Interactive comment on "Evaluating Climate Emulation: Unit Testing of Simple Climate Models" by Adria K. Schwarber et al.

dr. Rypdal (Referee)
kristoffer.rypdal@uit.no
Received and published: 25 October 2018

Dear Dr. Rypdal,

We want to begin by thanking you for taking the time to read our manuscript and provide comments. We have copied the unedited original comments in bold. Our point-by-point responses are provided in regular font, indented from the original comment for clarity. We will supplement our response with revised text after we have responded to all reviewers and following the ESD process.

General comments

This manuscript presents the responses of a set of climate variables in five different simple climate models (SCMs) to a selected set of impulses. The results of the global temperature response to one of these impulses (a step quadrupling of atmospheric CO2-concentration) is compared to the corresponding responses in an ensemble of CMIP5 Earth System Models (ESMs). The simple models belong to two categories: the idealized SCMs (AR5-IR and FAIR), and the comprehensive SCMs (Hector v2.0, MAGICC 5.3, and MAGICC 6.0).

We appreciate that you took the time to provide an accurate summary of our work.

Testing of simple models against more complex ones is interesting and relevant to ESD, but the interpretation of results are difficult, since it is not obvious that a complex model represents specific aspects of reality more correctly than a simple model.

We appreciate that you agree this work is interesting and relevant to ESD. Comparing simplified models to more complex models is a technique often utilized in the literature (e.g., Joos et al., 2013) and we also employ this technique. We compare the responses of idealized SCMs to comprehensive SCMs and comprehensive SCMs to CMIP5-class models. In our paper, we do not expect individual models to represent reality, but instead rely on the multi-model mean to ground our comparisons. It is well established that the multi-model mean behavior of the complex models replicates well a broad suite of observations (e.g., Figure 9.7, Flato et al. 2013). Our subsequent responses address this comment.

The paper does not seem to present novel concepts, ideas, tools or data. The concept of "unit testing" seems to be a misnomer here, as pointed out in the comment by dr. Nicholls.

We strongly believe this paper does present concepts that are new to the literature. Though fundamental impulse tests have been used in the literature, our manuscript employs these existing techniques in a novel way. This is the first study in the literature to rigorously evaluate SCMs using impulse-response tests. SCMs are widely used in the literature and in decision-making context, e.g., within Intergovernmental Panel on Climate Change (IPCC) Reports, coupled with Integrated Assessment Models. In fact, a paper describing a commonly used SCM, MAGICC 6.0, has been cited 371 times in the literature and policy contexts. Another model, the impulse response model used in the IPCC Fifth Assessment Report (AR5-IR), is heavily used by the scientific community to support decision making. Despite their importance, the fundamental responses of SCMs are not fully characterized and we provide a set of tests that we recommend as a standard evaluation suite for any SCM. Further, the U.S. National Academies of

Science (2016) specifically suggested that SCMs be, "assessed on the basis of [the] response to a pulse of emissions," which we do here.

We have added portions of the text above to the revised manuscript introduction to make a more compelling case for our work.

We address the comment about the phrase "unit testing" below.

The conclusions are not very clear, and the concluding section is very short.

We will expand the conclusion in the revised manuscript to include a discussion of Table 1, and we copied the revised text into this response below.

The authors do not present reflections around the assumptions underlying the conclusions.

We remind the reviewer that we are evaluating the behavior of models and their responses to fundamental impulse-response tests and are not providing information on the underlying mechanisms of the models. The underlying mechanisms are explored by the individual modelling groups in their publications, which we have cited in our manuscript.

Model parameters are not given and discussed (not even in the supplement), which has been a source of frustration and confusion for this referee.

We apologize for any confusion in our omission of model parameters. We agree that model parameters are very important for understanding how these models differ. We will add the model parameter files to the supplemental materials so that readers can more easily replicate our results.

Reasonable credit is given to related work.

Thank you for the positive comment.

The title should find another term than "unit testing".

We use the phrase "unit testing" with the understanding that this phrase is commonly used in software as we mentioned in the Supplement. Similar to meaning of "unit testing" in software, we are testing the SCM in the simplest way possible, by determining the impulse response of specific model sub-systems such as CO_2 and CH_4 gas cycles, and the forcing to temperature response of each model. Though we believe our use of the phrase is consistent with its use in software, as we replied to the Short Comment, we will update the language in the manuscript and title to "fundamental impulse tests" to avoid confusion.

The abstract reflects the content of the paper, apart from the term "unit testing".

