Response to Anonymous Referee #1 (our comments in blue)

Received and published: 1 July 2016

The authors propose a nonparametric emulator aimed at reproducing geoengineering scenarios. Using dynamical linear models, they propose a formulation of the emulator as a convolution of the forcing and the impulse responses, and test this approach for two geoMIP scenarios for some variables of interests.

The manuscript is overall well written and presents an interesting problem, but I believe that in its present form is not suitable for publication and, in order to be reconsidered, needs to be considerably improved in many parts. The proposed method would have considerable limitations if it is to be expanded beyond the narrow context of this work, e.g. annual averages for two model runs. Further, the validation setting is extremely limited, not based on any metric, and completely ignores the emulation uncertainty.

> We agree that we could do a better job of quantifying emulator accuracy (i.e., evaluating based on specific metrics); we will address this in revision.
>
> The description in the paper of how to extend results to cover sub-annual time-scales was poorly worded and thus misleading. The case of evaluating sub-annual variables in response to annual-mean forcing is already covered by the method (indeed we considered one such variable but did not include a plot in the paper, we will do so in revision). The description in the paper regarding sub-annual scales was intended to refer to the case where the forcing varies at sub-annual time-scales. Conceptually this is a trivial extension, but would require additional training data; this is not a limitation of our method per se, but only a limitation that information cannot be invented from nothing. (You cannot infer the response to seasonally-varying forcing from a simulation where the forcing was constant, no matter what method you apply.)
>
> Also note that our formalism for capturing the linear forced response is intentionally provided in a more general fashion than most previous research. The key distinction relative to the existing literature aimed at the similar problem for non-geoengineering forcing is that we only assume linearity, and do not choose to make additional ad hoc (and apparently unnecessary) assumptions regarding the form of the dynamics; in doing so it helps clarify the additional assumptions made elsewhere. We will reword to better articulate the relationship with existing approaches.
>
> We disagree regarding the limitation of the validation setting, as described in more detail below. The perceived limitation regarding annual averages is not a limitation but simply poor wording on our part. Validation on one forcing scenario is sufficient to demonstrate that linearity is a sufficiently useful approximation for the variables that we consider here. Since that is the only assumption we make, additional validation scenarios would not add any further value.

General comments

• The validation setting is extremely limited: the proposed approach is fit for the G1 scenario, and then used to extrapolate G2. Also, for the G1 scenario the emulator is likely to work well, since it consists of impulse functions. A considerable amount of work is needed to perform more tests under different forcing scenarios. While the geoMIP is limited in size, the CMIP5 or other large multi-model ensembles could be used to validate the forcing part of this emulator.

> We demonstrated that a *linear* emulator trained on one scenario successfully follows the behaviour of another. Note that in contrast with other studies looking at climate emulators for greenhouse gas forcing, the *only* assumption we make is linearity (and time invariance). By demonstrating that linearity holds for forcing levels up to 4xCO2 and for levels of solar reduction sufficient to compensate this, it will clearly also hold for *any* other smaller-amplitude forcing scenario and hence no value would be added by validating it on additional forcing scenarios. We will reword the paper to clarify this.

• The present version puts very little emphasis of the uncertainty of the scenario estimation. The validation essentially consists in eyeballing many plots of the emulator against the original computer model, with no attempt to quantify the fit or, most importantly, to assess how the internal variability of the model is reproduced by the emulator. The definition itself of 'climate variability' $n_i(t)$ is unclear. Are the authors assuming a white noise? Also, I would assume that this noise is independent for different variables, but it should be clearly stated.

We agree that we should do a better job of quantifying the fit, we will address this in revised manuscript.

We also need to clarify in the revised text that the emulator is intended to only capture the forced response, and is not intended to reproduce internal variability. We will also clarify the text describing the effect of the spectral characteristics of the climate variability. Insofar as we are only interested in capturing the forced response, we do not need to make any assumptions about the nature of the climate variability as it enters only as "noise" in our ability to estimate the forced response; we will say this as well.

• This approach will have significant limitations at finer temporal scales. The authors briefly discuss this when they mention how we can impose h = h(_,m). This solution is not straightforward, as a nonparametric estimation of 12 different impulse responses will require more scenarios (surely more than two) to have reliable estimates. The authors somewhat acknowledge it when they state that additional simulations would be required, but in an off-the-shelf ensemble such as geoMIP, where no more scenarios are readily available, this is a strong limit of this approach. This will be become even more evident for finer temporal scales, e.g. weekly or daily data.

As noted above, we apologize for badly worded text here that was potentially misleading. To clarify, if the *forcing* does not vary significantly over the course of the year, then emulating GCM response at finer temporal scales is not intrinsically more difficult for this or any other emulator, although the signal to noise ratio (SNR) will likely be poorer (a limitation that is intrinsic to the information contained in the training data, and has nothing to do with the method itself).

