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

Anonymous Referee #2
Received and published: 23 December 2018

Dear Referee #2,

We want to begin by thanking you for taking the time to read our manuscript. Though we disagree with the conclusions of your review, your comments have helped us clarify and improve the manuscript. We have copied the unedited original comments in bold and our responses are provided in regular font, indented from the original comment for clarity. We continued the number of figures and tables from the first review to be consistent when revising our manuscript.

SCMs are routinely used to emulate state of the art GCMs, and generally display reasonable (though not perfect) agreement when tuned specifically to do so. The authors themselves cite several papers relating to this which discuss strengths and weaknesses of such emulation. While of course SCMs can also be integrated with standard (default) parameter values to provide some guidance as to how the climate system may behave, these simulation will not encapsulate our uncertainty in the best parameter values to use. Furthermore, such simulations will depend greatly on how the default parameter values were chosen, which may differ between SCMs.

> The aim of this paper is not to validate any individual simple climate model (SCM), nor the range of parameters used in the SCMs, which are also explored in the literature cited in our manuscript as you note. Rather, we are evaluating the fundamental behavior of the simple models. However, we do agree that understanding the uncertainty associated with our results is important and based on the comments here and from Reviewer #1, we have now included the parameter files in the supplement so show the default parameters of the models.
>
> In our original supplement we conducted a simple sensitivity test for the 4xCO_2 concentration step experiment by changing the climate sensitivity values in the three comprehensive SCMs used in this paper. Based on your concerns, we have added some additional tests to the supplement of our paper by exploring a range of climate sensitivity values and ocean diffusivity values in MAGICC 6.0 under a unit pulse of CO_2 emissions and a unit pulse of CO_2 concentration.
>
> We selected climate sensitivity and ocean diffusivity values from the parameter ranges presented in Table 1B in Meinshausen et al., 2011. The values are the native MAGICC 6.0 parameters required to emulate complex models used in CMIP3 using three calibrated parameters (climate sensitivity, ocean diffusivity, and land/ocean warming). We provided the climate sensitivity and ocean diffusivity value ranges we explored in Table R2 below.

Table R2 MAGICC 6.0 parameter values from Meinshausen et al., 2011 Table 1B for sensitivity tests

Scenario	Climate sensitivity (K)	Ocean diffusivity (cm^2s^{-1})
Base Case	3.0	1.1
High Ocean diffusivity	3.0	3.74
Low Ocean diffusivity	3.0	0.50
High Climate sensitivity	6.03	1.1
Low Climate sensitivity	1.94	1.1

Figure R4 shows the global mean temperature response exploring the range of ocean diffusivity (Kz) (a) and global mean temperature response exploring the range of climate sensitivity (CS) (b) under a CO_2 emissions perturbation. Figure R5 shows the same results for under a CO_2 concentration pulse. Both figures illustrate that climate sensitivity has the greatest impact on the responses and in our manuscript, we accounted for this and used similar climate sensitivity values in SCMs where possible, unless otherwise noted in the supplemental figures.