Thank you for the comment. We addressed the use of the term "unit testing" in the response above and will instead use the phrase "fundamental impulse tests".

The presentation and language is adequate.

Thank you for providing comments on the structure of the paper.

<u>Specific comments</u>

FAIR is a generalization of AR5-IR to include state dependence of the carbon cycle (Millar et al., 2017). For the experiments shown in Figures 1 and 4 (temperature responses to CO2-forcing), the carbon-cycle module is not active, and from my understanding of the description of FAIR in Millar et al., 2015, the two models should be identical when temperature response to CO2 concentration is simulated. However, in both figures the responses of the two models are very different.

We do expect slight differences in the response of FAIR and AR5-IR to a unit forcing. According to Equation 8 in Millar et al., 2017, FAIR will have a differential response to change background CO_2 concentrations. By contrast, the AR5-IR response is independent of background concentration.

If the models are identical in this mode this can only arise from different choices of the time-constant parameters in the simulations of AR5-IR and FAIR. From the figures it looks like the time constants for temperature response in AR5-IR are those used originally by Myhre et al., 2015 (Table 8.SM.11, d1 = 8.5 yr and d2 = 409.5 yr), while in FAIR they look more like the choice of Millar et al. 2017 (d1 = 4.1 yr and d2 = 239.0 yr).

As we mentioned above, the FAIR and AR5-IR responses will differ. And we did use the time constant parameters representing the thermal equilibrium of the deep ocean (d2) and the thermal adjustment of the upper ocean (d1) from Myhre et al., 2013 rather than from Millar et al., 2017. We are testing the model responses as they would be 'out of the box' and only make modifications if required for the models to run, as was the case for Hector v1.1 to handle a $4xCO_2$ concentration step.

However, to address your comment we have included below additional model responses from the AR5-IR model using parameters from Millar et al., 2017. The parameter choices are available below in Table R1. We will add this information to the Supplement.

Table R1 Parameter values for the simple impulse-response model, AR5-IR

Parameter (Units)	Value – AR5-IR (from Myhre et al., 2013)	Value – AR5-IR-var (from Millar et al., 2017)	Guiding analogues
α (Wm^{-2})	5.35	5.395 (α = F2x/ln(2); F2x=3.74)	CO_2 RF scaling parameter
q_1 ($KW^{-1}m^2$)	0.631	0.41	Thermal adjustment of the upper ocean
q_2 ($KW^{-1}m^2$)	0.429	0.33	Thermal equilibrium of the deep ocean
d_1 (year)	8.4	4.1	Thermal adjustment timescale of the upper ocean
d_2 (year)	409.5	239.0	Thermal equilibrium timescale of the deep ocean

Figure R1 shows the temperature response from a CO_2 concentration impulse in several SCMs, including the AR5-IR response found using the Millar et al., 2017 time constants, which we refer to as "AR5-IR-Millar-parameters" in this figure. We note that the AR5-IR-parameters response is still not identical to FAIR because FAIR has a differential response to change background CO_2 concentrations.

