivity in soil carbon'. Instead, we assume that the rate of specific soil respiration (i.e. the soil respiration per unit of soil carbon) doubles for every 10K of global warming. Soil carbon is however also affected by changes in litter-fall, which is approximately equal to NPP on the timescales considered here. Soil carbon is therefore dependent on both CO2, through the CO2 sensitivity of NPP, and climate, through the temperature sensitivity of soil respiration. Varney et al. 2023 is used to justify our assumption of neglecting the temperature dependence of NPP on a global scale. In that paper, the individual sensitivities of NPP to CO2 and temperature are isolated using the C4MIP idealised simulations with CMIP6 Earth system models (Jones et al. 2016). This can be seen the most clearly in their Figure A2, where the total future change in NPP (left column) is seen to be primarily due to the sensitivities to CO2 (middle column) and not the sensitivity to temperature (right column).

I am not satisfied with how the assumption of no temperature dependency in the solubility of CO2 in the ocean is treated. I find this a rather strong assumption without citing previous work or giving a good indication on why this assumption is okay to make. From my perspective, writing in a temperature dependency for k should be doable, plus it would add an additional positive feedback to the system. E.g. use the equation of Weiss (1974) for K0, which as I understand it should be the 1/k parameter in your model. An assumption for salinity needs to be made, which I would say is more valid than the no temperature dependency assumption made now. However, if T1 and T2 are indeed temperature anomalies then adding the temperature dependency might be a bit more difficult. Looking at the values of k, k1 and k2 they appear to be taken at a T0 of 10C, so one possibility is to use T0 + T1 in the Weiss (1974) equation as temperature. Though since T1 also represents the atmosphere this might also not be a valid assumption.

You are correct that neglecting the temperature dependence of solubility is a strong assumption. We neglected it to keep the analysis simple and because the original formulation of our ocean

model (Glotter et al 2014) also made this assumption. Glotter et al offer an extension of their model accounting for the temperature dependency of solubility using essentially Weiss's equation for K0. Working with this extension algebraically is tricky as it makes a matrix element in the Jacobian non-zero. However, we can work with it numerically relatively easily. Doing so causes a small shift in the critical ECS from 10.9K to 10.2K.

As this effect is reasonably small and accounted for in the more complex model JULES/IMOGEN we feel justified in neglecting it in the simple model. We add some extra justification: *"Following Glotter et al 2014, we neglect the role of temperature change in carbon uptake, as this effect is smaller than the direct effect of increased CO2 concentrations (Arora et al 2020)."* and discuss this in the discussion.

As far as I understand it now, the carbonate chemistry is solved for by assuming alkalinity is equal to carbonate alkalinity. It is not clear in the text that this is assumed. Furthermore, the implications of this assumption are also not mentioned. For example, pH values are typically 0.15 – 0.20 lower using this method compared to more sophisticated methods (Munhoven, 2013). I suggest being clearer about this assumption and the implications of the assumption.

We clarify this, now adding to the paper: *"[…] carbonate alkalinity, Alk, in the ocean. We note that this assumption leads to lower ocean pH values than more sophisticated methods (Munhoven 2013)."*

Do the uptake rates in the ocean also capture processes related to the biological and carbonate pumps?

Although there is no explicit modelling of these processes, we can view the uptake rates as effective parameters that capture their effects. We add: *"These timescale parameters capture the effects of carbon pumps, which transport carbon to depth."*

What I think should be made more explicit is for what timescales this model is valid. This is also relevant for simulations with different CA* as shown in Fig. 5.

We agree, and write: *"Very slow carbon cycle processes, such as the flux of carbon into and out of the solid Earth, and variations in astronomical forcing are neglected. As a result this model is not valid for longer than millennial timescales."*

How I interpret the model is that CA* does not necessarily represent pre-industrial CO2 concentrations but the stable CO2 concentration on a certain timescale. Am I correct in this? If so, I suggest clarifying this.

