Response to Anonymous Referee #1

> Thanks again for your careful review; our responses are in blue below. We uploaded a response shortly after this review, but waited until the other reviews were complete before revising the paper; this comment is mostly redundant with our original response, but is now also updated based on the revisions made.

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 quantifying emulator accuracy (i.e., evaluating based on specific metrics) is essential; Table 1 now includes a comparison for each model and each variable of the root-mean-square difference between the emulator prediction and the simulation, compared with the natural variability.
>
> The description in the paper of how to extend results to cover sub-annual time-scales was poorly worded and thus quite 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 original paper,. It has now been included both because it makes this point, and because it also illustrates a breakdown of linearity (these two are unrelated). 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 clarified this in the introduction 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. (If the reviewer is used to dealing with nonlinear emulators, for example as used in model tuning, then it would be absolutely true that you can neither tune nor validate such an emulator from single simulations.)

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. Since linearity is the only assumption we make, then (i) the emulator can be uniquely specified from a single forcing simulation such as G1, and (ii) no value would be added by validating it on additional forcing scenarios. 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. We reworded 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 added a quantification of the fit; we agree that not doing so was an oversight in the original manuscript.

We also clarify that the emulator is intended to only capture the forced response, and is not intended to reproduce internal 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. This has been clarified in the manuscript.

• 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 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 have corrected this.

We focused on information about annual-average behaviour because it is indeed useful, both for geoengineering and more general climate science applications. We have added one sub-annual variable in revision; the annual-minimum sea ice extent. This is provided not to illustrate the ability to project sub-annual variables, but because it illustrates a case where nonlinearity is significant and the emulator does not perform well. (That this particular variable happens to be nonlinear is unrelated to the fact that this particular variable is at a sub-annual time scale.)

• 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. We added some text to this effect and appropriate references.

• 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 added both a more complete description, and a more formal pattern similarity assessment (rms differences).

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.) This is a not uncommon usage of the adjective, nonetheless we removed it from the title (shorter is better).

• 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.

> This point was not well worded in the original manuscript, and we have endeavoured to clarify what we meant, both here and in the text.
>
> And, of course the climate model is needed to generate training data for the emulator, and does not replace climate models. We also agree that the emulator is only useful so long as the training set is meaningful.
>
> As to whether the emulator prediction is a more accurate estimate for some specific scenario, that depends on the purpose. If the goal is to estimate the *forced* component of the response, isolated from natural variability, then 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 *for a given amount of computation*. (If the system were perfectly linear in its response to forcing, this is self-evident.) If computational power is unlimited, then of course the best answer for the forced response would come from a sufficiently large ensemble of GCM simulations of the specific scenario.
>
> 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. Given sufficient computation to conduct only 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, or simulating a higher SNR scenario and using an emulator to "scale" it.

• 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 added a comment and reference to acknowledge the breadth of application of emulators. However, it is only the use of emulators to deal with forcings that is directly relevant to the case here. (As a side note, 1990 probably doesn't count as "very recently" any more!)

• 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. The emulator is now given in terms of discrete sums; the continuous-time is still introduced in case some readers are more comfortable with it, and in particular, we motivate the influence of the

climate variability spectrum through Laplace transforms of the continuous-time convolution equation; readers are undoubtedly more comfortable with them than Z-transforms that would otherwise be needed.

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

Thanks; fixed.

• 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; fixed. 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.

Response to Anonymous Referee #2

Thanks for the comments; our responses are in blue below.

The authors used model results from Geoengineering Model Intercomparison Project (GeoMIP) to test the linearity of the climate response to external forcings. The authors first constructed a climate emulator based on a convolution of impulse response function using results from GeoMIP G1 simulations involving abrupt changes in atmospheric CO2 and solar irradiance. Then the authors used the climate emulator to predict climate consequences of the GeoMIP G2 simulations involving gradual change in atmospheric CO2 and solar irradiance. For climate variables including temperature, precipitation, and annual mean Northern Hemisphere sea ice extent, the emulator does a good job in reproducing climate model simulated temporal evolution and spatial distribution.

The use of impulse response function to emulate climate model results is not new. The novelty of this study is that it extends the application of impulse response function to the simulations involving both CO2 and solar forcing. This extension advances our understanding of climate response to external forcing, and in particular, climate response to solar geoengineering. The ms is well written. I recommend publication after the following issues are addressed:

1. The GeoMIP simulations are limited to a period of 50 years. Over longer timescales (several centuries), response from deep ocean dynamics would become important. Many aspects of ocean dynamics response (e.g., thermohaline circulation) are nonlinear. So the question is: To what extend the linear emulator would be valid in reproducing long-term climate response involving feedbacks from deep ocean dynamics?

> Agreed; we have added a comment to the manuscript regarding this point, including both a citation to the literature on AMOC nonlinearity (acknowledging the limitation in using 50-year training simulations), and to one study indicating that the net response (combining CO2 forcing and solar reduction) does not drift (which at least gives some confidence that the long-term climate response would not be radically different from the short-term, at least in one model).

2. A large part of the residual response of the hydrological cycle over land to solar geoengineering is due to the direct effect of increasing atmospheric CO2 on vegetation (stomatal, leaf area index, etc.), which cannot be offset by reduced solar forcing. Assumedly, this part of hydrological cycle response is nonlinear. This issue should be discussed.

> The reviewer raises an important distinction here, between whether the overall processes involved are nonlinear, versus whether the perturbation in the response is approximately proportional to a perturbation in the forcing (so double the forcing doubles the perturbation). As long as the nonlinear relationships in question are differentiable at the current equilibrium point, then by definition there is a linear first-order response, although the size of perturbation for which that is relevant is not a priori clear. We have endeavoured to clarify the wording regarding linearity in a few places, both at the beginning of section 2, and with a more thorough discussion of which variables have an apparently nonlinear response in which simulations (since the difference between emulated and simulated responses indicates nonlinearity, if for example G2 precipitation (suppressing the slow response) is well predicted but not the 1%CO2 simulation, then one can conclude that the fast response is relatively linear, but that nonlinearities arise in the slow response to precipitation.)

3. The method used to emulate spatial pattern of temperature and precipitation is not clear. How EOFs were constructed, selected, and applied to generate the spatial pattern of climate change? These should be elaborated.

Agreed; we have added section 2b to describe the spatial EOF analysis.

Response to Anonymous Referee #3

Thanks for the detailed review! This is very helpful; our comments are in blue below.

The paper entitled "Multi-model dynamic climate emulator for solar geoengineering" by MacMartin & Kravitz presents a simple numerical emulator of the complex GeoMIP models which could be used to discuss Geoengineering scenarios. This paper is well written, fairly straightforward, and is interesting – I believe – for the community.

One could wonder, however, if ACP is the best journal for publication, as a lot of technical detail regarding the modeling (i.e. establishing the response functions) is given, whereas the more physical aspects remain (maybe too) brief. Maybe GMD would have been a better choice. But that is ultimately an editorial issue. And I don't think this point alone prevents publication in ACP, especially as the physics is well understood and already published elsewhere. It goes in favor, however, of improving the narrative so that the reader can grasp both the modeling approach and the modeled physical processes.

We have added some text regarding physical processes (particularly with regards to fast and slow responses, and some comments regarding what the difference between observed nonlinear effects between G2 and 1%CO2 simulations implies about nonlinearities in fast and slow responses).

Ultimately, I do recommend publication, but provided the few points below are answered.

Major points:

1. As mentioned: the end of the paper can be improve. Specifically, while the beginning (the methods, mostly) is well documented, the last part (the results) appears too short. This creates a sort of frustration, as the reader realizes the emulator performs well but is not always sure what physical behavior/process is actually well emulated. A couple of sentences, here and there, to remind the reader of the main conclusion of already cited studies (e.g. Kravitz et al., 2015; Andrews et al., 2010) would help.

We agree with this comment and have added both references to previous literature and a more detailed description of the results and their physical interpretation.

2. The paper lacks an introduction to EOFs! There is a quite lengthy explanation of what IRFs are and how they are obtained, but almost nothing about EOFs in the methods section. This should be re-balanced as EOFs are presented at the end of the paper. Maybe the part on IRFs could be shortened a little so as to avoid a too lengthy methods section.

Agreed; we have added section 2b in the methods to discuss EOFs; insofar as EOF analysis is standard (in contrast to IRFs) in climate science, this section is shorter. As the IRFs are crucial to the paper, we have not shortened.

3. The analysis of the performance of the emulators is limited to looking at some plots. It would be better to have at least a few quantitative metrics, to better understand the emulators' performance. Metrics could be provided in a table, both for the IRFs (timeseries) and EOFs (spatial patterns).

