Many studies have shown that the global surface temperature response to abrupt forcing is well approximated by two (or sometimes three) e-folding timescales, with usually a single timescale representing the “slow” warming beyond a couple decades at least out to 150 years or so. The authors pose an interesting question of whether this slow adjustment is well approximated by a single timescale when considering longer simulations out to 1000 years, and whether that timescale changes with forcing level given nonlinearities in climate feedbacks (lambda) and time dependence of the climate’s effective heat capacity (c). It’s obvious that a single timescale shouldn’t work for the slow adjustment if you assume it’s given by something like tau = -C/lambda and allow c and lambda to change over time or with temperature. The authors demonstrate this using AOGCMs forced by 2x, 4x, and 8x CO2. In particular, they suggest that climate feedback nonlinearities (lambda changing with temperature) changes the slow mode adjustment timescale, and that changes in effective heat capacity also play a role, all of which makes sense.

Beyond this, I had difficulty reviewing what the authors have done, even after reading the manuscript several times. The writing is opaque and the many methods are not explained clearly. I generally trust that the authors know what they’re doing so I expect that the results will hold up once explained more clearly, but I currently am unable to make a recommendation regarding whether the study should be published. As written, I don’t think even interested readers (such as myself) will get much from it. I hope the authors will find these comments useful as they revise.

Unclear writing
- A more general introduction would be helpful. You really throw the readers into the deep end on L20-24 with two redundant sentences that don't really say what a “slow mode” is (long term climate change is determined by the slow mode, and the slow mode describes the temporal adjustment on long timescales). A better introduction would be something along the lines of what you have written on L51-55. And then a clear statement could be made about the aim of the study, e.g., the two-mode approximation has been shown to work well for single forcing level and out to 150 years, but it's unclear whether this still holds for longer timescales and multiple forcing levels given feedback nonlinearities.

- L55: By “associated with a radiative feedback” do you mean “different regions associated with different radiative feedbacks”?

- L58-74: I was getting lost at times whether you were talking about a timescale tau or the effective heat capacity C, and what the relationship is between them. I suggest rewriting to give a simple example first: C dT/dt = F + lambda*T, in which case tau = -C/lambda. Then you can point out that tau would not be constant if C is time dependent, which of course it is in climate models, or if lambda changes (e.g., the pattern effect or nonlinearities such that the feedback becomes lambda + a*T).

- L107-109: I don’t understand this sentence.

- L90-170 (Conceptual Insights): I found this section to be very confusing. You introduce a two-region model (equations 4 and 5) which later you use to analyze the slow response in GCMs. You could use this model to straightforwardly make your point that the response timescale (in this case tau_S = C_S/(lambda_S+a_S*T_S)) is not constant if C_S changes or for nonzero a_S. But instead, you introduce the two-layer model which is mathematically equivalent to equations 4 and 5 for the case a_F=a_S=0 (as shown in Geoffroy et al. 2013) but appears to have a very different form. You then make the confusing statement that “However, the parameters of the two-layer model modify the inertia of the slow mode. For instance, parameter for the efficiency of ocean heat uptake eta is an inertia parameter, and changes in ocean heat uptake cause C_S to increase or decrease”. It's unclear whether you mean that C_S can change with model parameters such as eta (which is obvious because C_S can be written as a function of eta), or whether you mean that C_S changes over time somehow (which is the topic of the paper, but not obvious from these equations). L156-167 confuse things further by introducing a new approximate definition of the slow mode which asserts a constant tau_S while noting that the full solution to equation (5) has a time varying C_s (presumably for the case of nonzero a?). But then Figure 1c suggests that tau_S is a constant function of a, which is confusing given that equation (5) suggests tau_S should change over time for nonzero a_S.

Overall, much more clarity is needed here about whether you are talking about C_S changing with model parameters or about changing over time. It’s also not clear that discussing the 2-layer model adds anything at all given that it seems to confuse things and you don’t use it outside of this section anyway.
- L205-206: I don’t understand this sentence. What does it add?