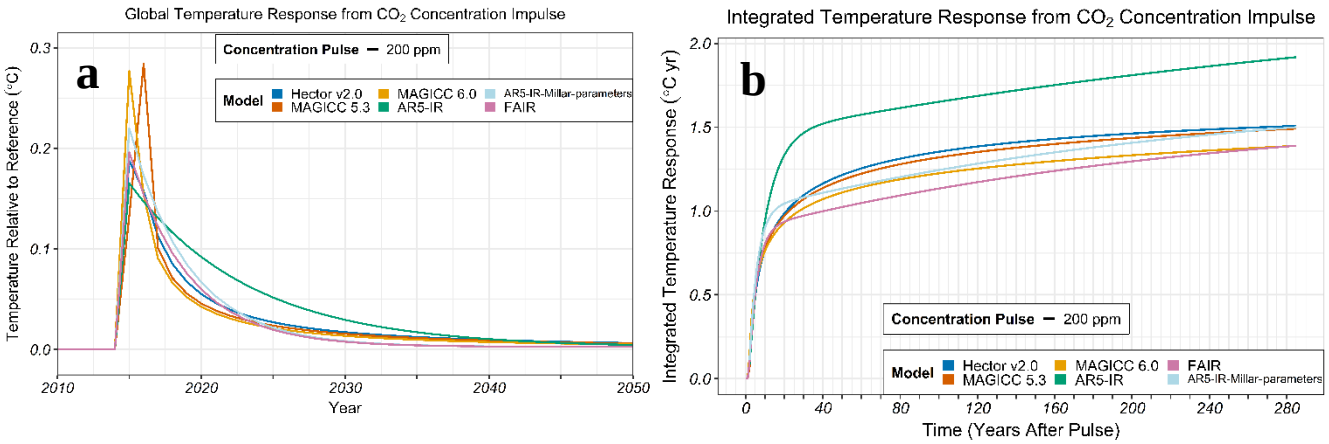

Figure R1 Global mean temperature response (a) and integrated global mean temperature response (b) from a CO_2 concentration perturbation in SCMs (MAGICC 6.0 – yellow, MAGICC 5.3 BC-OC – red, Hector v2.0 – blue, AR5-IR – green, FAIR – pink, AR5-IR-Millar-parameters –light blue). The time-integrated response, analogous to the Absolute Global Temperature Potential, is reported as 0-285 years after the perturbation.

Moreover, if I have got this right, then AR5-IR and FAIR are not only identical models in the simulations shown in Figures 1 and 4, they are also both linear (the nonlinearity in FAIR is in the carbon-cycle module).

The nonlinearity in FAIR is also present in the forcing module based on Millar et al., 2017 Equation 2.

For a linear response, the time-integrated temperature response shown in Figure 1b and the response to a step forcing shown in Figure 4 are identical, apart from a multiplicative constant depending on the relative strength of the forcings used in Figure 1 and 4. However, in Figure 1b the FAIR response curve is well below the AR5-curve, while in Figure 4 it is well above. For linear, identical models this is possible only if ratio between the climate sensitivities (ECS) of AR5-IR and FAIR is chosen larger in the simulations for Figure 1 than for Figure 4.

We used consistent ECS values throughout our experiments, unless otherwise noted, and we do want to thank you for your careful comments. We made an error in applying the 4xCO_2 concentration step in the AR5-IR model, which resulted in the response being significantly lower than it should have been. Figure R2 in our response provides the updated results and is consistent with Figure 1b. We have updated the manuscript and supplement to reflect the amended figure, and we note that this change does not impact our overall conclusions that, "Fundamental forcing tests, such as a 4xCO_2 concentration step, show that the SCMs used here have a faster warming rate in this strong forcing regime compared to more complex models. However, comprehensive SCM responses are similar to more complex models under smaller, more realistic perturbations (Joos et al., 2013)."

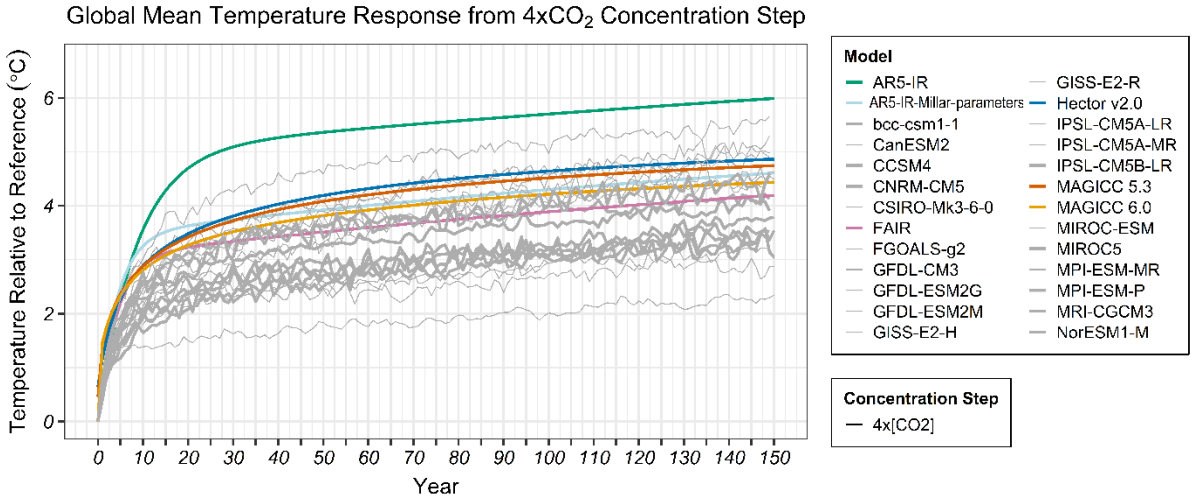

Figure R2 Global mean temperature response from 4xCO_2 concentration step in CMIP5 models (grey) and SCMs (MAGICC 6.0 – yellow, MAGICC 5.3 BC-OC – red, Hector v2.0 – blue, FAIR – pink, AR5-IR –green). A climate sensitivity value of 3°C was used in the SCMs and the thick lines represent CMIP5 models with an ECS between 2.5 - 3.5 °C.

