General comments

Repeating my comments from the second round review, 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.

My major concerns now focus on making the discussions of emulator quality quantitative, in particular removing vague terms, and the representation of one key reference.

The paper presents some very interesting and pertinent results and so, subject to revisions to fix the concerns above, would recommend it for publication.

Major concerns

1. I am happy to be corrected on this, but I don't think the representation of Calel & Stainforth (2017) in line 27 of page 2 is fair. The authors say that 'Calel & Stainforth (2017) highlighted the potential future role of PH99, however if and only is users invested in an application-specific re-callibration of PH99 as a valid future approach to emulation.' I cannot work out where this comes from. As far as I can tell, Calel & Stainforth highlight differences in the uses of PH99 in different IAMs and reveal how to resolve them but I can't see any comment about the need for 'application-specific re-callibration'. Perhaps something like this would be a solution, 'Calel & Stainforth highlighted the potential future role of PH99 and hence further validation of its behaviour is warranted.'

2. The discussion of emulator quality is somewhat varied. In places, the authors have been very careful and ensure that they quantified the meaning of terms such as 'a good emulator' or 'a sufficiently accurate' emulator but in other places these terms remain vague. Making sure that the meaning of all such 'quality' terms was clear would help make the point being made clearer to the reader, I feel.

Specific comments

1. The difference between your result and the Van Vuuren et al. (2011) result is still not that clear to me (i.e. why do you find the climate components responding too strongly whilst they found them responding too weakly). Is it simply because Van Vuuren et al. (2011) considered emissions driven results whilst you are considering forcing driven results and the emissions to radiative forcing steps are outweighing the forcing to temperature step? Or something else? Making some comment on possible reasons for the difference would help place this article in the context of other work.
2. The introduction is quite long, is it possible to split it or cut it somehow (some sections might be better included in the discussion rather than in the introduction)?
3. A quick check over your treatment of acronyms would be a nice improvement. Sometimes you introduce the term first, and the acronym next e.g. 'integrated assessment models (IAMs)' whilst other times you introduce the acronym first, and the term next, e.g. 'RCPs (representative concentration pathways)'. This is somewhat confusing and it would be nice to make it uniform.
4. Do you have any intention to make your analysis code available?

Technical corrections

page 1, line 14: 'emulator (accurate to within 0.1K for mitigation scenarios and the baseline scenarios RCP4.5 and RCP6.0) of these AOGCMs' → 'emulator of these AOGCMs (accurate to within 0.1K for RCP2.6, RCP4.5 and RCP6.0)'

page 1, line 15: 'time horizon' → 'time horizon (on the order of the time to peak radiative forcing)' or something which quantifies which time horizon you're talking about

page 1, line 16: 'We offer a method to re-interpret already published works based on the 1-box model accordingly.' → 'Accordingly, we offer a method to re-interpret already published works based on the 1-box model.'

page 1, line 18: 'claimed' → 'intented' (if that's what you actually mean)

page 1, line 22: 'would comply' → 'comply'

page 1, line 23: 'the most sophisticated' → 'sophisticated' (Earth System Models are arguably more sophisticated)

page 1, line 25: 'foster' → 'offer'

page 2, line 3: 'Van Vuuren' → 'In previous work, Van Vuuren'

page 2, line 26-27: adjust comment about Calel & Stainforth

page 2, line 29: ‘correctly’ is vague, although given it’s in an overarching question perhaps ok (as long as you define the term later)

page 2, line 32: ‘2target’ → ‘well below 2target’ (although not being a lawyer, I might be wrong on this one)

page 3, line 5: quantify ‘for years’ or use another term

page 3, line 7: ‘- inadequate’ → ‘- as inadequate’ (although the massive break with the dashes makes it hard to see exactly how the sentence is meant to fit together, perhaps re-write the whole sentence)

page 3, line 8: ‘sufficient’ → ‘0.1K’ (is this what you actually mean by ‘sufficient’?)

page 3, line 13: ‘but likely not beyond it’, can you check?

page 3, line 21: add comma after ‘For that reason’

page 3, line 28: ‘the former’ → ‘(ii)’ (or do you mean (i) and (ii))

page 4, line 1: quantify the size of the error

page 4, line 6: ‘model market.’ → ‘model market’.

page 4, line 14-16: ‘Among others, one of the most extensively used most parsimonious climate emulators is the 1-box global energy balance 15 model, Eq. (1), introduced by Petschel-Held et al. (1999), which projects the atmospheric GMT anomaly compared to its preindustrial level.’ → ‘PH99 projects the atmospheric GMT anomaly compared to its preindustrial level.’

page 4, line 28: ‘propose the 3-step’ → ‘propose a 3-step’

page 5, line 1: ‘good’ is vague, quantify or re-word

page 5, line 5: ‘such scenarios are not available for’ → ‘AOGCMs have not been run for 2-target-compatible scenarios for’

page 6, line 7-11: just to check, the reason you don’t use the historical period for validation is that you want to focus on purely projection emulation, not model validity in a range of forcing scenarios? This seems slightly odd to me, especially given you can always chose to compare temperature perturbations between convenient reference periods in later quantifications.