Agreed; we have added a table with calculation of the root-mean-square deviation between predicted and simulated results for each model and each variable, compared with the rms of climate variability.

4. This is more of a request, but it is maybe the most important point of my review. I believe the IRFs calculated by the authors should be provided as supplementary material. The paper would strongly benefit from it, as it would have much more impact on the modelers' community (and, therefore, it would be much more cited). This is especially true as the rationale behind the study is presented as being using those emulators in future studies of geoengineering scenarios. An Excel spreadsheet with one time-series per model and global variable should do it.

An excellent suggestion; we have included these in supplementary material.

Minor points:

l. 3: I suggest adding "further" to "without relying *further* on GCMs" and removing "for every possible pathway".

Good! Done.

l. 15: I find "be a more accurate estimate" than GCMs too strong. I would rather say "more cost-effective", especially as for GCMs the multi-model approach, as well as the multiple realizations, do compensate for the possible bias induced by natural variability. In the end, it is an issue of computation time requirement, not of accuracy.

This sentence has been deleted from the abstract, as putting in the appropriate caveats is too long for an abstract; we add the extra phrasing later in the text. We agree that if there was no computation time limit, then the GCM would be more accurate. However, given a finite amount of computation time, then it *is* possible that the emulator is indeed more accurate; this is too subtle for the abstract.

l. 21: I don't like the word "interpolation" here.

Agreed, changed.

l. 22: Change "fidelity" for something like: "spatial and temporal resolution".

The emulator in principle is capable of the same spatial and temporal resolution as the GCM, and even debating their relative accuracy as predictors of the forced-response would require a lengthy discussion. We haven't come up with a better word that is an accurate description of the advantages of the GCM.

l. 29: Define GeoMIP and explain briefly.

The definition and explanation are expanded towards the end of this section when the GeoMIP simulations are being explicitly referred to (this also benefits from your suggestion of moving figure S1 into the main text).

l. 38: Other variables such as precipitations are not always assumed to be strictly proportional to global mean temperature by simple models. E.g. some simple models use the relationship to GMT and RF by Andrews et al. (2010) for precipitations. Overall, I suggest being slightly less categorical.

Thanks for correcting our error; reworded.

l. 46-48: That sentence referring to Cao et al. (2015) should either be developed or removed. I found it incomprehensible.

Thanks – completely reworded to clarify.

l. 71: I suggest removing NPP of that study. See point below about figure S5.

Agreed, good point.

l. 93-94: I find that last sentence too brief: please develop.

We moved the important aspects of the sentence further down after the equations, where it is better motivated.

l. 113-114: I think a reminder that when the difference is done between these two simulations, you're assuming the system is linear.

Good call!

l. 144-150: The drawback of training over lower forcings would be a reduced domain of validity of the emulators, wouldn't it?

Not necessarily. Basically it's a trade-off between signal-to-noise ratio and nonlinearity; we reworded this paragraph to clarify.

l. 191-193: This drifting issue makes one wonder about the results of the study: : : Maybe this should be slightly expended. Can the drift be actually explained? How significant is it?

We are quite certain that the drift is due to initialization (that is, a few of the models were not fully spun up and continued to drift). If the starting conditions had been well documented, so that we could download the control run, this could be fixed; unfortunately this is not true and so we simply discarded those models where the drift was significant. (As a result, this is no longer an issue for this study.)

l. 202: Example of where one or two sentences could improve the paper. Explain/recall why there is a difference in the fast response.

Done!

l. 230-233: That sentence is a bit obscure. Is this a property of the IRFs or the GCMs? Develop.

Yeah, that was a pretty badly worded sentence. We clarified.

l. 234: Change "indicate" to "provide"?

Agreed!

l. 239-242: I honestly don't understand how the authors can claim that there is "no evidence of non-linearity". What would be the evidence? Do you mean that the nonlinearity is negligible, and therefore captured by the IRF?

Sentence is completely reworded in order to clarify. (If the nonlinearity was non-negligible, then the emulator based off a linear assumption and trained at one forcing

level would not have matched the response to a different forcing. We therefore conclude that nonlinearity is not too large, at least for these variables.)

l. 242-245: As in the abstract, I find this statement far too strong. It should be moderated. I would basically remove the sentence, unless actual proof can be provided: : :

The paragraph is reworded to better clarify exactly what we meant. Note that the observation is trivially true if indeed the model were perfectly linear; in general there is a trade-off between uncertainty introduced from natural variability, and errors introduced due to nonlinearity. Without conducting a large ensemble of G2 to provide a "truth" for the forced-response, it is not possible to separate these two errors and determine whether the emulated response actually does provide a better estimate of the forced response, or whether it is simply possible that it does (and again, the potential is trivially true, and that is the only statement made here, so we disagree that the statement is "too strong" though agree that it was badly worded!)

l. 256: Change "metrics" to "impacts"?

The word "impacts" is often associated with a specific meaning in some of the climate change literature, as the actual things humans care about; while there is often overlap with climate variables in a GCM, there might not be (e.g. vector-borne diseases), and so we consciously avoided the term "impacts".

l. 260: Again, moderate a little bit: more insight *on some aspects*. Maybe recall the computing-efficiency of the emulators. I believe this is definitely their most significant strength.

Agreed, changed.

l. 267: I fear the use of the word "moments" here may be confusing for the majority of the community. Maybe write "*statistical* moments", or expend or rephrase.

On reflection, including "statistical moments" was unnecessary to convey the point; we condensed to simply refer to extremes.

l. 281: It is always possible to develop emulators, except that they have to be nonlinear. So basically, the next step is to build box models with non-linear coefficients.

Reworded to clarify that one can always develop nonlinear emulators, although they would require either a priori assumptions or multiple forcing scenarios for training.

Fig.1: Check units. For this specific plot, the IRF units should be e.g. _C/[W/m2] I think. Check also units for precipitations.

Actually, not quite; the units here should be degrees C for a 4xCO2 (which is not the same radiative forcing in every model). We added this to the caption.

Fig.3: Units.

Thanks! (Oops.)

Fig.5: Needs a title over each map.

Thanks, done.

Fig. S1: Could be in main text.

Agreed, moved.

Fig. S5: Units are likely wrong. NPP should be tens of PgC/yr. But more importantly I suggest removing that plot on NPP. NPP is not a variable of the climate system stricto sensu, it is a variable of the carbon-cycle. NPP responds firstly to changes in atmospheric CO2, then to changes in climate and incoming radiation (at least in currentgeneration ESMs). The response to CO2 is strongly non-linear in intensity (can be captured with a simple log function, at global scale) and it is virtually instantaneous at the yearly time-scale. So here there is virtually no difference between the two simulations because NPP is basically responding to the annual atmospheric CO2. In short: the IRF approach is *not* the right approach for NPP: wrong driving variables, and wrong time-scale.

Agreed, NPP removed.

Manuscript prepared for Atmos. Chem. Phys.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 18 November 2016

~~Multi-model dynamic~~ Dynamic climate ~~emulator~~ emulators for solar geoengineering

Douglas G. MacMartin[1] and Ben Kravitz [2]

[1]Department of Mechanical and Aerospace Engineering, Cornell University, Ithaca NY, USA and Computing + Mathematical Sciences, California Institute of Technology, Pasadena CA, USA
[2]Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, WA, USA

Correspondence to: D. G. MacMartin (dgm224@cornell.edu)

Abstract. Climate emulators trained on existing simulations can be used to project the climate effects that would result from different possible future pathways of anthropogenic forcing, without relying further on general circulation model (GCM) simulations~~for every possible pathway~~. We extend this idea to include different amounts of solar geoengineering in addition to different pathways of greenhouse gas concentrations by training emulators from a multi-model ensemble of simulations from the Geoengineering Model Intercomparison Project (GeoMIP). The emulator is trained on the abrupt $4\times CO_2$ and a compensating solar reduction simulation (G1), and evaluated by comparing predictions against a simulated 1% per year CO_2 increase and a similarly smaller solar reduction (G2). We find reasonable agreement in most models for predicting changes in temperature and precipitation (including regional effects), and annual-mean Northern hemisphere sea ice extent, with the difference between simulation and prediction typically smaller than natural variability. This verifies that the linearity assumption used in constructing the emulator is sufficient for these variables over the range of forcing considered. Annual-minimum Northern hemisphere sea ice extent is less-well predicted, indicating ~~the limits~~ a limit of the linearity assumption. ~~For future pathways involving relatively small forcing from solar geoengineering, the errors introduced from nonlinear effects may be smaller than the uncertainty due to natural variability, and the emulator prediction may be a more accurate estimate of the forced component of the models' response than an actual simulation would be.~~

1 Introduction

Climate emulators have been used extensively to provide projections of climate changes for different anthropogenic forcing trajectories. These are trained based on a limited number of simulations with General Circulation Models (GCMs) and allow ~~interpolation~~ prediction of climate response for a much broader set of trajectories, trading the fidelity of a GCM simulation for computational efficiency. A similar approach could in principle be undertaken for projections of the

climate effects from solar geoengineering. Various solar geoengineering approaches have been suggested for intentionally influencing Earth's radiation budget, such as the injection of aerosols into the stratosphere (see, e.g., National Academy of Sciences, 2015). It is possible that such approaches may be considered in the future for reducing some amount of climate damages. However, any climate model simulation of geoengineering necessarily corresponds to some specific scenario, such as offsetting all of the global-mean-temperature change from other anthropogenic forcing, ~~(as in GeoMIP; Kravitz et al., 2011)~~ (as in GeoMIP; Kravitz et al., 2011, described in more detail below) . It is therefore useful to develop emulators that can use existing simulations in order to predict climate consequences both for different future trajectories of greenhouse gas forcing and for different possible choices regarding the level of geoengineering.