In section 3.3 (line 209) the authors write: "Differences between the model responses to a finite pulse (Fig. 1) and a large concentration step (Fig. 4) demonstrates the expected bias in AR5-IR under larger perturbations." This sentence shows that the authors attribute the different relative response between the two models in Figure 1 and 4 to nonlinear effects in FAIR. While FAIR has a weaker response on decadal time scales than AR5-IR under the the small temperature perturbations in Figure 1, the response is stronger than AR5-IR under the stronger forcing in Figure 4, i.e., if model parameters are unchanged, this amplification must be due to a strong nonlinear feedback. The authors need to clarify the source of this nonlinearity in FAIR.

We apologize for the confusion, which we believe it is resolved by updating Figure 4 in the manuscript with Figure R2 in this response. The source of the nonlinearity in FAIR is in the forcing component.

The total forcing response to CO2 and CH4 emission impulses shown i Figure 2 show quite small spread over the SCMs. Unfortunately the FAIR response is not plotted in that figure, but the AR5- response does not differ drastically from the comprehensive SCMs. This indicates that the carbon-cycle module of the idealized and comprehensive models behave rather similarly. The substantial difference between AR5-IR and the rest appears when the resulting temperature response is displayed in Figure 3a, and also in the temperature response to BC emission in Figure 3b. This is all consistent with Figure 1; the time constant d1 for the temperature response in AR5-IR is too high. Fitting a two-box model to the multimodel mean in the 16 member ESM-ensemble considered by Geoffroy et al., 2013 yields d1 = 4.1 yr, which is about half the e-folding time observed for AR5-IR in Figure 1a and 3b. This supports the assertion that the mismatch between AR5-IR and the other SCMs is just a question of a bad choice of model parameters.

As we mentioned above, we tested these models using their default parameter values unless a change was required for the model to successfully complete an experiment. Though we take the reviewer's point about the importance of parameter choice, we note that the definitions and meanings of each parameter are not consistent across the SCMs used in this manuscript. For example, using the ocean component as an example we find that the vertical diffusivity parameter is not defined in the same way across the comprehensive SCMs, and is completely absent from the idealized SCMs where it is implicitly represented by the parametrized ocean timescale values.

Since no use of observation data is made in this paper, the benchmark to assess the performance of the SCMs are the complex ESMs. The temperature response to a step in BC emission is claimed (in S12) to level off much more slowly in SCMs than in the NorESM model, suggesting that the SCMs do not capture aerosol dynamics correctly, but otherwise the comparison with ESM responses is limited to the ensemble of 4 × CO2 step forcing simulations. Unfortunately, the spread over the ensemble of ESM responses in Figure 4 is so large that it cannot be used to validate the SCMs.

We first point out that our primary purpose in this paper is to evaluate the fundamental behavior of the simple climate models. We do this by both comparing them to each other, and also, in the limited cases where this is possible, to more complex models (Joos et al., 2013). We compare against the suite of complex model results because it has been shown that the multi-model mean behavior of the complex models replicates well a broad suite of observations (e.g., Figure 9.7, Flato et al. 2013). Also see the next response, below.

In Figure S22, responses for the three comprehensive ESMs are plotted for two other ECS values, 2.1 and 4.7 degrees. For ECS=2.1, the results are in the mid-range of the ESM-ensemble, while for ECS=4.7 the responses are outside (above) this range.

We changed the ECS values in the SCMs to illustrate the effects of parameter selection on the model responses. We found that spanning the range of complex model ECS values still resulted in stronger SCM responses, which supports the conclusion in our main paper that the SCMs have a faster warming rate under strong forcing regimes compared to more complex models. We revised the supplemental text around Figure S22 to state this as well.

Table 1 reflects the underlying circular logic in this approach to model testing, a logic that seems to be quite prolific in the modeling community. The performance of the models are ranked according to their deviation from the mean of the three comprehensive SCMs. Is the conclusion that the model closer to this mean is the preferable one?

We have moved amended text from the supplement to the main paper to better describe the logic behind our conclusions as represented in Table 1. We do, indeed, find that – at least amongst the simple models examined – the physically based comprehensive SCMs generally respond better than more simplified models such as AR5 or FAIR. As we clarify in the text, this is largely a relative assessment of the responses between the SCMs.