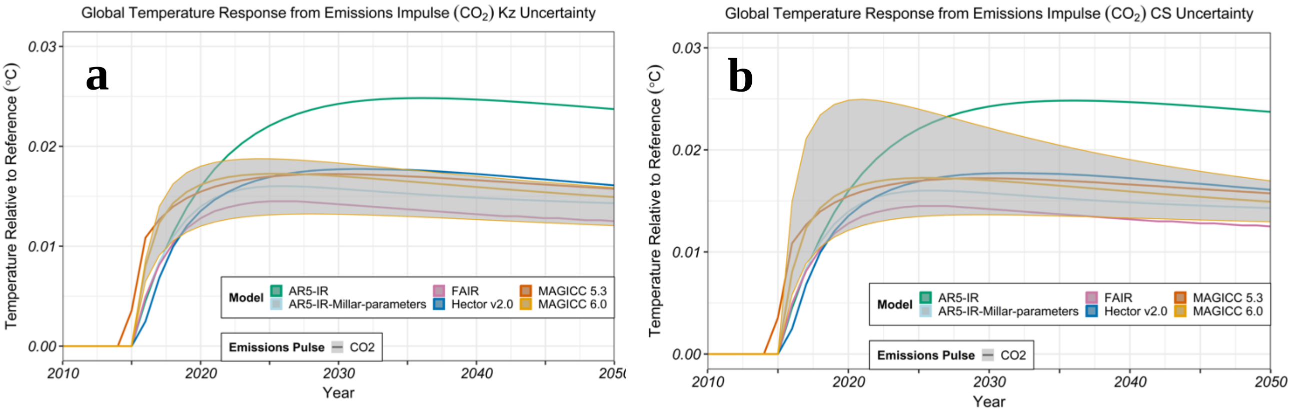

Figure R4 Global mean temperature response exploring the range of ocean diffusivity (Kz) (a) and Global mean temperature response exploring the range of climate sensitivity (CS) (b) from a CO_2 emissions perturbation in SCMs (MAGICC 6.0 – yellow, MAGICC 5.3 BC-OC – red, Hector v2.0 – blue, AR5-IR – green, FAIR –pink, AR5-IR-Millar-parameters –light blue). The grey shaded region in each figure shows the range in MAGICC 6.0 responses found using the Table R2 parameters. We note that the range of responses exploring CS (b) are normalized to account for the different climate conditions under difference CS values.

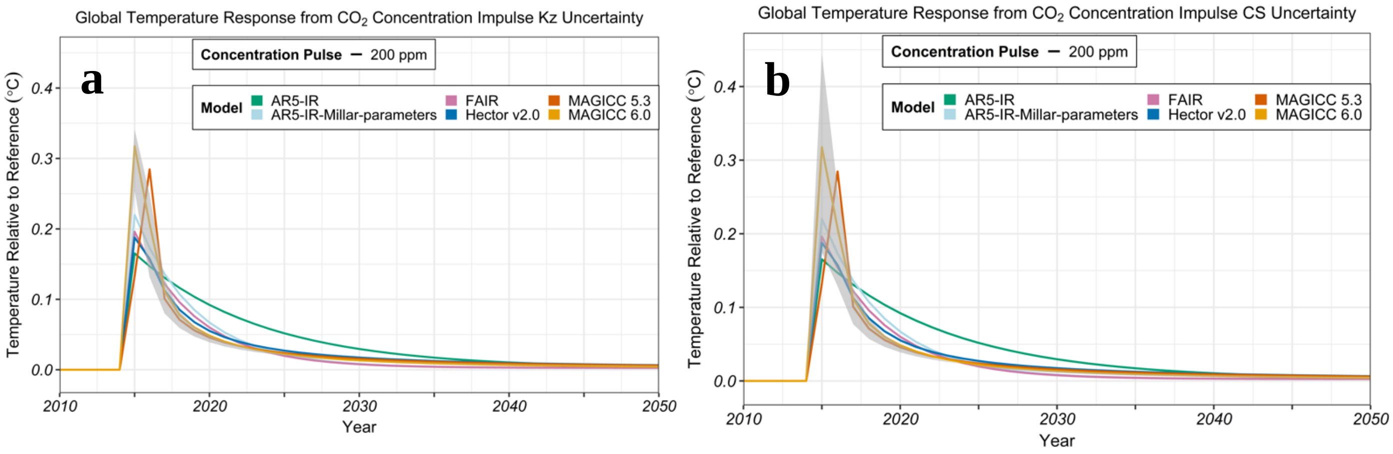

Figure R5 Global mean temperature response exploring the range of ocean diffusivity (Kz) (a) and Global mean temperature response exploring the range of climate sensitivity (CS) (b) from a CO_2 concentration pulse in SCMs (MAGICC 6.0 – yellow, MAGICC 5.3 BC-OC – red, Hector v2.0 – blue, AR5-IR – green, FAIR –pink, AR5-IR-Millar-parameters –light blue). The grey shaded region in each figure shows the range in MAGICC 6.0 responses found using the Table R2 parameters. We note that the range of responses exploring CS (b) are normalized to account for the different climate conditions difference CS values.