The simplest emulator approach is pattern scaling (Santer et al., 1990; Mitchell, 2003; Tebaldi and Arblaster, 2014), where a predictive dynamic model is used only for the time-evolution of the global mean temperature (either from energy balance approaches or estimated directly from GCM simulations), and the temperature at every spatial location is assumed to vary with the same time evolution as the global mean – that is, that the pattern of temperature change is not itself a function of time. Other variables, such as precipitation changes, are also assumed to ~~scale with~~ depend only on the global mean temperature ~~, so that~~ and on radiative forcing (Andrews et al., 2010) ; the only "memory" in the emulator ~~is~~ in this case remains embedded in the dynamics of the global mean temperature response. Extending this, Cao et al. (2015) assume precipitation depends on global-mean-temperature and not just instantaneous CO_2 concentrations but also solar reduction, allowing for a different "fast" response to these different forcings, but again maintaining global mean temperature as the sole dynamic predictor. Additional spatial patterns can also be included to capture other forcing agents including aerosols (Schlesinger et al., 2000; Frieler et al., 2012) .

Of course, not all of the climate system responds to forcing with the same time-constants. Pattern scaling can be improved upon by introducing additional dynamic variables, such as land-sea temperature contrast (Joshi et al., 2013), multiple empirical orthogonal functions (EOFs) of temperature (Holden and Edwards, 2010; Herger et al., 2015), or by including many more spatial degrees of freedom to better predict regional effects (Castruccio et al., 2014). ~~Additional spatial patterns can also be included to capture other forcing agents including aerosols (Schlesinger et al., 2000; Frieler et al., 2012) . Cao et al. (2015) include the climate response to a solar reduction in a dynamic emulator, with global-mean-temperature as the sole dynamic predictor (in addition to instantaneous forcing).~~ The use of only one or a few dynamic variables (or predictors) is ultimately constrained by the difficulty in estimating the dynamic response of additional variables in the presence of climate variability due to low signal-to-noise ratio.

The primary assumption typically made in developing a climate emulator for predicting climate response is that the ~~climate~~ response is sufficiently linear and time-invariant (LTI). (We are explicit about our usage of the terms *linear* and *non-linear* in Section 2 below.) Success with emulators

illustrates that linearity can be a reasonable approximation, although the accuracy of this assumption will depend on the variable and the level of applied forcing (e.g., Tebaldi and Arblaster, 2014). The response of any LTI system to any time-varying forcing can be described by a convolution between the impulse response function that describes the system dynamics and the exogenous forcing; see equation (1) in Section 2 below, and also Åström and Murray (2008, Sec. 5.3) or Ragone et al. (2015, eq. 2). "Training" ~~the~~ a linear emulator amounts to estimating the impulse response from one or more simulations. Nonlinear approaches to emulators are used in other aspects of climate modeling, such as model tuning and parametric uncertainty analysis (Neelin et al., 2010) , but such investigations are beyond the scope of this manuscript.

We start from the same LTI assumption here as in the references above, but extended to include solar geoengineering. The spatial patterns of the responses to solar and greenhouse gas forcing will not be the same, leading to regional differences in outcomes (Ricke et al., 2010; Kravitz et al., 2014, 2015), nor are the precipitation responses the same ~~(Bala et al., 2010)~~ (Bala et al., 2010; Andrews et al., 2010) , nor necessarily the time-evolution of the responses (Cao et al., 2015). All of these factors are important to capture if the emulator is to be useful in understanding climate effects of strategies that include solar geoengineering. We therefore only make an LTI assumption, and do not start with any additional *a priori* assumptions on the form of the dynamics~~, and begin by considering~~ . We thus consider independent predictors for each variable. For estimating the spatial temperature and precipitation response, we employ an EOF-based approach (as in Herger et al., 2015) with a common set of EOFs constructed from both CO_2-forced and geoengineering simulations. In addition to temperature and precipitation, we also consider Northern hemisphere sea-ice extent~~(see Supplementary Material for net primary productivity; NPP)~~; the minimum extent over the year provides an example where linearity is not a good assumption.

We use simulations from the Geoengineering Model Intercomparison Study (GeoMIP, Kravitz et al., 2011) where solar reduction is used as a proxy for any approach that reduces incoming shortwave radiation. ~~By training the emulator on one set of simulations and validating on a second , we can evaluate the fundamental assumption of linearity~~ Linearity and time-invariance are the *only* assumptions we make in developing the emulator. The emulator can therefore be uniquely specified based on a single simulation for each model. The assumption of linearity can then be evaluated by comparing predictions with a second simulation for a different forcing trajectory; deviations between these result from nonlinearity, and conversely, agreement validates linearity being a reasonable approximation. Section 2 describes the methodology and simulations used, and the resulting emulator and validation are given in Section 3.

2 Approach

The expectation that an emulator calibrated to match the GCM response to one climate forcing pathway can also do so for a different pathway is typically based on the assumption that the response to forcing can be reasonably approximated as linear and time-invariant (LTI). Consider a system forced by both time-dependent ~~forcing~~ forcing $f(t)$ from changes in atmospheric greenhouse gas concentrations and time-dependent forcing $g(t)$ from solar ~~geoengineering~~geoengineering. For any variable $z_i(t)$, define $z_i^f(t)$ as the response to forcing $f(t)$ with $g(t) = 0$ and $z_i^g(t)$ as the response to forcing $g(t)$ with $f(t) = 0$, where the response is defined in each case as the difference relative to the initial state, and neglecting natural variability. The system is *linear* if for any scalars α and β, the response to the combined forcing $\alpha f(t) + \beta g(t)$ is the same linear combination of the individual responses, $\alpha z_i^f(t) + \beta z_i^g(t)$. Note that in general, even if the system is linear, the ratio of any two variables will vary with time simply because different variables respond at different rates; that is, for any forcing scenario, there is not in general some constant μ such that $z_i(t) = \mu z_j(t)$ for all time (a plot of $z_i(t)$ against $z_j(t)$ will not be a straight line if these variables respond with different time constants). The usage of the word nonlinear to express this latter idea is distinct from the concept of the dynamic system itself being linear or nonlinear. By a dynamic system we simply mean that $z(t)$ depends on past values of the forcing $f(t)$ or $g(t)$ in addition to the current values. ~~A linear system can be characterized purely by its~~

The climate system as a whole is highly nonlinear. However, the response to a ~~Dirac delta function; this is the impulse response~~perturbation about the current state may be close to linear; if the perturbation is sufficiently small then linearity will be a good approximation.

2.1 Impulse Response

For an LTI system forced by both time-dependent $f(t)$ and $g(t)$, the response of any variable $z_i(t)$ can be expressed in terms of a convolution between the input time-series and the system impulse response functions as

$$z_i(t) = \int_0^t h_i^f(\tau) f(t-\tau) d\tau + \int_0^t h_i^g(\tau) g(t-\tau) d\tau + n_i(t) \quad (1)$$

where $h_i^f(t)$ is the impulse response due to greenhouse gas forcing and $h_i^g(t)$ the impulse response due to solar reductions; these will not in general be identical, nor in general the same for any choice of output variable z_i. The variable $n_i(t)$ is included to ~~denote~~ capture the effects of climate variability. Because the emulator is designed to capture the forced response, the actual character of $n_i(t)$ is unimportant in defining the emulator. The system is time-invariant if ~~$h(\tau)$~~ $h_i^f(\tau)$ and $h_i^g(\tau)$ in equation (1) ~~does~~ do not depend explicitly on the current time t; some possible exceptions are noted in Section 4. ~~Herein we consider only annual-mean variables. The same formalism as in equation (1) also applies for predicting the seasonal-dependence of the~~ Note that the response of a linear ~~(but~~

~~time-periodic) system , where~~ system can be completely characterized by the impulse response~~in general also depends on the time of year (e. g., $h = h(\tau, m)$ for month m) and~~ ; knowing the impulse response is thus sufficient to predict the response to any forcing trajectory. The same formalism would also apply for predicting the response to seasonally-dependent forcing, but of course additional training simulations would be required~~to estimate the seasonal-dependence of h.~~ .