"By using fundamental impulse tests, we found that idealized SCMs using sums of exponentials often fail to capture the responses of more complex models. SCMs that include representations of non-linear processes, such as FAIR, show improved responses, though these models still do not perform as well as comprehensive SCMs with physically-based representations. Fundamental forcing tests, such as a $4xCO_2$ concentration step, show that the SCMs used here have a faster warming rate in this strong forcing regime compared to more complex models. However, comprehensive SCM responses are similar to more complex models under smaller, more realistic perturbations (Joos et al., 2013).

It is not possible to compare these fundamental responses with observations, and it is even more difficult to compare SCMs with the more complex models at decadal time horizons due to internal variability (e.g. Joos et al., 2013, Figure 2a). However, it is common in the climate modeling literature to use the multi-model mean as a base comparison. In fact, the CMIP5 multi-model mean has been shown to capture observational trends (among other climate variables) better than any individual complex model (Flato et al. 2013).

Thus, we use the comprehensive SCM multi-model mean to compare to the individual model responses. It is our conclusion that the model response closer to the multi-model mean is more accurately representing that particular response pattern. We illustrate this assumption by using the scale developed for Table 1, which generally uses the time-integrated temperature response percent difference from the comprehensive SCM average. We set the scale based on the range in percent differences found in our analysis: ••• : 0-10% difference, •• : 10-20% difference, and • : 20-30% difference from the comprehensive SCM average (S9).

For example, we assign the comprehensive SCM responses to a CO_2 concentration impulse a three (•••) because the responses are within 10% of the comprehensive SCM average. The idealized SCMs, FAIR v1.0 and AR5-IR, have greater differences and are given a two (••) and a one (•), respectively.

Under the $4xCO_2$ concentration step experiment, we can compare the SCM response to more complex models from CMIP5. We assign MAGICC 6.0 a three (•••) because it appears to respond more reasonably under stronger forcing conditions than the other SCMs. We assign Hector v2.0, MAGICC 5.3, and FAIR a two (••) because these SCMs have initially quicker responses to an abrupt 4xCO2 concentration increase compared to the ESMs. We assign AR5-IR a one (•) because it has a slower response to an abrupt 4xCO2 concentration increase and is insensitive to changing background concentrations.

For CH_4 emissions impulses, we use the difference from the comprehensive SCM average to rate the responses. Unlike the 100GtC CO_2 and $4xCO_2$ step experiments, we cannot compare the SCM responses to more complex models, therefore, we are more lenient in our performance assignment against the comprehensive SCM average. CH_4 is a well-mixed GHG and, therefore, we expect that the climate system response to CH_4 concentration perturbations will be similar to that for CO_2. However, it would be useful to evaluate in more complex models if the simple representation of chemistry in the comprehensive SCMs adequately represents the time evolution of CH_4 concentrations in response to a change in emissions.

Finally, we assign ratings to the SCM responses to aerosols. We do not explicitly conduct aerosol experiments other than BC because the responses of the SCMs to other aerosols will be similar to their response to BC. We do not have a definitive reference for the time-dependent response to aerosol forcing perturbations. Instead, we rate the SCMs using the difference from the average of both MAGICC models, which both differentiate aerosol forcing between land and ocean, which results in a faster overall climate response to aerosols as compared to greenhouse gases (Shindell et al., 2014).

In the case of BC, we note that all SCM response ratings should be reduced from the values shown because they do not accurately represent the temporal response to a BC step found in an ESM (S12). A more definitive evaluation of climate system responses to aerosol perturbations would be useful. This would require additional GCM simulations to step emission changes for various aerosol species and/or forcing mechanisms. There are currently two studies that have conducted this test, one study specifically investigated NorESM's response to black carbon (BC) perturbations (Sand et al., 2016) and a more recent study that conducted similar BC perturbations in CESM (Yang et al., 2018 *in discussion*).