This is an important point, so we clarify this by writing *"Although C_A^* was introduced as the*

amount of carbon stored in the pre-industrial atmosphere, we can generalise this to the case where C_A^* represents a potential atmospheric carbon steady state, and measure temperature anomalies relative to the global temperatures the value of C_A^* implies. We can then determine how stable or unstable that state would be. This allows us to consider how the amount of carbon in the atmosphere affects the stability of the system."

Currently there is only a few sentences on the assumptions in the discussion. I think the assumptions and the implications of these assumptions, as well as the timescales involved in this model, should be more thoroughly discussed in Section 4.

The discussion has been substantially rewritten. We discuss more fully the assumptions made as well as the timescales the simple model operates on. For example, we discuss the effect of our assumptions about the value of alkalinity, CO2 solubility and that our assumptions will break down after the bifurcation.

Description of IMOGEN/JULES

I think the description of the model setup, including assumptions, especially with regards to the carbon cycle model, could be more extensive. For me specifically I would like to know more how IMOGEN works and whether the coupling between IMOGEN and JULES is in one or both directions

We have extended the description of JULES/IMOGEN, focussing on the coupling between them. We add: *"Uptake of carbon by the oceans in IMOGEN depends both on ocean temperatures and atmospheric CO2 concentrations."* and *"with ocean heat uptake parametrised diffusively"* and *"JULES and IMOGEN are fully coupled. IMOGEN provides JULES with meteorological forcing, which affects the terrestrial carbon cycle modelled by JULES. JULES then provides land-atmosphere carbon fluxes to IMOGEN, as the terrestrial land stores evolve. IMOGEN combines this flux with the ocean-atmosphere carbon flux to update atmospheric CO2, which leads to changes in climate and thus feeds back to the meteorological forcing given to JULES."*

Results

It is also stated in Section 4, but I think it would be good to also note it in Section 2 that the limit cycle shows behaviour in which the model assumptions, including the timescales resolved, are not valid anymore.

We add: *"It should be noted that the amplitude and period of this limit cycle are large enough that the system has been pushed into a regime where the assumptions in the model are no longer strictly valid."*

I think it is very important to very explicitly spell out the physical mechanism behind the instabilities and the limit cycle in the conceptual model.

We add: *"The limit cycle emerges because carbon is transferred back and forth between the land and the atmosphere. When there are sufficient stocks of terrestrial carbon, the positive feedback between heterotrophic respiration and global warming moves carbon from the land to the atmosphere. As there is now less carbon in the land, heterotrophic respiration decreases and, due to the CO2 fertilisation effect, NPP increases. This means that carbon now instead flows back into the land from the atmosphere, and so the cycle can begin again."*

JULES is a more sophisticated model, if the mechanism in JULES is similar to the mechanism in the conceptual model, the results will be much more powerful. However, the mechanism in JULES is not really discussed. Is the underlying mechanism that causes the instability in JULES the same as in the conceptual model? This comparison is essential for me to lend credibility to the results of the conceptual model.

In a model as complex as JULES, it is difficult to tell exactly what the cause of a particular instability is. However, by plotting supplementary figures S4 & S5 which show the changes in soil and ocean carbon during these experiments, we can conclude that the rise in CO2 comes from the carbon in the soils, rather than the oceans. Soil carbon in JULES is set in a similar way to the simple model as the balance between litterfall (essentially NPP) and respiration. As a CO2 perturbation increases NPP in JULES, the loss of soil carbon must be driven by increases in respiration. Respiration in JULES is parametrised in the same way as in the simple model with a Q10 form, therefore we can have confidence that the same instability mechanism is at work.

To aid credibility and support interpretation of our calculations, we add: *"This unstable growth in JULES/IMOGEN is associated with a decrease in global soil carbon (see supplementary figure S4). Furthermore, soil carbon and heterotrophic respiration in JULES are modelled in a similar way to how we have modelled them in the simple model, suggesting that the instability in IMOGEN/JULES and the instability in the simple model share a common mechanism."*

There is no discussion about how certain it is that the JULES simulations above 11K ECS are actually unstable. In the conceptual model there are internal oscillations on longer timescales than the duration of the JULES simulations. Would it be possible to extend one of the simulations with e.g. another 5000 years to be a bit more certain that the model is moving towards a runaway state?