If the climate system were indeed LTI, then equation (1) would hold for any variable (temperature, precipitation, etc~~) either at any one location or projected onto any particular spatial pattern.~~ .), at global or regional scale, and whether annual-mean or at a shorter time-scale, although the degree to which the forced response can be estimated in the presence of natural variability will vary with spatial and temporal scale, as will the influence of nonlinearities. We consider variables evaluated once per year (e.g., annual-mean, or September sea ice extent), and equation (1) can be cast in discrete-time to predict the response in year k as

$$z_i(k) = \sum_{j=0}^{k} h_i^f(j) f(k-j) + \sum_{j=0}^{k} h_i^g(j) g(k-j) + n_i(k) \tag{2}$$

To estimate the impulse response for CO_2 forcing, we use the difference between the abrupt $4\times CO_2$ simulation and pre-industrial simulation for each of the models participating in GeoMIP. To estimate the impulse response for solar reduction, we use the G1 simulation from GeoMIP, in which the CO_2 concentration was quadrupled and insolation decreased to approximately maintain radiative balance and hence global mean temperature (see Figure ~~S1~~1). The difference between G1 and the $4\times CO_2$ simulations thus gives the response to an abrupt change in solar forcing, assuming linearity. Note that each model separately chose the level of solar reduction $g_{4\times}$ required to balance the forcing from increased atmospheric CO_2, so that the percent solar reduction in G1 varies from model to model based on the efficacy of solar forcing in that model; see Table S1. Define

$$f(t) = \log_2\left(\frac{\mathrm{CO_2}(t)}{\mathrm{CO_{2,ref}}}\right) \div 2 \tag{3}$$

$$g(t) = -\left(\frac{\mathrm{Solar}(t) - \mathrm{Solar_{ref}}}{\mathrm{Solar_{ref}}}\right) \div g_{4\times} \tag{4}$$

where $\mathrm{CO_2}(t)$ is the time-varying atmospheric CO_2 concentration and $\mathrm{Solar}(t)$ is the solar irradiance. The $4\times CO_2$ experiment then corresponds to forcing $f(t) = 1$, $t \geq 0$ and $f(t) = 0$, $t < 0$ with $g(t) = 0$, while the GeoMIP G1 simulation uses the same $f(t)$ but with $g(t) = 1$, $t \geq 0$.

Substituting into Equation (~~1~~2) then for any variable ~~$z_i(t)$~~ $z_i(k)$, the difference ~~$z_i^{4\times}(t)$~~ $z_i^{4\times}(k)$ between its value in $4\times CO_2$ and preindustrial is given by

$$z_i^{4\times}(tk) = \int_0^t \sum_{j=0}^{k} h_i^f(\tau j)\, d\tau + n_i(tk) \tag{5}$$

and the difference $z_i^{G1}(t)$ between its value in G1 relative to 4×CO_2 is

$$z_i^{G1}(tk) = \int_0^t \sum_{j=0}^{k} h_i^g(\tau j) d\tau + n_i(tk) \tag{6}$$

from which we can estimate

$$\hat{h}_i^f(tk) = \frac{d}{dt} z_i^{4\times}(tk) - z_i^{4\times}(k-1) \qquad \text{and} \qquad \hat{h}_i^g(tk) = \frac{d}{dt} z_i^{G1}(tk) - z_i^{G1}(k-1) \tag{7}$$

The impulse responses ~~$h_i^{f,g}(t)$~~ $h_i^{f,g}(k)$ could be estimated from the time-series of any forced simulation, but take particularly simple form from these step response simulations. (A linearly increasing forcing scenario such as a 1% per year increase in CO_2 also leads to a simple form, with the continuous-time impulse response proportional to the second derivative of the 1%CO_2 response.)

These impulse response estimates are "noisy" due to natural variability. Various approaches could be used to reduce the influence of natural variability, such as

1. Using multiple ensemble members or multiple forcing scenarios (as in Castruccio et al., 2014, for example),

2. Only considering spatial averages by computing the global mean as in pattern scaling, projecting onto EOFs as in Herger et al. (2015), or averaging over specific spatial regions as in Castruccio et al. (2014),

3. Applying temporal filtering to smooth high-frequency noise in $\hat{h}$ or fitting $h(t)$ to some estimated functional form such as semi-infinite diffusion for global mean temperature (Caldeira and Myhrvold, 2013) or a multiple-exponential (Castruccio et al., 2014) or

4. Finding some less-noisy predictive variable, such as global mean temperature, to use as the predictor of other, noisier variables (effectively what is done in predicting the regional precipitation or temperature response in any pattern scaling analysis).

Choosing simulations with high forcing levels to train the emulator (4×CO_2 and GeoMIP G1) increases the "signal" of the forced-response relative to the "noise" of climate variability. This choice allows us to make useful predictions at lower forcing levels without the need for introducing ~~arbitrary a priori~~ additional assumptions on the functional form of the dynamics, such as that every field simply scales with global mean temperature. The penalty for this choice is that the high forcing will exacerbate any nonlinear effects; this choice precludes, for example, useful predictions of the Northern hemisphere annual-minimum sea ice extent (see Section 3 below), which would require that a lower-forcing simulation be used to train the emulator.

A frequency-domain perspective is useful to understand how the "noise" due to climate variability affects the emulator predictions. The Laplace transform of Equation (1) transforms the convolution

into multiplication:

$$\mathcal{L}(z_i) = \mathcal{L}(h_i^f)\mathcal{L}(f) + \mathcal{L}(h_i^g)\mathcal{L}(g) + \mathcal{L}(n_i) \tag{8}$$

$$= H_i^f(s)F(s) + H_i^g(s)G(s) + N(s) \tag{9}$$

where the Laplace transform of the impulse response, $H_i(s) = \mathcal{L}(h_i)$, is the *transfer function* between that input and that output; capital letters will denote the Laplace transform of $h(t)$, $f(t)$ and $g(t)$. ~~The~~ (The discrete-time formalism in equation (2) could similarly be analyzed with a Z-transform; we use the continuous-time formulation here as readers are more likely to be familiar with it.) The impulse response could thus equivalently be estimated by first taking the Laplace transform of the input and output, computing the ratio, and computing the inverse transform. Consider for example the response to increased CO_2 (the estimation for solar reduction is analogous), where the emulator is trained on the input ~~$f_t(t)$~~ $f_e(t)$ and used to predict the response to ~~forcing~~ a different forcing time-series $f_p(t)$, with Laplace transforms ~~$F_t(s)$~~ $F_e(s)$ and $F_p(s)$. The transfer function estimate used by the emulator is

$$\hat{H}_i^g(s) = H_i^f(s) + \frac{N(s)}{F_t(s)}\frac{N(s)}{F_e(s)} \tag{10}$$

and hence in the frequency domain the response predicted by the emulator for input forcing $F_p(s)$ is

$$\hat{Z}_i = Z_i(s) + N(s)\frac{F_p(s)}{F_t(s)}\frac{F_p(s)}{F_e(s)} \tag{11}$$

That is, climate variability in the simulation used to train the emulator leads to an error in the prediction that depends on the ratio of frequency content in the forcing signals between training and prediction simulations. Because a "step" change in the input such as in the abrupt $4\times CO_2$ simulation has more signal energy at low frequencies than high (Laplace transform proportional to $1/s$), it leads to a better estimate of the output response at low frequencies than at high frequencies; the high-frequency estimation errors due to natural variability manifest as "noise" on the estimated impulse response (see Figure 2 for an example). However, the smoothly varying radiative forcing input due to a 1% per year increase in CO_2 has even less energy at high temporal frequencies than the step input (Laplace transform proportional to $1/s^2$). Thus training an emulator on a "step" input simulation and then using it to predict the results from a smoothly-varying forcing trajectory will result in relatively noise-free emulator predictions, despite the apparent high-frequency "noise" in the impulse response. Note that the GeoMIP G2 simulation (described at the beginning of the next section) has an abrupt change in the solar forcing at year 50 (see Figure ~~S1~~1), and the emulated responses to this "step" change in forcing are, as expected, noisier than those due to the smooth forcing changes over the first 50 years of G2.