		Model				
Impulse	**Species**	**Hector v2.0**	**MAGICC 5.3**	**MAGICC 6.0**	**FAIR v1.0**	**AR5-IR**
Forcing	CO_2 impulse	•••	•••	•••	••	•
	$4xCO_2$ step	••	••	•••	••	•
GHG Emissions	CO_2	•••	•••	•••	••	•
	CH_4	•••	•••	•••	--	••
Aerosols*	SO_2, BC	••	•••	•••	--	•

*Table 2: Summary of SCM Performance. The performance scale is generally based on the maximum percent difference in time-integrated temperature response compared to the relevant reference (generally comprehensive SCM average in SI 9). ••• : 0-10%, •• : 10-20%, • : 20-30% difference (SI13). * This ranking refers to aerosol response in general, which do not differ substantially for different aerosol types in these models. For BC specifically, all ratings should be reduced since none of the SCMs accurately represent the temporal response for BC seen in ESMs (Sand et al., 2016) (SI12).*

There are numerous benefits to using simplified models, but the selection of the model should be rooted in a clear understanding of the model responses (see Table 1). Our work illustrates the necessity of using fundamental impulse tests to evaluate SCMs and we recommend that modeling communities adopt them as a standard validation suite for any SCM. Given that idealized SCMs are biased in their response patterns, more comprehensive SCMs could be used for many applications without compromising on accessibility or computational requirements."

I note, however, that the ESM responses plotted seem to be smaller than typically reported for ESMs. Some of the model runs are also present in the ensemble of Geoffroy et al., 2013, and two of them are possible to recognize in the cloud of response curves. These are the MIROC5 and GISS-E2-R. The MIROC5 run has a characteristic oscillation in the response which is easy to detect in the cloud, and GISS-E2-R is the lower curve in the cloud. For both the temperature values seem to be scaled down by a factor around 0.7 compared with the corresponding curves in Fig. 2 of Geoffroy et al., 2013. The authors should clarify this discrepancy. I notice that if the cloud is adjusted by such a factor, the comprehensive SCM curves (for ECS=3.0 degrees) in Figure 4 will appear much more centered within the range of the ESM cloud.

Conducting impulse tests with complex models is computationally expensive, illustrated by the few studies employing this technique to understand the responses of models. We cite the Sand et al., 2016 study that specifically investigated NorESMs response to black carbon (BC) perturbations (Sand et al., 2016). We now include another study that conducted similar BC perturbations in CESM (Yang et al., 2018 *in discussion*). Other stylized CMIP5 experiments, such as the 1% CO_2 concentration experiment, are not

included in our comparison because we do not consider them to be impulse response tests. It is not possible to cleanly extract the impulse response from the 1% experiments. The CMIP5 4xCO_2 concentration step experiment is mathematically related to impulse responses, so are a reasonable comparison, particularly because these are the largest suite of such tests conducted in complex models, which is the reason we highlight these results in the paper.

Geoffroy et al., 2013 reported the 4xCO_2 concentration step temperature change relative to the 150-year temperature mean from the corresponding pre-industrial control run. For comparison to the simple models, we report the drift corrected (see S3) 4xCO_2 concentration step temperature change relative to the start of the 4xCO_2 concentration run. Therefore, there will be a difference in the temperature reported. We included this additional information in the revised supplement to clarify the difference in the way modeled temperature change is reported.

Figure R3 shows the global mean temperature response from the 4xCO_2 concentration step experiment for the 20 CMIP5 models used in our comparison following the Geoffroy et al. (2013) procedure of reporting the 4xCO_2 concentration step temperature change relative to the 150-year temperature mean from the corresponding pre-industrial control run. The responses reported in Figure R3 are consistent with Geoffroy et al. (2013). We expanded the number of complex models and updated the supplementary materials accordingly.