- L237-239: I suggest you cite papers showing this.

- L245: I don’t follow this sentence. What do you mean by “the slow mode emerges from heat uptake”?

- L283-284: I don’t understand this sentence. Doesn’t a_S represent feedback temperature dependence on long timescales? Then how can it be less strong than that?

- L302-305: Getting lost here. I think you are saying that your theoretical predictions don’t work because the effective heat capacity is not constant, which makes sense. But can you show this, rather than simply suggesting it? Is there a way to account for changes in heat capacity separately from feedback nonlinearity in the response timescale?

- L345-346: How are these “mathematical terms”? This is an unclear explanation that I don’t follow.

- Discussion: I was hoping by this point to have some clarity about whether additional timescales are needed to model the slow response (beyond year 21). The results seem to suggest that a single timescale is not good enough, which makes sense. But is this true only because of feedback nonlinearities, or also because of effective heat capacity changes? Could theory be saved by adding one (or more) additional slow timescales? I do not know what the takeaways are here.

- L434-436: This is an important summary point which greatly helps to clarify the purpose of your study. But I could not tell you where these points were shown in the paper. Which figures show this clearly?

- L445-446: Wasn’t this the point of this study?

Unclear methods
- L181-189: A schematic would really help here. I think you are describing Gregory plots for the fast and slow mode, but I was not able to follow your methods without sketching out what the Gregory plots would look like for fast and slow modes separately and their combination via the equations on L182.

- L183-184: All of your results follow from this choice to separate fast and slow modes at year 21. How did you choose this separation year (other than following earlier papers, e.g., DOI: 10.1175/JCLI-D-14-00545.1)? What did you find when you “explored the separation of the fast and slow mode”? Do your results depend on this choice at all? And, how do you actually define T_S and N_S for the slow mode? Do you take values of T and N at year 21 and subtract them from all subsequent years of the T and N timeseries to calculate T_S and T_N, or something else?

- L190-195: More detail is needed here, for example:

  - How do you define and calculate the “background feedback parameter lambda”? With what runs? Over what timescales? Is it assumed to not change between forcing levels, and is this a good assumption? I imagine that if different forcings produced different patterns of warming, that would result in different feedbacks from pattern effects rather than global temperature nonlinearities, but has this been accounted for somehow?

  - How is T(infinity) estimated from transient simulations?

  - I got lost trying to figure out what the three equations were for each forcing level, and how you went about fitting for all the parameters. I suggest writing this out explicitly for readers to follow.

  - Do you assume that a_F and a_S are the same, or can they be different?

  - How do you calculate effective radiative forcing in all the runs? (you mention this later, but it should be stated clearly here.)

- L208-211: I don’t follow the method you describe. I understand you do regression over different lengths of years (from 5 to 20), but what does it mean to “apply subsequently bootstrapping by replacement of the forcing estimates in order to generate the details of a continuous probability distribution”?

- L214-215: There are of course a way to more precisely estimate forcing using fixed-SST
simulations with CO2 increased.

- L222-229: Again, a schematic would help show what you mean here. It’s not obvious how all this works without showing the reader.

- L233-236: Is it a linear extrapolation a good assumption given the feedback nonlinearities you find? How off might your estimates be? It’s also unclear what you did with bootstrapping here again.

- Figure 4: I did not get much out of Figure 4. What are we supposed to learn here? Are we to take away that the energy balance model predictions match the AOGCM responses or not?

- L278: What does it meant to solve for an equation? Do you mean a specific set of the parameters in the equation? By what method?

- L341-343: It’s plausible that it’s differences in feedback temperature nonlinearity causing the differences in the temporal changes in tau between models, as you suggest. But can you show this? Could changes in effective heat capacity not also play a role?

- Figure 6: Again, it’s unclear what to take away from this figure. That the energy balance model doesn’t replicate the AOGCM output? What is learned?

- Section 5.2: I have admittedly run out of steam here, but I don’t know what the point of this section is or follow its methods.