2.2 Spatial analysis

For predicting the spatial pattern of the forced response, we estimate impulse responses not for every individual grid cell in each GCM, but only for the spatial response projected onto the first few empirical orthogonal functions (EOFs). For each model, EOFs are constructed from the area-weighted spatial temperature and (separately) the precipitation response. For each variable, and for each model, a single set of EOFs is constructed using output from both the 4×CO_2 and G1 simulations, leading to a description of the form

$$T(x,y,t) = \sum_{i=1}^{m} \Phi_i(x,y)\psi_i(t) \tag{12}$$

where Φ_i are the spatial basis functions (EOFs) and ψ_i the corresponding principal components (projection of $T(x,y,t)$ onto each Φ_i for any particular forcing scenario); the basis set Φ_i are thus unchanged across the different forcing mechanisms and temporal trajectories. Truncating the set of EOFs provides a maximally efficient basis for describing the spatial pattern of the response, capturing any pattern strongly excited by either one or both forcing mechanisms. In general, only the first few principal components are distinguishable from climate variability and have any predictive capability (Figure S4) and we retain $m = 4$ throughout. The first pattern, corresponding to the highest variance in the simulations, is similar to the long-term pattern of global warming; choosing $m = 1$ would thus be analogous to pattern scaling. Including additional EOFs captures both the differences in how the climate responds to solar versus CO_2 forcing, as well as differences between the short- and long-term pattern of response for either forcing (i.e., that not everything responds at the same rate). Temperature EOFs for one model are shown in Supplementary Material Figure S1, where the second EOF captures the equator-to-pole differential warming that is a robust signature of compensating a CO_2-induced global mean temperature rise with a solar reduction, while EOFs 3 and 4 capture Northern hemisphere and global patterns of land temperature, which change more rapidly than ocean temperatures in response to forcing.

The impulse responses can then be separately estimated for each principal component as before from the 4×CO_2 and G1 simulations, and the time series of ψ_i for any other forcing scenario estimated. Equation 12 is then used to construct the estimate of the spatial response.

3 Results and validation

The impulse responses $h_i^f(t)$ and $h_i^g(t)$ are estimated for a number of different variables from the abrupt 4×CO_2 and G1 simulations as described above. The impulse-response based emulator for CO_2 forcing without any solar reduction can be validated by comparing the predictions with the simulations for a 1% per year increase in CO_2 (1%CO_2). To validate the emulation of solar reduction, we use the GeoMIP G2 scenario, in which CO_2 levels increase at 1% per year, and for the first

50 years, the solar reduction is gradually increased to balance this forcing. This uses the same ratio of $g(t)$ to $f(t)$ as in G1 for each model. After 50 years, the solar reduction is returned to zero so that only the radiative forcing from the CO_2 remains (see Kravitz et al. (2011) and ~~Supplementary Figure S1~~ Figure 1 for a schematic of the forcing in the G1 and G2 simulations). Several of the climate models that conducted experiments G1 and G2 exhibit significant drift in the absence of net radiative forcing, ~~presumably due to initialization issues~~due to the initialization state not being in equilibrium. These models are not considered further, leading to a total of 9 models considered here (Table S1).

The impulse response functions for predicting the global mean temperature and precipitation responses to either CO_2 or solar forcing are shown in Figure 2, averaged over all of these climate models (see Supplementary Material for tabulation of these and other impulse responses for each model). As expected these are "noisy" estimates due to natural variability. Note that while the temperature response characteristics are similar (aside from the sign) for increased CO_2 and reduced insolation, the precipitation response differs. The impulse response of precipitation clearly highlights that while CO_2 and solar reduction have a similar "slow" response (changes in precipitation that result from changes in temperature), they have quite different "fast" responses (rapid atmospheric adjustments in the climate system before temperature has time to adjust). The fast response is related to different amounts of radiative forcing absorbed by the atmosphere ~~(e.g., Andrews et al., 2010) .~~ that affect stability and convection (e.g., Andrews et al., 2010) . For CO_2-forcing this leads to an initial precipitation response of the opposite sign to the long-term slow response; while solar reductions might largely compensate for the slow response there will be residual differences due to the differential fast response. Comparing impulse response functions between models may also be useful to identify differences in dynamics (Figure S1).

Figure 3 validates the ability of the impulse response formulation in equation (1) tuned from the $4\times CO_2$ and G1 simulations to correctly predict the global mean temperature response from the $1\%CO_2$ and G2 simulations. ~~Figure 4 shows the corresponding plots for global mean precipitation. Similar results are shown in the supplementary material (Figures S2–S3) for the temperature or precipitation difference between land and ocean.~~ Linearity has previously been argued as a reasonable assumption for temperature and precipitation responses (Kravitz et al., 2014, and references therein)~~and since~~ . Since that is the only assumption made in constructing the emulator, the error in estimating the forced response arises only from natural variability and from nonlinearity. The difference between GCM-simulated and emulator-predicted trajectories is similar to the standard deviation of natural variability in many models; see Table 1. Cases where the predicted and simulated responses agree to within the limit imposed by natural variability indicates that nonlinear effects are small relative to variability, and hence this analysis also ~~validates that assumption for these variables and~~ illustrates the utility of a linearity assumption at these forcing levels. ~~Note that the~~

~~difference between GCM-simulated and emulator-predicted trajectories is typically less than the standard deviation of natural variability.~~

Figure 4 shows the corresponding plots for global mean precipitation. The deviation between emulated and simulated responses are higher for some models here than for temperature, though the estimation errors are close to the limit due to natural variability for many models. Note that since G2 suppresses global mean temperature changes, it largely suppresses the slow (temperature-dependent) precipitation response (there will still be some effect from regional temperature changes). This suggests that in models such as GISS-E2-R, HadCM3, or MIROC-ESM where the G2 emulation is notably better than the emulation of the 1% CO_2 simulation, larger nonlinearities in the precipitation response arise in the slow rather than fast response to precipitation.

Similar results are shown in the supplementary material (Figures S2–S3) for the temperature or precipitation difference between land and ocean; the only notable case where the error from nonlinearity exceeds natural variability is in the GISS-E2-R prediction of land-sea precipitation differences in the 1% CO_2 simulation. While it is not our purpose to evaluate mechanisms of nonlinearity in the climate models, this type of analysis may be useful input into such research.

Northern hemisphere sea ice extent is an example of a variable that is both highly relevant for assessing possible future scenarios, yet one in which a nonlinear response to forcing might be expected. The $4\times CO_2$ forcing is large enough that September sea ice is nearly lost in all models, and thus an emulator trained off of this simulation will do a relatively poor job at predicting the reduction in annual-minimum sea ice extent from smaller forcing; see Figure 5. However, despite the obvious nonlinearity in the annual-minimum extent, the annual-mean sea ice extent does behave sufficiently linearly in most models, even at this large a forcing level, so that the $4\times CO_2$ simulation can be used to train a useful emulator. This is illustrated in Figure 6.

Finally, Figure 7 illustrates the ability to capture the spatial response. One of the concerns raised regarding the use of solar geoengineering is that the response from ~~turning down the sun~~ solar reduction does not perfectly compensate that from increased CO_2, resulting in some regional differences in temperature and precipitation responses (Ricke et al., 2010; Kravitz et al., 2014). It is therefore valuable to assess whether the emulator can capture some of the regional variation in the response between CO_2 and solar forcing. ~~For each model, EOFs are constructed from the spatial temperature response using both $4\times CO_2$ and G1 simulations to construct a single set of combined EOFs. In general, only the~~ As described earlier, the regional response is predicted using EOF analysis and estimating the forced-response for the first few principal components~~are distinguishable from climate variability and have any predictive capability (Figure S4)~~. Figure 7 plots the model-mean temperature and precipitation responses averaged over years 41-50 of the G2 simulation for both the simulation and the emulator prediction. ~~Note that only after averaging over 10years and 9 climate models are the temperature and precipitation changes due to~~ The G2 ~~– which is designed~~

~~to have near-zero top-of-atmosphere radiative forcing – statistically significant at the grid-cell level. The differences here between the simulated and emulated responses are not significant at the grid cell level, although greater spatial averaging may indicate some statistically significant regional differences. In particular, the emulator appears to slightly underpredict the amount of~~ simulation, like G1, results in overcooling of the tropics and undercooling of the poles. The emulator slightly underpredicts the residual Arctic warming in G2, likely due to the nonlinearity associated with sea ice albedo feedback at the $4\times CO_2$ forcing used in training the emulator. ~~Beyond this feature, it is difficult to assess with certainty to what extent the differences between the simulated and emulated regional responses are due to nonlinearities or simply due to natural variability. There is no evidence of nonlinearity in either the ability of the emulator to capture differences between land and ocean temperatures or precipitation (Figure S2 and S3), nor in capturing the first few principal components of the response (Figure S4).~~ The area-weighted spatial root mean square (rms) of the difference between emulated and simulated responses is also shown in Table 1, normalized at each grid cell by the standard deviation of interannual climate variability. Where the rms value is close to unity implies that the errors introduced by assuming linearity are not limiting the emulator predictions; the Arctic nonlinearity contributes to the larger rms errors in temperature prediction for many models.