page 6, line 29-31: great bit of sensitivity analysis

page 7, line 6: ‘the RCP2.6’ → ‘RCP2.6’

page 7, line 10-13: comments about Paris Agreement and relevance of 0.5K difference could come out much earlier than Results. Perhaps introducing this earlier would help you set out what acceptable ‘error thresholds are’ and make the scales you’re talking about throughout the paper clearer

page 7, line 23: ‘whereby’ → ‘where’

page 7, line 24: ‘so’ → ‘sufficiently’, delete ‘suitably’ (as you quantify later anyway)

page 7, line 25: ‘0.14’ → ‘0.14K’

page 8, line 1-2: ‘Before diving into our suggestions, it might be worthwhile to first take a look at one of the existing options. (However, a reader mainly interested in our improved method of utilizing PH99 might directly move on to Subsection 4.2.)’ → ‘Before diving into our suggestions, we examine one of the existing options (a reader solely interested in our improved method of utilizing PH99 can move straight onto Subsection 4.2.)’

page 8, line 4: ‘the ECS’ → ‘ECS’

page 8, line 11-17: Can you make some comment on how the TCR is calculated using the Lorenz approach, do you just leave it constant?

page 8, line 19: heading has typo, ‘AOGMC’ → ‘AOGCM’

page 9, line 11: ‘better’ → ‘best’

page 9, line 17: ‘Please note’ → ‘Note’

page 9, line 26-27: Delete ‘Hereby we presuppose that a 2-box model emulates an AOGCM qualitatively better than a 1-box model.’ I don’t think it adds anything here and you have good discussions elsewhere which explain why you are using a 2-box model at all.

page 10, line 6: ‘perspective on’ → ‘projections under’

page 10, line 11: ‘find a’ → ‘approximate a’

page 10, line 14: ‘both summing up to’ → ‘sum equal to’

page 11, line 14: ‘resulting in’ → ‘where’

page 12, line 2: what does ‘exact’ refer to here, do you mean ‘2-box’?

page 12, line 18-19: ‘In Section 5.1 we derived an analytic explanation why a naïve transfer of an AOGCM’s ECS and TCR to PH99 leads to a too large maximum GMT when driven by a mitigation forcing scenario.’ → ‘In Section 5.1 we derived an analytic explanation for why a naïve transfer of an AOGCM’s ECS and TCR to PH99 results in a maximum GMT which is too large when driven by a mitigation forcing scenario.’

page 12, line 19: delete ‘could’

page 12, line 20: ‘good’ is vague, quantify/reference somewhere else/make clearer

page 12, line 25: ‘a order-of-magnitudes’ → ‘an order-of-magnitude’ or → ‘an orders-of-magnitude’

page 13, line 1: ‘Quite’ → ‘On’

page 13, line 9: this is fine for exploration but a fairly brute force way of making the models agree as far as I can tell

page 13, line 13: 'an as' → 'a'

page 13, line 14: 'We cannot' → 'Hence we cannot'

page 13, line 23: 't_1' → 't_1, i.e. on the order of the time to peak forcing' (remind the reader what t_1 means)

page 14, line 1: delete 'would like to focus on our main finding and'. If you want to make this the main finding, I think you need to re-structure the article as at the moment the main finding is definitely the improved method to transfer AOGCM ECS/TCR onto PH99 ECS/TCR, with this explanation of why such a new method is necessary being a nice bit of supporting analysis.

page 14, line 8: delete 'i.e. for the upcoming 200 years vs. the time horizon thereafter', it adds nothing

page 14, line 12-15: nice quantification

page 14, line 28: 'utilizig' → 'utilizing' (probably worth checking whether you are using American or English spelling throughout, if English then 'utilizing' → 'utilising')

page 14, line 34: delete 'even'

page 15, line 2: missing bracket after '2014'

page 15, line 18: delete 'a first version of'

page 15, line 20-21: '1-box-based model. (Hereby we assume that a 2-box model mimics an AOGCM better than a 1-box model.)' → '1-box-based model (assuming that a 2-box model mimics an AOGCM better than a 1-box model).'

page 15, line 24: 'sensible' → 'useful' (the irony of me correcting this is not lost given I wrote sensible in my previous review, I just think that in the paper useful, or even applicable, fits better)

page 15, line 28: 'the explanation of which' → 'for which the explanation'

page 15, line 30: 'equivalennt' → 'equivalent'

page 15, line 1: delete 'rather would like to'

page 15, line 4-7: perhaps switch full stops for semi-colons in your numbered phrases i..e 'we propose the following steps: (i) By comparison with more sophisticated, multi-box climate modules it should be tested again whether the effect of a transient climate sensitivity (and TCR) alone could explain our observed PH99- AOGCM discrepancy. (ii) Future discussions with the AOGCM community should illuminate to what extent the further explanations we suggested might also apply, thereby potentially reducing the need to correct for PH99. (iii) An' → 'we propose the following steps: (i) By comparison with more sophisticated, multi-box climate modules it should be tested again whether the

effect of a transient climate sensitivity (and TCR) alone could explain our observed PH99- AOGCM discrepancy; (ii) Future discussions with the AOGCM community should illuminate to what extent the further explanations we suggested might also apply, thereby potentially reducing the need to correct for PH99; (iii) An'

page 17, line 13: 'insure' → 'ensure'

page 25, line 1: I think you could combine Figures 3 and 4

page 30, line 1: I think you could simply reference this figure rather than including it in full

page 35, line 1: I think you could simply reference this figure rather than including it in full

page 36, line 1: could you combine this with Figures 3 and 4