Baseline Evaluation of the Impact of Updates to the MIT Earth System Model on its Model Parameter Estimates

Alex G. Libardoni[1], Chris E. Forest[1,2], Andrei P. Sokolov[3], and Erwan Monier[3]

[1]Department of Meteorology, Pennsylvania State University, University Park, Pennsylvania, USA
[2]Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania, USA
[3]Joint Program on the Science and Policy of Global Change, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Correspondence to: Chris E. Forest (ceforest@psu.edu)

Abstract. For over twenty years, the Massachusetts Institute of Technology Earth System Model (MESM) has been used extensively for climate change research. The model is under continuous development with components being added or updated. To provide transparency in the model development, we perform a baseline evaluation of the newest version by comparing model behavior and properties to the previous model version. In particular, the impacts resulting from updates to the land surface

5   model component and the input forcings used in historical simulations of climate change are investigated. We run an 1800-member ensemble of MESM historical climate simulations where the model parameters that set climate sensitivity, ocean heat uptake, and the net anthropogenic aerosol forcing are systematically varied. By comparing model output to observed patterns of surface temperature changes, the linear trend in the increase in ocean heat content, and upper-air temperature changes, we derive probability distributions for the three model parameters. Furthermore, we run a 372-member ensemble of transient

10   climate simulations where model forcings are held fixed, absent an increase in carbon dioxide concentrations at the rate of 1% per year. From these runs, we derive a response surface for transient climate response and thermosteric sea level rise as a function of climate sensitivity and ocean heat uptake. We compare the probability distributions and response surfaces derived using the current version of MESM to the preceding version to evaluate the impact of the updated land surface model and forcing suite. We show that the probability distributions shift towards higher climate sensitivities and weaker aerosol forcing in

15   response to the new forcing suite. The climate response surfaces are relatively unchanged between model versions, indicating that the updated land surface model has limited impact on temperature evolution in the model.

1   Introduction

Equilibrium climate sensitivity (ECS), the equilibrium global-mean surface temperature change due to a doubling of atmospheric carbon dioxide concentrations, is a climate system property that has been widely studied and strongly influences future

20   climate projections. One of the complexities of ECS is that it is a function of many feedbacks and processes that act on different spatial and temporal scales. In particular, the lapse rate, water vapor, cryosphere, and cloud feedbacks play especially critical roles in determining the climate sensitiivty (Bony et al., 2006). Given its influence on future climate change, many studies using a range of methods have attempted to estimate ECS.

One class of studies estimates ECS directly from observations using a global energy budget approach (Gregory et al., 2002; Otto et al., 2013; Lewis and Curry, 2014; Masters, 2014). These studies calculate probability distributions of ECS from estimates of global mean surface temperature change, the heat stored in the ocean, and changes in radiative forcing, along with the associated uncertainties in their measurements. A second class of studies use simplified climate models such as

5 Earth System Models of Intermediate Complexity (EMICs) or energy balance models (e.g., Forest et al., 2002; Knutti et al., 2003; Forest et al., 2008; Libardoni and Forest, 2013; Olson et al., 2013; Johansson et al., 2015). Taking advantage of the computational efficiency of the simplified models, these studies run large ensembles with a range of climate sensitivity values in addition to other relevant factors, such as the ocean diffusivity and a measure of the net aerosol forcing. By comparing model runs to observations and evaluating how well individual model runs match the past, estimates of ECS and other parameters are

10 then presented as probability distributions.

Transient climate response (TCR) provides a second metric for estimating future climate change and is defined as the global mean surface temperature change at the time of carbon dioxide (CO_2) doubling in response to CO_2 concentrations increasing at the rate of 1% per year. CO_2 doubling occurs in year 70 of this scenario, making TCR a shorter-term assessment of climate change than ECS. Unlike ECS, which requires reaching an equilibrium state, TCR is estimated while the climate system is still

15 adjusting to a time-dependent forcing. There is a constant evolution in the strength and activity of processes and feedbacks in both the atmosphere and the ocean as the climate system adjusts to reach equilibrium. Due to the long time scales required to reach equilibrium, Allen and Frame (2007) argue that we should focus on estimating TCR, which is more policy-relevant than ECS. Estimates of TCR can be made from current historical observations and are more meaningful on the decadal time scale, whereas even if the equilibrium response is known, it may never be reached. However, even if more focus is placed on TCR

20 than ECS, the two are still closely linked. Warming on time scales relevant to estimating TCR is related to the sensitivity of the climate system to external forcings and the coupling between the atmosphere and the ocean. When considering atmosphere-ocean interactions, we know that TCR depends on both climate sensitivity and the rate at which heat is mixed into the deep ocean (Sokolov et al., 2003; Andrews and Allen, 2008).