This raises an interesting observation. If it is purely the forced-response that is of interest, then a single GCM simulation of a low-forcing scenario such as G2 leads to uncertainty in the estimate due to natural variability. While the most accurate estimate would be obtained by averaging over a sufficiently large ensemble, this may not be achievable for computational reasons. The emulator provides a computationally-efficient alternative. Because the emulated response is based on simulations with roughly three times higher radiative forcing, and because the process of its construction suppresses high-frequency natural variability ~~, it is potentially a~~ (equation 11), the estimate of the forced-response that it provides has less uncertainty due to natural variability, at the cost of increased errors from nonlinearity. It is thus possible that, given only sufficient computation to conduct a single simulation, the emulated response based off of G1 could be a more accurate representation of the forced-response to G2 ~~in the models~~ than that obtained from the actual G2 simulation. This is trivially true if indeed the response was perfectly linear; in general there is a trade-off between errors due to nonlinear effects and the uncertainty introduced by variability.

4 Discussion

Climate emulators provide a powerful tool for assessing any proposed future pathway of mitigation choices (including carbon dioxide removal) and different levels of geoengineering. For example, solar geoengineering could be used only to limit peak warming as part of an "overshoot" scenario in which atmospheric CO_2 concentrations peak and subsequently decline as net-negative carbon emissions reduce concentrations (Long and Shepherd, 2014; Tilmes et al., 2016). A limited, temporary

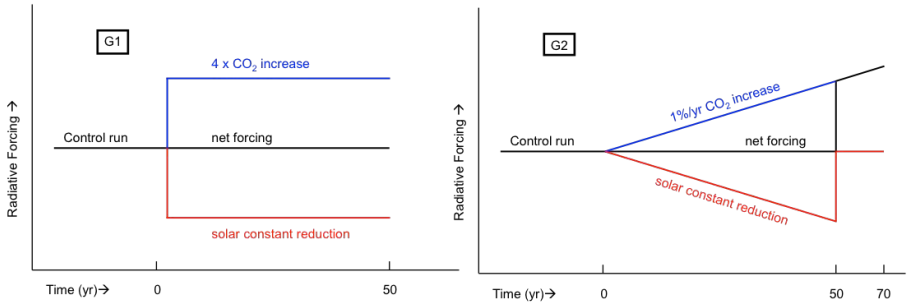

Figure 1. Schematic of GeoMIP G1 and G2 simulations, from Kravitz et al. (2011) .

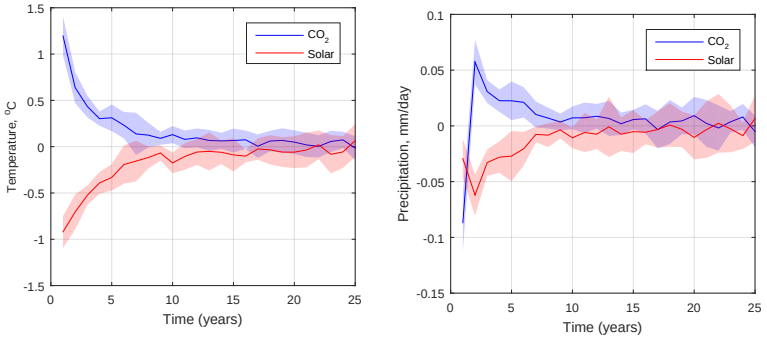

Figure 2. Estimated impulse response for CO_2 and solar forcing, for global mean temperature and precipitation, averaged over all 9 models (table S1); the inter-model standard deviation is shown by the shaded bands. While these impulse response functions are "noisy", predictions made using them are less so, particularly for forcing levels much smaller than those used in estimating these functions. Note for precipitation the robust "fast" response to increased CO_2 has the opposite sign as the "slow" response. Temperature and precipitation units are given as the response for a quadrupling of CO_2. (See Supplementary Material including Figure S1 for individual model impulse respones functions.)

deployment has also been described as a way to reduce the rate of warming (Keith and MacMartin, 2015; MacMartin et al., 2014). These types of limited-deployment scenarios are motivated in part by recognizing that solar geoengineering sufficient to reduce global mean temperature to preindustrial levels could lead to significant regional disparities and other risks, while a deployment that only partially reduces global mean temperature might decrease some metrics of climate change everywhere (Kravitz et al., 2014).

By training emulators on a standard set of simulations, such as GeoMIP, that have been conducted by multiple modeling centers, any ~~future~~ proposed scenario such as these can be readily evaluated with multiple models. This ~~can provide more~~ yields a computationally-efficient method for providing insight into the robustness of conclusions~~than detailed simulations with any single model~~. (Of course, any collection of models is an ensemble of opportunity, with interpretation challenges as a statisti-

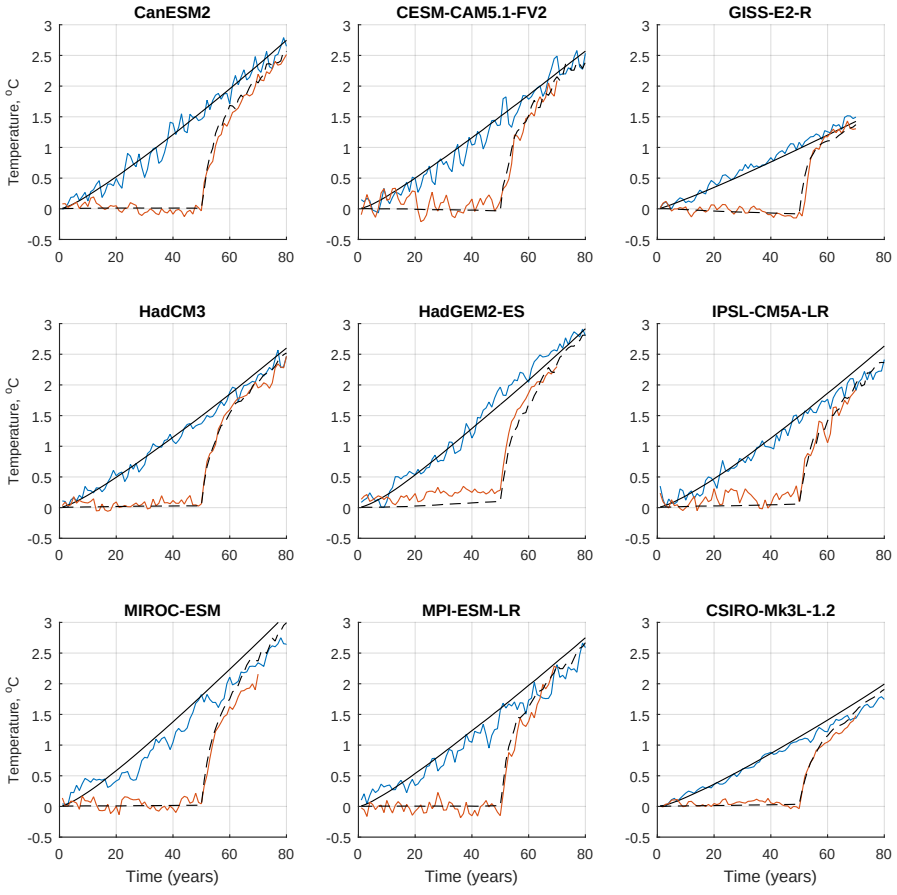

Figure 3. Simulated and predicted global mean temperature, both for a 1% per year increase in CO_2 (blue curves) and for GeoMIP experiment G2 (red), for each of the climate models considered here. The predicted response using the emulator is given by black lines, solid for the 1% CO_2 case and dashed for G2.

cal sample; see, e.g., Collins et al. (2013), Section 12.2, for a thorough discussion.) The emulator used here assumes that the climate system response can be sufficiently well approximated over the range of forcing levels of interest by the output of a linear system. For many variables, the analysis here indicates that this is a sufficiently good assumption, with the difference between simulated and emulated responses ~~smaller than~~ similar to the standard deviation of natural variability. There are many more variables that may be of interest~~, including higher moments to predict extremes~~; similar analysis as here could be used to assess whether a linear assumption is or is not sufficient for projecting the response of any variable beyond those considered here. The GeoMIP simulations are also of limited duration, and nonlinearities may arise at longer time-scales due to changes in ocean dynamics, for example (Bouttes et al., 2015) .