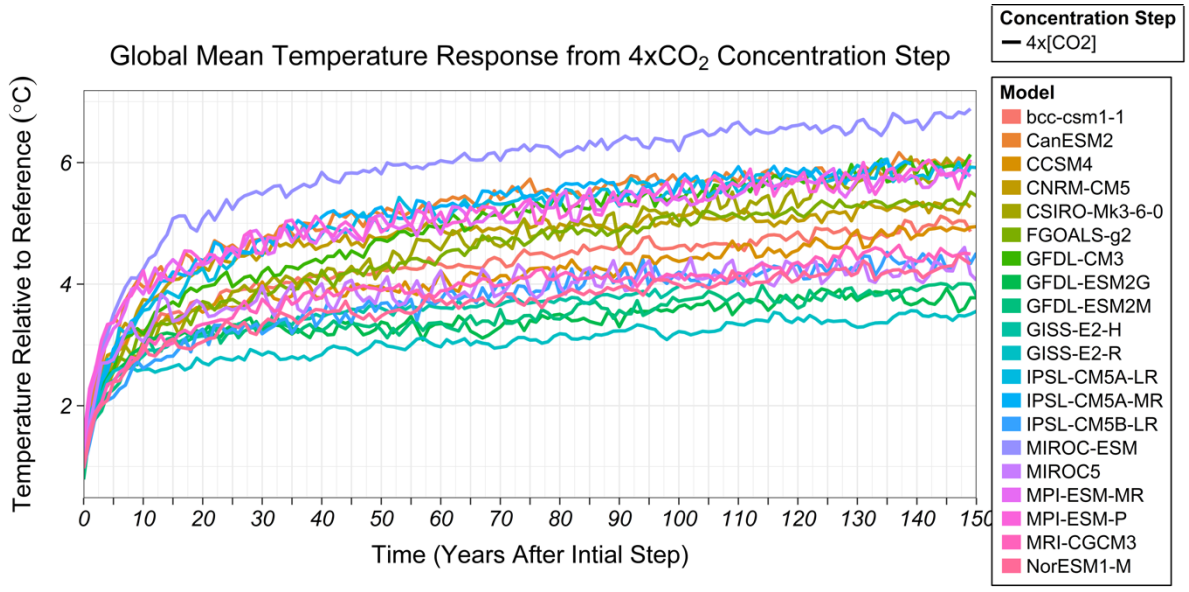

Figure R3 Global mean temperature response from 4xCO_2 concentration step in 20 CMIP5 models.

I cannot see where it is shown in the paper that comprehensive SMCs fail to capture response timescales of ESMs to CO2 forcing. This is not apparent in Figure 4.

To clarify, in Figure 4 of our manuscript the rate of temperature response from the SCMs immediately following the 4xCO_2 step is generally faster than the rate of temperature response from the ESMs. We also illustrate this in Figure S22 where we will expand the discussion in the revised manuscript, as we mentioned above. From this, we conclude that some SCMs do not capture the response timescales of ESMs.

Finally, I would urge the authors to discuss more explicitly unspoken assumptions underlying their conclusions, and also to make more explicit reference to the results from which these conclusions are drawn. For instance, in the abstract one can read:

<u>Line 17:</u> "While idealized SCMs are widely used, they fail to capture important global mean climate response features, which can produce biased temperature results."

Our language was vague in the abstract and we provide revised text to more explicitly reference our results.

"We find that while idealized SCMs are widely used, they fail to capture the magnitude and timescales of global mean climate responses under emissions perturbations, which can produce biased temperature results."

Since observations are not used in this study, the underlying assumption is that increased model complexity yields more correct results for global response features. This is not obvious. All climate models must be parametrized and constrained against observation. This means parameter fitting, and increased complexity increases the chance of overfitting. Complex models, and ESMs in particular, will to a great extent be parametrized against observations of local processes and not on the global responses. The large spread in the global responses of ESMs is a clear indication that they cannot be used as a substitute for observation of global responses.

We disagree with the reviewer that the large spread in ESM global mean temperature responses means they are not useful. While some climate studies benefit from using observations, we cannot employ observations to compare with impulse response tests, as we mentioned above. As noted previously, ESMs are constrained by more detailed representations of the relevant physics (e.g. energy balance, heat transport, etc.) and the multi-model mean of ESMs does a better job of matching observations than any individual ESM. The suite of ESMs results are, therefore, one of the best (albeit not perfect by any means) tools by which we can compare SCMs.

<u>Line18:</u> "Comprehensive SCMs, which have non-linear forcing and physically-based carbon-cycle representations, show improved responses compared to idealized SCMs."

Again, a simple model fitted to observation can represent reality better than a more complex model fitted to observation, because overfitting of a complex model may weight real physical processes in an unrealistic manner.

While it is true that a simple model may fit observations better than a more complex model, we do not agree that this is an indication that the fit represents a better representation of reality. This may also mean that, due to a lack of physical constraints in an overly simplified model, a good fit is obtained for the wrong reasons. We again point out the long-standing finding that the multi-model mean for CMIP-class models better represents reality as compared to any individual model. This finding indicates that the physical processes represented in these models (some explicit, some parameterized) are providing meaningful constraints on the behavior of the coupled system.

In our experience, the overall results of these global models, such as global temperature change, are not fitted to observational datasets. Instead, individual components are developed and tested against appropriate observations (e.g., top of atmosphere radiative flux, cloud properties, laboratory measurements, etc.), which provides an emergent, aggregate model behavior (albeit, dependent on the properties of these numerous sub-systems.). Every GCM is wrong, at least in some specific aspects, but the evidence suggest that the behavior of these models taken together is a useful overall constraint on Earth system responses (Flato et al. 2013).