One EMIC that has been extensively used in studies estimating ECS and TCR is the Earth system component of the Mas-

25 sachusetts Institute of Technology (MIT) Integrated Global Systems Model (IGSM, Sokolov et al., 2005). Forest et al. (2002, 2006, 2008) and Libardoni and Forest (2011, 2013) estimated the joint probability distribution for climate sensitivity and other model parameters in IGSM. Each study used similar, but not identical, versions of IGSM with changes both to key components of the model and to the input data used to force the model. Climate change diagnostics were also modified in the studies. The Earth system component of IGSM has undergone further development and a new, updated version incorporated into the

30 integrated framework. This study serves as a baseline evaluation of how probability distributions for the model parameters change as a result of updating the Earth system component. More specifically, we investigate the impact of (1) the structural changes to the model, (2) the historical datasets used to force the model, and (3) the sampling strategy used to vary the model parameters.

In the past, "IGSM" has been used to reference both the fully integrated model as well as the standalone Earth system

35 component. We follow this convention and refer to the older version of the Earth system model as IGSM, and we refer to the

updated version of the model as the MIT Earth System Model (MESM). In this study, we provide a transparent method of testing and accounting for how the simulated behavior and probability distribution functions change in response to the recent model development. We derive a new joint probability distribution by closely following the methods of Libardoni and Forest (2011) to show the impact that the new version of the model has on the parameter estimates and find that the new version of

5 the model leads to higher climate sensitivity estimates in addition to shifts in the distributions of the other model parameters. The effects on the parameter distributions due to changing observations and temperature metrics will be addressed in future papers to separate their impacts from those due to changes to the model framework alone. We also show here how the emergent behavior of MESM compares to the older IGSM by running a new set of transient simulations and calculating how the response surfaces for TCR and thermosteric sea level rise depend on ECS and ocean diffusivity.

10 In Section 2, we give a brief description of the MIT modeling framework and the differences between IGSM and MESM. We describe the process for deriving the joint probability distribution function used in Libardoni and Forest (2011) and the modifications implemented in this study in Section 3. Parameter distributions and response surfaces are presented in Section 4. In particular, we test whether changes in the distributions and responses are due to reducing the number of model diagnostics, the sampling of the parameter space, or changes in the model structure and input forcings. We present our conclusions in

15 Section 5.

2 Model

The coupled atmosphere-ocean-land model of the updated MIT Earth System Model (Sokolov et al., 2018) replaces the version described in Sokolov et al. (2005). The first update to the model was the incorporation of a new land surface model. The Community Land Model (CLM) version 3.5 (Oleson et al., 2008) replaced CLM version 2.1 to improve estimates of the

20 surface heat balance in the model. A second update to the model was an adjustment to the radiative forcing of non-CO_2 greenhouse gases in the radiation code. The adjustment was made to match the calculations used in the Intergovernmental Panel on Climate Change (IPCC) experiments and produces weaker forcing for those constituents. Additionally, the forcings used to drive the model until now (Forest et al., 2006) were extended and, in some cases, new data sources were used. Greenhouse gas concentrations and stratospheric aerosols from volcanic eruptions were obtained from the National Aeronautics and Space

25 Administration Goddard Institute for Space Studies modeling group forcing suite. The procedure for updating the greenhouse gas emissions from Hansen et al. (2007) and the volcanic aerosol forcing from Sato et al. (1993) was described in Miller et al. (2014). Updates included incorporating data from more observational sources and extending the length of the datasets. Sulfate aerosol loading from Smith et al. (2011) was extended to 2011 by Klimont et al. (2013). The Kopp and Lean (2011) solar irradiance dataset replaced the Lean (2000) dataset. Lastly, the ozone concentration database developed by the Atmospheric

30 Chemistry and Climate initiative (AC&C) and Stratospheric Processes and their Role in Climate project (SPARC) ozone concentration database (Cionni et al., 2011) that was developed in support of the Coupled Model Intercomparison Project phase 5 (CMIP5) replaced the concentration data used in Forest et al. (2006). The concentrations in the dataset, hereafter referred to as AC&C/SPARC, drive the tropospheric and stratospheric ozone forcing in the radiation code. In Section 4, we

show the differences between the old and new datasets for those forcings where the data sources have changed, namely solar and ozone.

Three model parameters that impact the climate system response are easily modified in MESM. These parameters are the effective climate sensitivity (ECS), the effective ocean diffusivitiy (K_v), and the net aerosol scaling factor (F_{aer}). ECS is
5 changed by adjusting the strength of the cloud feedback at different levels in the model (Sokolov, 2006; Sokolov and Monier, 2012). K_v represents the vertical diffusion of heat anomalies into the deep ocean by all mixing processes and tends to be larger than typical ocean diffusivity values which represent the diffusion of heat alone (Sokolov et al., 2003). The mixing pattern is prescribed spatially with stronger mixing in the polar regions and weaker mixing near the equator. K_v represents the global mean diffusion rate and the spatial pattern is scaled to obtain the desired value. The anthropogenic aerosol forcing
10 used in the model is prescribed by a latitude-dependent pattern that differs over land and ocean and is used as an estimate of all unmodeled forcings in the simulations (Forest et al., 2001). This pattern is held fixed spatially but scaled temporally by estimated emissions of sulfur dioxide. F_{aer} sets the amplitude of the pattern in the 1980s. By choosing a set of the three parameters, $\theta = (ECS, K_v, F_{aer})$, we simulate different climate states.