Finally, note that the results herein were obtained using simulations that reduce the solar constant as a proxy for any solar geoengineering approach. ~~The~~ While this is clearly a useful first step, the

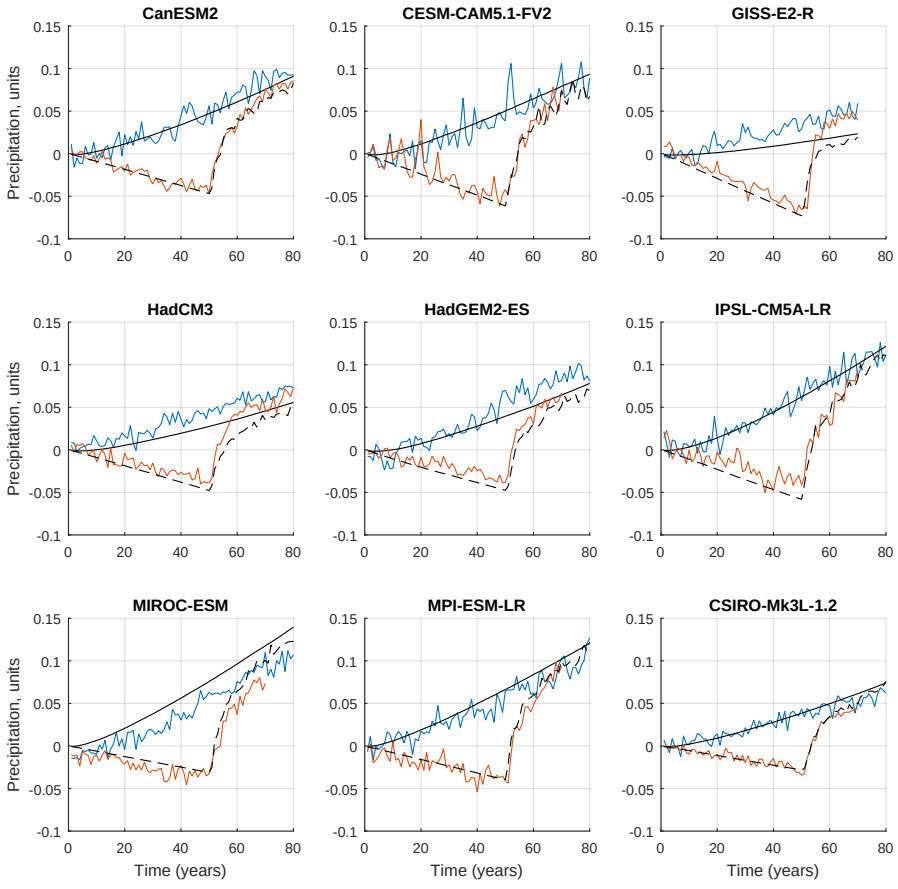

Figure 4. As in Figure 3 but for global mean precipitation. Simulated and emulated response are shown for 1% per year increase in CO_2 and GeoMIP experiment G2 for each of the climate models considered here.

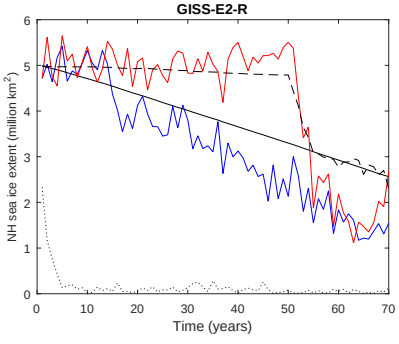

Figure 5. As in Figure 3 but for Northern Hemisphere annual-minimum sea ice extent. Simulated and emulated response are shown for 1% per year increase in CO_2 and GeoMIP experiment G2 for one model, GISS E2-R. The dotted line shows the response for the abrupt $4\times CO_2$ simulation. The relatively poorer emulator prediction for the 1% CO_2 case in particular illustrates that the linearity assumption does not hold for all relevant climate variables.

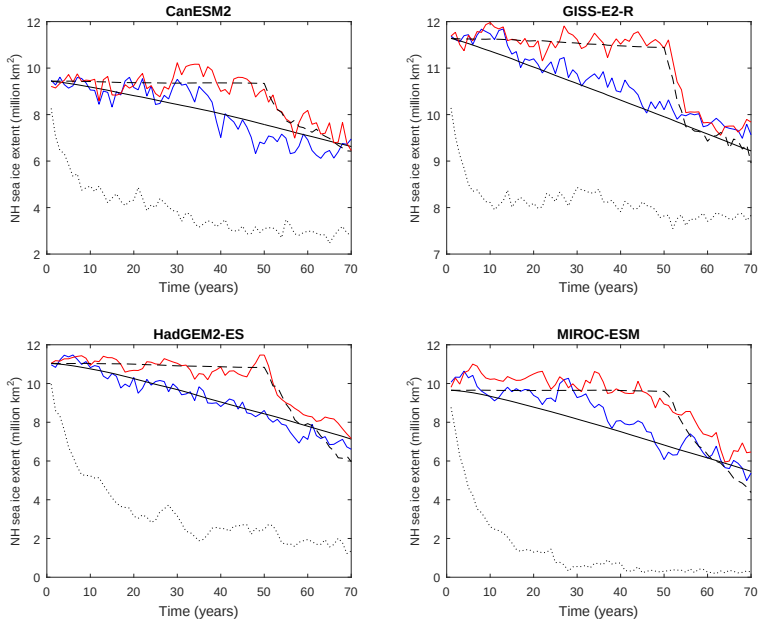

Figure 6. As in Figure 3 but for Northern Hemisphere annual-mean sea ice extent. Simulated and emulated response are shown for 1% per year increase in CO_2 and GeoMIP experiment G2 for several of the climate models considered here; the dotted line shows the response for the abrupt $4\times CO_2$ simulation.

Figure 7. Temperature (left) and precipitation (right) averaged over years 41-50 of G2 simulation and averaged over all 9 models. The upper row shows the simulated results; the lower row shows the prediction based on a spatial emulator developed using 4 EOFs for each model. As noted elsewhere, the robust response to increasing CO_2 and reducing insolation to maintain zero global mean temperature difference is a net reduction (overcompensation) of global mean precipitation (Bala et al., 2010), and an overcooling of the tropics and an undercooling of the poles (Kravitz et al., 2013). The latter is an artifact of a latitudinally-uniform reduction in sunlight, and could be better managed by increasing the forcing at high latitudes relative to low (Kravitz et al., 2016).

	Global-mean Temperature		Global-mean Precipitation		Annual mean NH sea ice		Spatial rms Temperature		Spatial rms Precipitation	
Model	1%CO_2	G2	1%CO_2	G2	1%CO_2	G2	1%CO_2	G2	1%CO_2	G2
CanESM2	1.0	0.5	1.3	0.5	1.8	1.3	1.4	1.2	1.0	0.7
CESM-CAM5.1-FV	1.4	1.0	1.0	0.7	-	-	1.8	1.2	1.5	1.2
GISS-E2-R	1.1	1.2	3.0	2.0	1.6	1.4	2.2	1.8	2.2	1.3
HadCM3	1.3	1.3	3.2	1.3	-	-	2.2	1.9	1.3	1.2
HadGEM2-ES	1.4	1.7	2.7	1.3	1.0	1.7	2.4	1.7	1.0	0.8
IPSL-CM5A-LR	1.2	1.9	1.2	1.7	-	-	2.6	2.0	2.0	2.1
MIROC-ESM	1.8	0.7	1.3	1.1	2.2	1.3	4.0	1.5	2.4	1.4
MPI-ESM-LR	1.7	0.8	1.3	0.9	-	-	2.6	1.2	1.5	1.1
CSIRO-Mk3L-1.2	1.4	1.7	1.2	0.7	-	-	4.1	1.3	3.7	0.8

Table 1. Root-mean-square (rms) deviation between simulation and emulator prediction. For first three (scalar) variables, temporal rms is computed over years 31–50, normalized by the standard deviation of interannual natural variability. For spatial response, the area-weighted rms is computed after normalizing by variability at each grid cell (that is, the spatial rms of the deviation as measured in standard deviations of natural variability).

climate effects from any specific technology, such as stratospheric aerosol injection (SAI) or marine cloud brightening (MCB) will differ (e.g., Ferraro et al., 2015) both due to the different mechanism of radiative forcing, and the different spatial pattern of radiative forcing (the latter being at least partially a design choice; Kravitz et al., 2016). Further, while linearity appears to be a reasonable assumption in these climate models for predicting the response of many climate variables to an imposed solar reduction, it may be a poorer approximation for SAI, for example. Nonlinearities will occur in aerosol size distribution (Heckendorn et al., 2009; Niemeier and Timmreck, 2015), as well as due to changes in the stratospheric circulation that result from the aerosols (Aquila et al., 2014); time-invariance might also not hold if, for example, time-varying stratospheric chlorine concentrations (which affects the aerosol impact on ozone) are considered part of the "system" rather than a forcing. It is unclear how significantly these will affect the ability to develop emulators for this technology.