We acknowledge, however, that vertical ocean diffusivity has a large impact on ocean heat uptake and we do note that this parameter selection also impacts the responses in the SCMs, particular under a CO_2 emissions pulse (Meinshausen et al., 2011). However, the SCMs we compare in our paper either do not have the same definitions of vertical ocean diffusivity, as is the case for the comprehensive SCMs, or ocean diffusivity is not directly represented In the models, as is the case for idealized SCMs. For our purposes, therefore, we kept the ocean diffusivity values at their default values within the comprehensive SCMs. By exploring the uncertainty in ocean diffusivity, we have, in fact, bolstered the main conclusions of our manuscript.

For completeness we also acknowledge that Meinshausen et al., 2011 spanned ranges of land/ocean warming contrast (RLO) in the three-parameter calibration described in Table 1B of their manuscript. And again, the SCMs either use the same values of RLO, as is the case for both versions of MAGICC, or this parameter is not represented in the idealized models. In fact, from our work using impulse response test to characterize SCMs, we concluded that SCMs without differential warming do not correctly capture the response pattern to BC perturbations.

Given that the GCMs disagree substantially amongst themselves, I do not understand the purpose of this paper in comparing the outputs of standard SCM instances to themselves and GCM output. It is inevitable that these will not match closely when the SCM parameters are set to standard values, and I do not think it is straightforward to attribute such differences to structural limitations of the SCMs without first checking that they cannot be explained by parameter choices.

We remind the reviewer that we are not attempting to emulate GCMs in our paper. Instead, we evaluate SCM responses by comparing the models to themselves and also, in the limited cases where this is possible, to more complex models.

One key purpose of this paper is to determine the fundamental response of these models by conducting impulse response tests (as recommended, for example, by a recent report by the NAS). This has not been done before, and this alone provides useful information on the behavior of these models, how this differs between the models, and the magnitude of those differences. Given the extremely widespread use of these models, this is a critical task.

We go beyond just comparing these models to each other, by comparing against the suite of complex model results were this is possible. We do this 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). Further, comparing simple models to complex models is a common technique employed in the literature (e.g., Joos et al., 2013). For example, in the abrupt 4 X CO2 regime, we find that the SCMs as a group initially respond more quickly than the GCMs. We conclude from this that there must be some physical processes represented in the GCMs that buffers the initial response in this regime that is lacking in the SCMs.

From our exploration of the range of ocean diffusivity values and climate sensitivity values above, we have found that the response differences we noted in our conclusions cannot be explained by parameter choice alone. We appreciate that the reviewer brought up this important question, and we will amend our paper discussion based on these results but note that our main conclusions remain. In fact, we believe that illustrating the ranges in responses further bolsters our claim that impulse response tests are needed to fully understand model behavior.

Of course in the simplest of cases one might show that a complex curve output by a sophisticated SCM/GCM simply cannot be explained by a very simple parametric form, but even here it would be appropriate to explore how close a fit could be obtained.

Fitting individual simple models to more complex models is generally explored by the individual SCM development teams, and we cite papers from the Hector, MAGICC, and FAIR model development teams which explore their respective model’s ability to fit a GCM (or the multi-model mean) with a given set of parameters. While emulation is outside the scope of this paper, but to address the reviewer’s comments, above we have expanded our impulse tests to include an uncertainty test which relies on parameters derived from GCM emulation experiments using MAGICC 6.0. We will also add a discussion of fitting SCMs to more complex models to the paper to address the points raised by the reviewer.

One could reasonably compare SCM responses amongst themselves when tuned to each other or to some common target (either observational or GCM-based). However, this has not been performed here. While in some experiments the sensitivity parameter has been set to a common value of 3, other model parameters appear to differ between the SCMS and were apparently set to standard values which were probably chosen by the SCM authors for a variety of reasons. Thus it is not possible to determine how much of the differences in response are due to model structure, and how much is the result of using different parameter values/tuning strategies.

We remind the reviewer that the goal of our paper is to evaluate the SCMs, as we mentioned above, and we ultimately suggest a suite of fundamental impulse-response tests using realistic backgrounds for use in SCM development. We make this clearer in our revised manuscript by including some of the text mentioned above, such as:

"In our paper, we evaluated the SCMs by comparing the models to themselves and also, in the limited cases where this is possible, to more complex models. 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)."

