General comments

Khabbazan and Held's paper highlights the nuances which must be considered when using a one box energy balance model for climate projections (the form they focus on is the one presented by Petschner-Held (1999), herein PH99, but any similar one-box model would exhibit the same behaviour). Their major conclusion is captured in the last paragraph of the paper, specifically that callibrating PH99 is 'much more involved than previously assumed' and hence 'future users should carefully consider whether they actually want to use PH99, or whether they prefer a less parsimonious solution'. On top of this, they also present a lovely bit of analysis which shows why a one box model must use a lower ECS than a two-box model if the two are going to respond similarly to a strong mitigation radiative forcing scenario over an ~200 year timescale.

Having already reviewed the paper in its first iteration, my major concerns now focus on its presentation and communication. In particular, having read the comments by reviewer 3, I am also not convinced that the current form of the paper will allow it to have its maximum impact. Suggestions to address this are below.

Despite this, I now feel that the paper presents some very interesting and pertinent results and so, subject to major revisions to fix a couple of errors, as well as to make the structure and messaging suitably clear for readers outside the field, would recommend it for publication.

Major concerns

1. The paper is still extremely difficult to read (I acknowledge the irony of this comment given that my review is probably also difficult to read). It has some extremely good, and pertinent, points to make but the style means they are far less accessible than they should be. I think a full rewrite is required if this paper is to have the impact that it should. Given that all of the science is done, that rewrite should not take too much time. However, before the paper is published, it should be proofread by at least a few other people in the group as the number of errors/incomprehensible passages which remain in this revision suggest that this step has not yet been suitably taken.

2. The authors include statements such as 'Over all, the results show that PH99 would be well trained by being calibrated to any RCP scenario.' alongside '[PH99]'s ECS and TCR are re-interpreted as effective, scenario-class-specific values'. These two comments are contradictory; a model cannot simultaneously both be appropriate for use across a wide range of scenarios and require recallibration for each different scenario class. The message would be made far clearer if the authors were to chose some metric of 'emulation accuracy' and use that throughout to explain how

well (or not) PH99 is emulating the target timeseries rather than using vague statements such as 'excellently emulate', 'appropriately mimic' and 'suitably mimic'. Alternately, the authors could simply present their results and make no comment on the suitability of PH99 as an emulation tool, leaving the reader to make up their own mind about whether it's 'suitably accurate' or not.

3. At the end of the paper, the use cases for PH99 are still not clear. How much faster are one box models like it than two box models? Are there any cases where the increased speed of the one box model over a two box model justifies its use, given how much harder it is to callibrate and its degrading performance outside its callibration range? Put another way, what sort of tradeoffs are made in terms of speed and performance? For example, what is the difference between using e.g. MAGICC, which runs in about a second but has excellent emulation ability, and PH99, which runs in some shorter amount of time (I'd guess a thousandth of a second or even much less) and has rapidly degrading emulation ability outside its callibration scenario type and time horizon.

4. The discussion of 'constant effective oceanic heat capacity, h' is wrong. See discussion below.

Possible ways to help the paper structure

Generally, I feel that the paper's argument gets broken at awkward times and that this makes its contributions unclear. I also think that its derivations are unclear and would be greatly helped by detailed supporting appendices/supplementary material.

The literature review in the introduction and the discussion of the Lorenz curve in Section 4 are introduced in a way which is particularly jarring to the overall flow of the paper. They interrupt the main flow, which makes it harder to tell which contributions the paper is making and what is existing literature. Shifting such sections around so that they're standalone would make it much easier to see what the contributions of this paper are, and then later compare to existing work.

The derivations are at times impenetrable for all but the most dedicated readers. Adding appendixes which walk the readers through the analytic derviations much more slowly will ensure that readers from outside the field have a much better chance of understanding what is going on and what the papers' equations mean.

Keeping the comment above in mind, Section 5 needs to be re-written. Section 5.1 makes an extremely important point but it is currently very hard work to get there. Section 5.2 may also make an important point, but even after multiple reads I am still not sure what it is (that you can do the same conversion between

2-box and 1-box models as between AOGCMs and a 1-box model, but that it's not as good?). Some better balance between a plain English summary of what the equations mean, more explanation of the steps taken in the derivation (perhaps in an appendix) and the mathematical derivation itself needs to be made. Perhaps Section 5 could stick to a plain English summary, leaving the mathematical details to an appendix/supplementary material?