If the intent is to capture the response to seasonally-varying forcing, then unless arbitrary assumptions are made regarding the seasonal dependence of the impulse response, one would need at least as many independent forcing scenarios as degrees of freedom of the seasonal response (i.e., 12 if one wants to distinguish how the response depends on monthly-varying forcing). This is not a limitation of the formulation we use (it would be a trivial extension), rather it is an intrinsic limitation on the knowledge of the response that holds for any such approach. Of course, for evaluating climate change in response to different pathways of greenhouse gas forcing and solar geoengineering, the forcing varies only slowly from year to year, so that the additional training data is not needed.

We did a poor job of articulating the distinction between these two cases, and will correct this.

We focused on information about annual-average behaviour because it is indeed useful, both for geoengineering and more general climate science applications although we could certainly include some sub-annual variables in revision (the only one we looked at in writing the paper was annual-minimum sea ice extent for which nonlinearity is significant and the emulator does not perform well; we will add this to the revised paper).

• The results and the discussion do not mention model differences, and most importantly what do they mean. Does the emulator estimate different impulse responses for different models? I would expect so, and I would expect these differences to convey information on how the models differ. For example, HadCM3 and HadGEM2-ES will likely display similar responses as both models are released from the Hadley Centre.

Our intent was to develop an approach for emulating climate models, not for describing the differences between them, for which there is already an abundant literature on the general differences between climate models in terms of processes they include, differences in how they respond to climate change, and differences in how they respond to geoengineering. We will add some text to this effect and appropriate references, although given the vast breadth of this sort of literature, the number of citations we can include will naturally be limited..

• The part on grid-scale emulation must be extended. Firstly, the methodology is unclear: a clear explanation of how were the EOFs selected must be presented, either in the main text or in the supplement. Secondly, as before, a more formal assessment of the pattern similarity is needed, as eyeballing figure 5 is not enough to convince that the emulator is performing well.

We agree that this section was too terse. We will add both a more complete description, and a more formal pattern similarity assessment.

Specific comments

• Title: what the authors present is not a multi-model emulator, in the sense that it independently fits each model and does not assume interdependencies.

We agree that there are multiple ways of interpreting the phrase "multi-model". We meant it only in the sense that the end result is a set of emulators (i.e., in the same sense that CMIP or GeoMIP are "multi-model" ensembles.) We will clarify this in the revised version.

• pag. 1 l.16-17. The claim that the 'emulator prediction may be a more accurate estimate [...] of the models' response than an actual simulation' is very questionable. The emulator is not meant to replace a climate model, it's just a faster approximation that is used to explore the input space in a computationally efficient manner. While emulators are arguably a useful tool for calibration and, as in this case, scenario extrapolation, they cannot replace the physics of the climate model and they are useful only as long as the training set from the climate model is meaningful.

On the final point, we of course agree completely – the climate model is needed to generate training data for the emulator, and an emulator cannot completely replace climate models. We also agree that the emulator is only useful so long as the training set is meaningful. We will ensure that these points are clear in the revised version.

As to whether the emulator prediction is a more accurate estimate for some specific scenario, that depends both on the purpose and on the input/output response of the dynamic system being emulated. Our goal is to estimate the *forced* component of the response, isolated from natural variability, and in doing so we approximate the response as linear. A single GCM simulation of a particular scenario will not give a perfect estimate of the forced response, due to the presence of natural variability, while the emulated response, with an emulator trained from a simulation at higher forcing amplitude, will introduce some error due to nonlinearity, but *less* uncertainty due to natural variability. *Fundamentally one is simply trading off the uncertainty in the forced response that comes from superimposed natural variability from the uncertainty that comes from nonlinearity*. The best answer for the forced response would come from a sufficiently large ensemble of GCM simulations of the specific scenario, but given sufficient computation for one single simulation, then it is not a priori clear whether the best estimate of the forced response in a particular scenario is obtained by simulating that particular scenario… it may well be true that simulating at a higher forcing amplitude, to give a higher SNR, and then scaling the response, would indeed give a better estimate. (If the system were perfectly linear in its response to forcing, this is self-evident.) We will revise the manuscript to make this point more clearly.

• pag 1. l.19-20. Actually, emulators are much more popular in model calibration and local sensitivity analysis of physical parameters then in projections of anthropogenic forcings. Only very recently this methodology have been extended to deal with forcings. This introductory part must be rewritten with a more extensive literature review on traditional emulators.

We can add some references to acknowledge the prior history. However, it is only the recent extension to deal with forcings that is directly relevant to the case here.

• pag. 4, eq (1) and onwards. It is somewhat inappropriate to represent the emulator as a convolution given that the authors are effectively using just annual averages. A reformulation in terms of discrete sums is necessary.

Agreed that we should describe in terms of discrete sums; the continuous-time derivation was provided only because we felt that some readers might be more comfortable with it. We will fix this.

• pag. 4, line 101. h(_) was never defined.
• pag. 6, line 161. Poor choice of pedix in $f_t(t)$, please reformulate.

Thanks; we will clarify notation

• Figures. What is the unit measure of precipitation? Also, are the all figures expressed as anomaly with respect to a reference value? If so, what is it?

Oops. Sorry, final version of figures were generated but didn't get included! Not sure how that happened or passed final proof-reading. Units are in mm/day, and are all in anomalies with respect to the preindustrial control values.