Unfortunately due to computational and structural limitations within the model, we cannot run the JULES model for longer than we do. For example, the impulse-response formulation within IMOGEN (which assumes a maximum run length of 5000 years) involves at each timestep

integrating over the history of the model, so adding extra timesteps becomes progressively harder. We have run simulations for very high ECS (45K) which blows up on this 5000 year timescale, rather than oscillates, however this does not rule out the possibility that what we see in Fig 6 is the start of a stable oscillation. To reflect this, we add: *"The CO2 changes are all consistent with exponential growth or decay, although it is possible we are detecting a very long period oscillation."* We also discuss this shortcoming in the discussion (please see below).

Discussion

As mentioned in the previous comments, I would like a more thorough discussion on the assumptions in the model and their potential effect on the results and conclusions. Also, a discussion on the used parameter values from Table 1 would add value I think (how certain/realistic are these values? How sensitive is the model to their values?).

What is still missing, in my opinion, is how these results compare to what is found in the literature. In the introduction already a few studies were named. Studies focusing on the marine carbon cycle are for example, Rothman (2019) and Boot et al. (2022). There could also be a connection made to paleo events, e.g. the Paleocene-Eocene Thermal Maximum (PETM). For some more conceptual work see e.g. Arnscheidt and Rothman (2021). The literature mentioned here are just suggestions and do not have to be included.

We have substantially rewritten the discussion section, including a discussion of the model limitations and the parameters used. We also try to situate the study in terms of other work, particularly with regards to paleoclimate.

The CO2 fertilisation effect plays a central role in the results, but I didn't see a reference to what is actually realistic. Do we know what the values are for ESMs?

Although the value varies between ESMs, the value in the table is typical. We use the value estimated by JULES. For example, Varney 2023 shows that the change in NPP for UKESM (of which JULES is a part) is close to the ensemble average for CMIP6 models. We add: *"The size of the CO2 fertilisation effect differs across CMIP6 ESMs, although the JULES prediction is close to the ensemble mean (Varney 2023)."*

Specific and Technical Comments

Figure 1: 'Increased CO2 solubility' should be 'Decreased CO2 solubility' I guess. You could also include ocean acidification in there as a positive feedback.

The figure has been modified.

Line 23: I suggest rewriting this sentence, specifically the 'even here' part.

It now reads: *"The carbon cycle has played an important role here, absorbing about half of the anthropogenic emissions of CO2 through land and ocean carbon sinks (Canadell 2021)."*

Line 42: Would it make sense to mention quantitative results from the Cox et al. (2006) study here?

We add*: "respiration increasing by a factor of 6 or more for every 10K of warming"*

Line 53: 'and and'.

Thank you.

Figures (general): all figures (except Fig. 1) miss certain text on the labels and the tick marks.

Apologies - the figures were mangled when uploaded to ESD.

Figure 3: I suggest mentioning explicitly which panel (top or bottom) represents without CO2 fertilisation and including CO2 fertilisation. I also suggest switching the order of the two for two reasons: (1) the case with CO2 fertilisation is mentioned first in the text, and (2) in Fig. 4 it is also switched, i.e. first with fertilisation then without.

Done

Line 206: I guess the reference should be to Figure 3 not Figure 2.

Corrected

Figure 5: As mentioned earlier, I would not call CA* 'pre-industrial' but something like steady state concentration (though CA is I think not defined as a concentration).

We make sure to refer to CA* as a steady state concentration when not discussing the pre-industrial state.

Figure 6: Can you explain a bit more what the dotted line is, and how it is determined? I'm not sure whether it is necessary to include an elaborate discussion in the text or caption, but I'd like to know a little bit more about it.

If JULES is linearly unstable, we would expect the change in CO2 concentrations to rise exponentially after an initial transient, i.e. Delta CO2 = A exp(gamma t). Writing this as log Delta CO2 = log A + gamma t we can regress log Delta CO2 against time to extract gamma and A. With these assumed parameters, we plot in the dashed line Delta CO2 = A exp(gamma t), which is a straight line on a log plot. We add: *"In dotted lines we plot fits of exponential growth or decay to the latter portion of the time series."*