Use of effective oceanic heat capacity

Equation (7) is

$$\mu = \frac{Q_2}{h \ln 2}$$
$$\rightarrow h = \frac{Q_2}{\mu \ln 2}$$

given that $\mu \ln 2$ is the radiative forcing due to a doubling of CO_2 concentrations for PH99 (herein Q_{2PH99}) then it is clear that h is not an 'effective oceanic heat capacity' as stated in the text but rather a dimensionless scaling factor which is the ratio of the forcing due to a doubling of CO_2 in the AOGCM to the forcing due to a doubling of CO_2 in PH99 (the units of h make this clear). Hence the correction which is given to allow the authors to use Forster et al's radiative forcing timeseries with PH99 is then

$$Q_{PH99}(t) = \frac{Q(t)_{Forster}}{h} = Q(t)_{Forster} \frac{Q_{2\,PH99}}{Q_{2\,AOGCM}}$$

where here we have used Q to denote forcings, rather than a combination of F and Q as in the paper, for consistency.

Hence what the authors are actually doing is simply rescaling the entire forcing timeseries by the ratio of the AOGCM's and PH99's forcing due to a doubling of CO_2. For non-CO_2 forcing, which is important in strong mitigation scenarios, this seems completely inappropriate as it assumes some connection between CO_2 and non-CO_2 forcing which may or may not be there.

The question I ask here is why the authors are using equation (1) as a starting point. Why not simply start with equation (5), or some variant thereof, which takes total forcing as input rather than CO_2 only forcing e.g.

$$\frac{d\Delta T}{dt} = \frac{Q(t) - \lambda \Delta T}{d}$$

where d is some time constant of the response.

This would remove the need for the factor of h and make the entire analysis much clearer and easier? I'm almost certain the conclusions would be unchanged (as this is still a one box model with two free parameters) and you wouldn't need this awkward scaling factor throughout.

Specific comments

1. 'the most parsimonious SCM, PH99, ensures maximum transparency', this is only true if all the parameter settings etc. used to run the model are documented, a practice which isn't yet commonplace. I think solving the transparency problem is related to more than just model choice, namely accessibility of code, documentation of reproduction steps. A more accurate statement may be that, 'the most parsimonious SCM, PH99 (and variants thereof), ensures maximum comprehensibility'.

2. The discussion around the impact of directly prescribing AOGCM ECS values to PH99 rather than callibrating PH99 is, in my opinion, confusing. ECS has a strict definition and if you ran PH99 until it reached equilibrium, its ECS would be the same as the AOGCMs'. Hence researchers who have directly transferred AOGCM ECS's to PH99 have not sampled 'effectively higher ECS', they have sampled the ECS they prescribed. What they haven't done is sample the intended centennial climate response. They have sampled a response which is higher than that of the corresponding AOGCM.

As a result, I think the implications of the paper would be much clearer if the comments about sampling higher ECS values than intended were all rephrased to focus on sampling higher climate responses to radiative forcing than intended. This would reinforce the important conclusion of this paper and also highlight that using ECS to characterise climate response on a few hundred year timescale is fundamentally flawed, given that ECS takes on the order of a thousand years to emerge.

I think this would also facilitate a simpler comparison with Van Vuuren (2011)'s findings. The models Van Vuuren considered were sampling lower temperature responses. You have shown that the opposite problem emerges when researches directly transfer AOGCMs' ECS values to PH99, i.e. researchers were inadvertantly sampling higher temperature responses than they intended to. Section 5 shows that this is not because they are sampling a higher ECS (they are, by construction, sampling the same ECS and would see the same long-term response) but because PH99 has a fundamentally different response shape to an AOGCM and hence ECS alone does not allow you to easily move between the two (reflecting the paper's main conclusion). You show that the solution to this is to adjust PH99's ECS, sacrificing agreement in the long-term response in

order to gain agreement in the centennial response (which is sensible given it is more often than not the timescale of interest).

Technical corrections

page 1, line 9: 'MIND and PAGE' → 'MIND and PAGE, widely used in policy making'

page 1, line 9: delete 'recent' (in fact I think you could delete the entire sentence, not needed in abstract)

page 1, line 12: 'although the model was validated in the past', I have no idea what this is referring to

page 1, line 13: 'overestimate mitigation needs and costs' → 'sample a higher range of climate responses than they intend to'? Whilst it's clear that this increases estimated costs, how much is not that clear and so talking about overestimates seems potentially premature. I'm on the fence about this one so this is more of a thought than a recommendation.

page 1, line 14: 'produced by' → 'resulting from'

page 1, line 15: 'good emulator' → 'good emulator (accurate to within 0.1K)' (it might be more accurate than this, my point here is that having a quantification of 'good' in the abstract would be useful)

page 1, line 19: delete 'on the question'

page 1, line 22: 'larger ECS than claimed' → 'larger climate response to forcing than claimed' (by definition the ECS is unchanged, your results show that it's just that the centenial timescale response is too high for a one box model with the same ECS as a two-box model)

page 2, line 2: 'would' → 'may'

page 2, line 4: delete 'Currently'

page 2, line 3: delete the first 'thousand' and 'were' → 'is'

page 3, line 4: 'Third' → 'Thirdly'

page 3, line 6: C missing after 3 degrees

page 4, line 5: delete 'might be in order', either you believe the note should be there, or you don't (in which case you can delete the entire paragraph)

page 4, line 23: somethign has gone wrong with the brackets at the end of the line (one too many or too few, I can't tell)

page 4, line 26: delete '(The right-hand side of the equation has been obtained by utilizing Eq. (2).)'