The sensitivity tests we have added to the paper do provide useful general information on how parameter choices might influence model responses, which addresses part of the question posed above. (We note also, that in response to reviewer 1, we have also added an example of how the idealized AR5 model response changes with a change in parameters.) Our text will be amended to reflect these results. However, as we noted above, due to structural differences in the SCMs it is, in general, not possible to operate the models with identical parameter values. This reinforces the importance of conducting fundamental impulse response tests to quantify the behavior of the SCMs.

I would also question whether the relatively unrealistic abrupt tests are a useful diagnostic tool for the model behaviour. While I accept it can be interesting to characterise the response to idealised forcing scenarios, it may be that the differences are much less significant when more realistic scenarios are applied, and the authors acknowledge this point in their conclusions.

There is a long history of doing just such idealized, abrupt tests to evaluate model behavior. The CMIP5 4xCO_2 concentration step experiments are the largest suite of impulse response tests conducted in complex models, for example, which is the reason we highlight these results in the paper.

The impulse response tests conducted enable us to uncover differences in model behavior that are not apparent when running standard, multi-emission scenarios. Indeed, one of the important uses of SCMs is to conduct model experiments were there may be relatively small changes in emissions between two scenarios. Because SCMs do not exhibit internal variability, such experiments can be used to quantify such changes. Impulse response tests also allow us to understand, on a more fundamental level, differences between SCMs that have been found comparing simulations with more conventional scenarios (e.g., van Vuuren et al., 2011).

Thus, this analysis does not sufficiently advance our understanding of the behaviour of SCMs, and I am sorry to say that I cannot support publication of this manuscript in ESD.

We believe this paper does present novel and useful results that are new to the literature. Though fundamental impulse tests have been used a few times in the literature in this context, our manuscript employs these existing techniques in a new manner. 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. 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. Additionally, *we provide a set of tests that we recommend as a standard evaluation suite for any SCM.*

Reviewer #2: As a minor comment, the "unit testing" terminology seems inappropriate, the test here is rather more comprehensive than such a term usually implies, and furthermore there does not appear to be any clear criteria for success or failure.

We received several comments on our use of the phrase "unit testing". Though we believe our use of the phrase is consistent with its use in software, as we replied to the Short Comment and Reviewer #1, we will update the language to "fundamental impulse tests" to avoid confusion.

Citations

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C. and Rummukainen, M.: Evaluation of Climate Models, Clim. Chang. 2013 Phys. Sci. Basis. Contrib. Work. Gr. I to Fifth Assess. Rep. Intergov. Panel Clim. Chang., 741–866, doi:10.1017/CBO9781107415324, 2013.
Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K. and Knutti, R.: Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks, J. Clim., 27(2), 511–526, doi:10.1175/JCLI-D-12-00579.1, 2014.

Joos, F., Roth, R., Fuglestvedt, J. S., Peters, G. P., Enting, I. G., Von Bloh, W., Brovkin, V., Burke, E. J., Eby, M., Edwards, N. R., Friedrich, T., Frölicher, T. L., Halloran, P. R., Holden, P. B., Jones, C., Kleinen, T., Mackenzie, F. T., Matsumoto, K., Meinshausen, M., Plattner, G. K., Reisinger, A., Segschneider, J., Shaffer, G., Steinacher, M., Strassmann, K., Tanaka, K., Timmermann, A. and Weaver, A. J.: Carbon dioxide and climate impulse response functions for the computation of greenhouse gas metrics: A multi-model analysis, Atmos. Chem. Phys., 13(5), 2793–2825, doi:10.5194/acp-13-2793-2013, 2013.

Meinshausen, M., Raper, S. C. B. and Wigley, T. M. L.: Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 - Part 1: Model description and calibration, Atmos. Chem. Phys., 11(4), 1417–1456, doi:10.5194/acp-11-1417-2011, 2011.

National Academies of Sciences, Engineering, and M.: Assessment of a Near-Term Update to the Social Cost of Carbon, edited by T. N. A. Press, Washington, DC., 2016.

van Vuuren, D. P., Lowe, J., Stehfest, E., Gohar, L., Hof, A. F., Hope, C., Warren, R., Meinshausen, M. and Plattner, G. K.: How well do integrated assessment models simulate climate change?, Clim. Change, 104(2), 255–285, doi:10.1007/s10584-009-9764-2, 2011a.