These impulse response tests allow us to determine the underlying dynamics of SCMs so as to better elucidate any potential issues with later analysis using these models. For example, a SCM with a faster overall temperature response to a forcing would return a different implied value of any fitting parameters (such as climate sensitivity) than a model with a slower fundamental response.

Line 20: "Even some comprehensive SCMs fail to capture response time scales of more complex models under BC or CO2 forcing perturbations."

The BC case may be true, but is based on one single simulation in NorESM.

There are now two studies that have conducted a BC emissions impulse in complex models (i.e., Sand et al., 2016 and Yang et al., 2018) and cited them above. We noted above that the Sand et al., 2016 study specifically investigated NorESMs response to black carbon (BC) perturbations (Sand et al., 2016), while another conducted similar BC perturbations in CESM (Yang et al., 2018 *in discussion*). Further, Shindell et al. (2014) concluded that without accounting for regional warming and feedbacks, simple models could overestimate aerosol impacts, though we note that some models such as MAGICC 5.3 and MAGICC 6.0 do have differential land-ocean and North-South hemisphere forcing.

Line 21: "These results suggest where improvements should be made to SCMs."
It would be very helpful if explicit improvements were suggested.

We avoided adding explicit suggestions on areas where SCMs could be improved because modeling groups have a variety of reasons for implementing different features and components in their models. We stated in our manuscript that "Given that idealized SCMs are biased in their response patterns, more comprehensive SCMs could be used for many applications without compromising on accessibility or computational requirements." Some modeling groups favor answering certain scientific questions versus flexibility versus computational intensity differently, for instance, and the purpose of our paper is to explore mechanism for assessing those differences to inform users. Nonetheless, we expanded the conclusion in our response to more fully discuss the scale used Table 1 and we believe this expanded discussion suggests areas of improvement.

Technical comments

The reference to Chapter 8 in IPCC AR5 WG3 (Myhre et al., 2013) for a description is not very user friendly. It took me a lot of time to identify the relevant part of that chapter and the corresponding Supplement.

We have added additional details in our citations of Chapter 8 in the IPCC AR5 for clarity. The manuscript and supplement have been updated.

In the main manuscript reference to sections, tables, and figures in the supplement are named SI1 etc., while in the supplement itself they are referred to as S1 etc. Be consistent.

Thank you for identifying this error. We have updated the manuscript to be consistent with the supplement.

On pages 58 and 61 in the supplement is referred to Figure 5 in the main paper. This figure does not exist.

Thank you for identifying this error. The reference should be to Figure 4, and we apologize for any confusion this might have caused. The supplement has been updated.

Citations

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason and M. Rummukainen. (2013). Evaluation of Climate Models. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA

Geoffroy, Olivier, et al. "Transient climate response in a two-layer energy-balance model. Part I: Analytical solution and parameter calibration using CMIP5 AOGCM experiments." *Journal of Climate* 26.6 (2013): 1841-1857.

Joos, F., Roth, R., Fuglestvedt, J. S., Peters, G. P., Enting, I. G., Bloh, W. V., ... & Friedrich, T. (2013). Carbon dioxide and climate impulse response functions for the computation of greenhouse gas metrics: a multi-model analysis. *Atmospheric Chemistry and Physics*, *13*(5), 2793-2825.

Millar, R. J., Nicholls, Z. R., Friedlingstein, P., & Allen, M. R. (2017). A modified impulse-response representation of the global near-surface air temperature and atmospheric concentration response to carbon dioxide emissions. *Atmospheric Chemistry and Physics*, *17*(11), 7213-7228.

Myhre, G., Shindell, D., Bréon, F. M., Collins, W., Fuglestvedt, J., Huang, J., ... & Nakajima, T. (2013). Anthropogenic and natural radiative forcing. *Climate change*, *423*, 658-740.

Sand, M., Berntsen, T. K., Von Salzen, K., Flanner, M. G., Langner, J., & Victor, D. G. (2016). Response of Arctic temperature to changes in emissions of short-lived climate forcers. *Nature Climate Change*, *6*(3), 286.

Shindell, D. T. (2014). Inhomogeneous forcing and transient climate sensitivity. *Nature Climate Change*, *4*(4), 274.

Yang, Y., Wang, H., Smith, S. J., Ma, P. L., & Rasch, P. J. (2017). Source attribution of black carbon and its direct radiative forcing in China. *Atmospheric Chemistry & Physics*, *17*(6).