page 5, line 6: delete '(see)' (or fix whatever was meant to be there)

page 5, line 8: delete extra comma before the first full stop

page 5, line 12: 'multiply it', not clear what 'it' is

page 5, line 14: changing from T dot to dT/dt halfway through the paper is extremely confusing, please choose a convention and stick with it. Similarly for F and Q throughout.

page 5, line 15: I don't understand what this sentence means

page 5, line 22: 'any' → 'all' and delete 'above' (or add something above, there's nothing there at the moment)

page 5, line 32: 'beyond CO2' → 'which includes non-CO_2 forcing'

page 6, line 3: why is 1881-1910 used as pre-industrial? Seems an odd choice for pre-industrial for a perturbation model as it's clearly not a pre-industrial period (although this may not matter in the end)

page 6, line 9: lack of quantification of 'tolerable' (see discussion above) makes it very hard to know what you mean here

page 6, line 16: if you need to delete something, this paragraph on drift could go. Whilst it is nicely done, it is also a fairly obvious point that PH99 itself acts as a low-pass filter

page 6, line 28: 'APGCM' → 'AOGCM' (find-replace the whole document to check for others which might have slipped through the cracks)

page 7, line 1: start of sentence is missing

page 7, line 11: 'consists in' → 'is'

page 7, line 14: (subject to discussion of h above), you may as well tune h too

page 7, line 20: delete 'In this regard,'

page 7, line 24: lack of quantification of 'suitably mimic' (see discussion above)

page 7, line 31: looking at Lorenz et al. here interrupts the flow of your paper. I would present your results first, then present the results of Lorenz et al. afterwards as a comparison.

page 8, line 18: The sentence starting with 'therefore' doesn't make sense. Do you mean that you are going to present analytic expressions which calculate α and μ from the inferred ECS and TCR?

page 8, line 30: turn 'square' into ^{2}

page 8, line 33: turn 'square' into ^{2}

page 9, line 8: swap this sentence with the next one so that you keep the advantages distinct

page 9, line 11: delete 'the ETE not only yields a better approximation;'

page 9, line 15: ‘cubic fit’ → ‘ETE’?

page 9, line 16: ‘For the sake of brevity. . . ’, maybe better phrased as ‘Given the explorations already done and their performance, we leave explorations beyond the linear approximation for future research’.

page 9, line 17: ‘beyond’ → ‘beyond a’

page 10, line 4: ‘t1’ → ‘t_1’ (twice), similarly for k2 and k1 throughout the text in this section

page 10, line 17: I have no idea how you got to this line, more explanation needed (if only in an appendix)

page 10, line 20: ‘being inferred from Eqs. (4) and (7)’, do you mean ‘by solving Eqs. (4) and (7)’ for the forcing given above?

page 11, line 2: ‘recapitulate’ → ‘note, by definition,’

page 11, line 5: put manipulation in an appendix

page 11, line 18: a1 formatted incorrectly

page 11, line 22: ‘1/2 degree’ → ‘0.5degreeC’

page 12, line 12: ‘hereby’ → ‘where’

page 13, line 7: delete ‘If the reader will join us in exploring this line of reasoning,’

page 13, line 18: ‘an at’ → ‘at’ and ‘least 2-box’ → ‘least a 2-box’

page 13, line 20: ‘excellent emulation’ without quantification is confusing (see discussion above)

page 13, line 23: discussion in terms of ECS only is confusing (see comments above)

page 14, line 11: ‘excellently emulate’ without quantification is confusing (see discussion above)

page 14, line 31: ‘parsimonious’ misspelt

page 15, line 4: ‘recapitulate’ → ‘rearrange’

page 15, line 17: ‘whereby’ → ‘where’

page 15, line 21: ‘appropriately mimic’ without quantification is confusing (see discussion above)

page 15, line 22: ‘Columns’ → ‘column’ and ‘5th’ → ‘5’

page 15, line 23: ‘Over all, the results show that PH99 would be well trained by being calibrated to any RCP scenario.’ is a direct contradiction to the paper’s statements such as ‘its ECS and TCR are re-interpreted as effective, scenario-class-specific values’. I would delete this sentence entirely as I think you’ve shown it’s not true.