Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2018-63-RC1, 2018

Earth System Dynamics Discussions

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

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 CO_2-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).

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.

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.

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

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

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

Reasonable credit is given to related work.

The title should find another term than “unit testing”.

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

The presentation and language is adequate.

Specific comments

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 CO_2-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 CO_2 concentration is simulated. However, in both figures the responses of the two models are very different. 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, $d_1 = 8.5$ yr and $d_2 = 409.5$ yr), while in FAIR they look more like the choice of Millar et al. 2017 ($d_1 = 4.1$ yr and $d_2 = 239.0$ yr).

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). 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.

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.

The total forcing response to CO_2 and CH_4 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 d_1 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 $d_1 = 4.1$ 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.

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 $4 \times CO_2$ 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. 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. 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.

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?

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."

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.

<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.

<u>Line 20</u>: "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. I cannot see where it is shown in the paper that comprehensive SMCs fail to capture response timescales of ESMs to CO_2 forcing. This is not apparent in Figure 4.

Line 21: “These results suggest where improvements should be made to SCMs.”

It would be very helpful if explicit improvements were suggested.

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.

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.

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

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2018-63, 2018.