Second report from reviewer #1

The essence of my critique of the the manuscript by Schwarber et al. is contained in he following comment from my first report:

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 authors' response to this is:

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 necessarily 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 a broad suite of observations better than any individual model (e.g., Figure 9.7, Flato et al. 2013). Our subsequent responses will also address this comment.

Unfortunately, I do not think this justification is correct and rests on a flawed interpretation on results in the literature, including Fig. 9.7 in Flato et al. 2013 (Chapter 9 in the IPCC AR5 report). Below, I will present my arguments.

Figure 9.7 in Flato et al (2013) deals with the RMS-difference between space-time global seasonal-cycle climatology of models and observations. This means that in every grid cell the monthly climatology is computed based on the years 1980-2005 to produce a mean annual cycle for this period in the model and in the observation (reanalysis), and the RMS-difference is produced. There are two features to notice: (i) The metric for comparing model with observation is based on the full space time-field, not the global average as done in the present manuscript (MS). (ii) The metric measures the RMS-difference over the annual cycle in historical runs/observation over a 25-year period, while Schwarber et al. measure the percentage difference of the time-integrated response of pulsed forcing experiments over 100 or 20 years. Hence the data compared and the metric used in Fig. 9.7 and in Schwarber et al. have very little in common.

The feature of Fig. 9.7 which Schwarber et al. use as justification is that the RMS-difference seems to be smaller for the so-called mean model than for any of the individual models, and that – with respect to this specific metric – the mean model is the better representation of reality. This has been shown empirically to hold true for many other model fields, not only for the annual cycle, but I have never seen it demonstrated for the long-time response of the global mean temperature for an impulse or step forcing.

It would actually have been a groundbreaking result, if this could be shown to be true, because the metric used by Schwarber et al. applied to the 4xCO2 step-forcing experiments would effectively measure the equilibrium climate sensitivity (ECS). If it were true that the mean model (the ensemble mean of the individual model experiments) is closer to reality than any of the individual models in this metric, then we would know that the ECS of the mean model is very close to thetrue ECS, and all the problems we have with the uncertainty in the ECS-estimate would evaporate.

A theoretical result explaining many observations like those in Fig. 9.7 was published by Annan and Hargreaves, J. Climate, 4537 (2011). It rests on the assumption that the observed reality and models are drawn from the same statistical distribution, but does not assume that this distribution is centered around the observation. They compute the probability that an ensemble member is closer to reality than the ensemble mean and show that it is generally small if the dimension of the data vector is large (se Figure 2 in that paper). For small effective dimension, however, this is no longer true. The metric used by Schwarber et al. measures only one number, the integrated response after 100 (or 20) years, so the data vector has dimension 1. This explains why one cannot use the ensemble mean of the complex models as "the truth" when assessing the performance of the simple models.

It probably will not help much to use a higher-dimensional data set to characterize the model solutions, since the simplest models are completely determined by a rather small number of model parameters, which renders the effective dimension small.

I have a number of reservations also with other aspects of the manuscript and the authors' response, but the problem I have discussed above is so serious that I cannot recommend publication.