Author contributions. DGM and BK designed the study, conducted the analysis, and wrote the paper.

Acknowledgements. We thank all participants of the Geoengineering Model Intercomparison Project and their model development teams, CLIVAR/WCRP Working Group on Coupled Modeling for endorsing GeoMIP and the scientists managing the Earth System Grid data nodes who have assisted with making GeoMIP output available. The Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC05-76RL01830. This work was partially supported by Cornell University's David R. Atkinson Center for a Sustainable Future (ACSF).

References

Andrews, T., Forster, P. M., Boucher, O., Bellouin, N., and Jones, A.: Precipitation, radiative forcing and global temperature change, Geophys. Res. Lett., 37, 2010.

Aquila, V., Garfinkel, C. I., Newman, P. A., Oman, L. D., and Waugh, D. W.: Modifications of the quasi-biennial oscillation by a geoengineering perturbation of the stratospheric aerosol layer, Geophys. Res. Lett., 41, 2014.

Åström, K. J. and Murray, R. M.: Analysis and Design of Feedback Systems, Princeton, 2008.

Bala, G., Caldeira, K., and Nemani, R.: Fast versus slow response in climate change: implications for the global hydrological cycle, Clim. Dyn., 35, 423–434, 2010.

Bouttes, N., Good, P., Gregory, J. M., and Lowe, J. A.: Nonlinearity of ocean heat uptake during warming and cooling in the FAMOUS climate model, Geophysical Research Letters, 42, 2409–2416, 2014GL062807, 2015.

Caldeira, K. and Myhrvold, N.: Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration, Env. Res. Lett., 8, 2013.

Cao, L., Bala, G., Zheng, M., and Caldeira, K.: Fast and slow climate responses to CO_2 and solar forcing: A linear multivariate regression model characterizing transient climate change, J. Geophys. Res. Atmos., 120, 12 037–12 053, 2015.

Castruccio, S., McInerney, D. J., Stein, M. L., Crouch, F. L., Jacob, R. L., and Moyer, E. J.: Statistical Emulation of Climate Model Projections Based on Precomputed GCM Runs, J. Climate, 27, 1829–1844, 2014.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A., and Wehner, M.: Long-term Climate Change: Projections, Commitments and Irreversibility, 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, edited by Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Ferraro, A. J., Charlton-Perez, A. J., and Highwood, E. J.: Stratospheric dynamics and midlatitude jets under geoengineering with space mirrors and sulfate and titania aerosols, J. Geophys. Res. A, 120, 414–429, 2015.

Frieler, K., Meinshausen, M., Mengel, M., Braun, N., and Hare, W.: A scaling approach to probabilistic assessment of regional climate change, J. Climate, 25, 3117–3144, 2012.

Heckendorn, P., Weisenstein, D., Fueglistaler, S., Luo, B. P., Rozanov, E., Schraner, M., Thomason, L. W., and Peter, T.: The impact of geoengineering aerosols on stratospheric temperature and ozone, Env. Res. Lett., 4, 2009.

Herger, N., Sanderson, B. M., and Knutti, R.: Improved pattern scaling approaches for the use in climate impact studies, Geophys. Res. Lett., 42, 2015.

Holden, P. B. and Edwards, N. R.: Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling, Geophys. Res. Lett., 37, 2010.

Joshi, M. M., Turner, A. G., and Hope, C.: The use of land-sea warming contrast under climate change to improve impact metrics, Clim. Change, 117, 951–960, 2013.

Keith, D. W. and MacMartin, D. G.: A temporary, moderate and responsive scenario for solar geoengineering, Nature Climate Change, 5, 201–206, 2015.

Kravitz, B., Robock, A., Boucher, O., Schmidt, H., Taylor, K. E., Stenchikov, G., and Schulz, M.: The Geoengineering Model Intercomparison Project (GeoMIP), Atm. Sci. Lett., 12, 162–167, 2011.

Kravitz, B., Caldeira, K., Boucher, O., Robock, A., Rasch, P. J., , Alterskjær, K., Karam, D. B., Cole, J. N. S., Curry, C. L., Haywood, J. M., Irvine, P. J., Ji, D., Jones, A., Lunt, D. J., Kristjánsson, J. E., Moore, J., Niemeier, U., Schmidt, H., Schulz, M., Singh, B., Tilmes, S., Watanabe, S., Yang, S., and Yoon, J.-H.: Climate model response from the Geoengineering Model Intercomparison Project (GeoMIP), J. Geophys. Res., 118, 8320–8332, 2013.

Kravitz, B., MacMartin, D. G., Robock, A., Rasch, P. J., Ricke, K. L., Cole, J. N. S., Curry, C. L., Irvine, P. J., Ji, D., Keith, D. W., Kristjansson, J. E., Moore, J. C., Muri, H., Singh, B., Tilmes, S., Watanabe, S., Yang, S., and Yoon, J.-H.: A multi-model assessment of regional climate disparities caused by solar geoengineering, Env. Res. Lett., 9, 074 013, 2014.

Kravitz, B., MacMartin, D. G., Rasch, P. J., and Jarvis, A. J.: A new method of comparing forcing agents in climate models, J. Climate, 28, 8203–8218, 2015.

Kravitz, B., MacMartin, D. G., Wang, H., and Rasch, P. J.: Geoengineering as a Design Problem, Earth Systems Dynamics, 7, 469–497, 2016.

Long, J. C. S. and Shepherd, J. G.: The strategic value of geoengineering research, Global Environmental Change, 1, 2014.

MacMartin, D. G., Caldeira, K., and Keith, D. W.: Solar geoengineering to limit rates of change, Phil. Trans. Royal Soc. A, 372, 2014.

Mitchell, T. D.: Pattern Scaling: An examination of the accuracy of the technique for describing future climates, Climatic Change, 60, 217–242, 2003.

National Academy of Sciences: Climate Intervention: Reflecting Sunlight to Cool Earth, The National Academies Press, 500 Fifth St. NW, Washington DC 20001, 2015.

Neelin, J. D., Bracco, A., Luo, H., McWilliams, J. C., and Meyerson, J. E.: Considerations for parameter optimization and sensitivity in climate models, Proc. National Academy of Sciences, 107, 21 349–21 354, 2010.

Niemeier, U. and Timmreck, C.: What is the limit of climate engineering by stratospheric injection of SO_2?, Atmos. Chem. Phys., 15, 9129–9141, 2015.

Ragone, F., Lucarini, V., and Lunkeit, F.: A new framework for climate sensitivity and prediction: a modelling perspective, Clim. Dyn., 2015.

Ricke, K. L., Granger Morgan, M., and Allen, M. R.: Regional climate response to solar-radiation management, Nature Geoscience, 3, 537–541, 2010.

Santer, B. D., Wigley, T. M. L., Schlesinger, M. E., and Mitchell, J. F. B.: Developing climate scenarios from equilibrium GCM results, Tech. rep., MPI Report Number 47, 1990.

Schlesinger, M. E., Malyshev, S., Rozanov, E. V., Yang, F. L., Andronova, N. G., Vries, B. D., Grubler, A., Jiang, K. J., Masui, T., Morita, T., Penner, J., Pepper, W., Sankovski, A., and Zhang, Y.: Geographical distribution of temperature change for scenarios of greenhouse gas and sulfur dioxide emissions, Technol. Forecast Soc. Change, 65, 167–193, 2000.

Tebaldi, C. and Arblaster, J. M.: Pattern scaling: Its strengths and limitations, and an update on the latest model simulations, Climatic Change, 122, 459–471, 2014.

Tilmes, S., Sanderson, B. M., and O'Neill, B.: Climate impacts of geoengineering in a delayed mitigation scenario, Geophys. Res. Lett., 43, 8222–8229, 2016.