3 Methods

15 In this section, we present an outline of the methodology used to derive the joint probability distribution function for the model parameters and highlight the changes implemented between this study and previous studies using IGSM. We follow closely the methods of Libardoni and Forest (2011) with two notable changes. First, we run the model for θs that sample individual parameters over a wider range and on a more regular grid. Climate sensitivity is sampled from 0.5 to 10.0 °C in increments of 0.5 °C by adjusting the strength of the cloud feedback, the square root of ocean diffusivity is sampled from 0 to 8 cm s$^{-1/2}$ in
20 increments of 1 cm s$^{-1/2}$, and the aerosol forcing amplitude is sampled from -1.75 to 0.5 Wm^{-2} in increments 0.25 Wm^{-2}. By choosing this sampling strategy, we have increased the number of runs from 640 with IGSM to 1800 runs with MESM, widened the range of parameter values sampled, and increased the density of model runs within the parameter space (Figure 1).

As a second change, we reduce the number of diagnostics used to evaluate model performance. We omitted the upper-air temperature diagnostic because it is highly correlated with the surface temperature diagnostic (Lewis, 2013). This leaves two
25 temperature diagnostics for evaluating model performance: (1) decadal mean surface air temperature anomalies from 1946-1995 with respect to a 1906-1995 climatology in four equal-area zonal bands, and (2) the linear trend in global mean ocean heat content from 1955-1995 in the 0-3 km layer. As in Libardoni and Forest (2011), we use five surface temperature datasets (Jones and Moberg, 2003; Brohan et al., 2006; Smith et al., 2008; Hansen et al., 2010) and one ocean heat content dataset (Levitus et al., 2005) as observations.
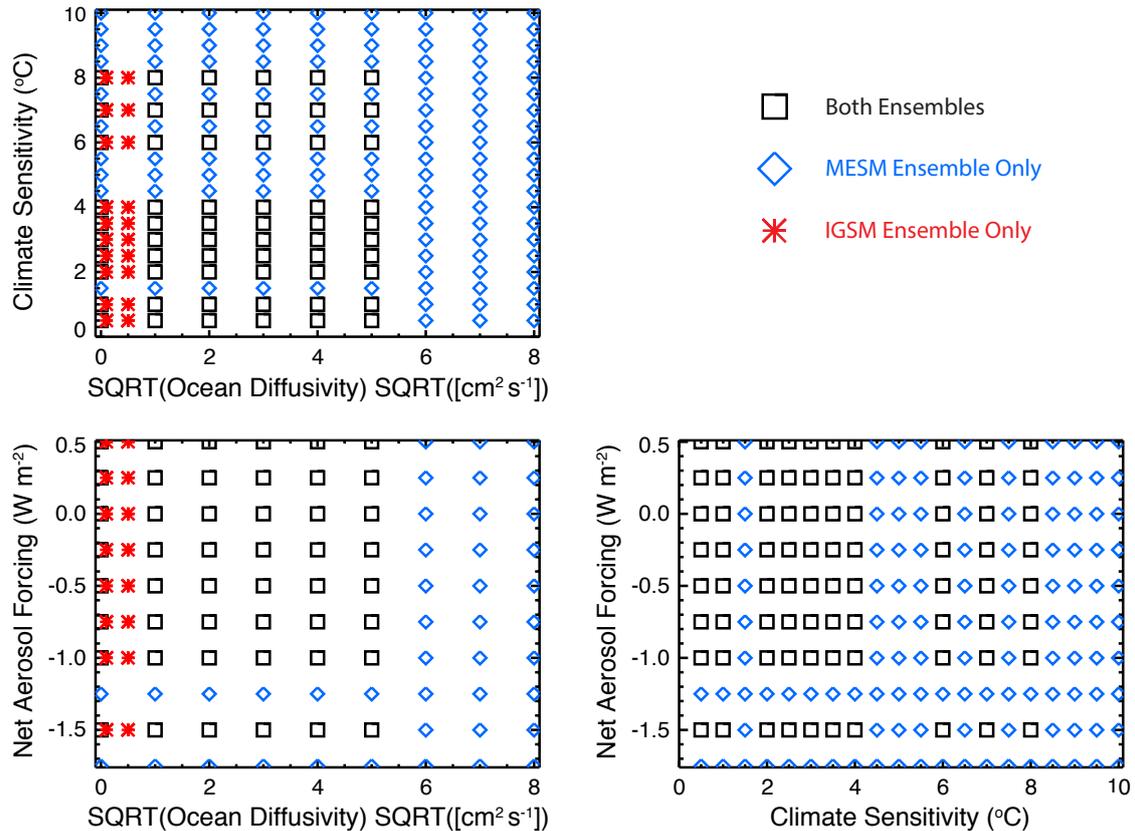
Figure 1. Parameter pairings where the models have been run. Points in black are common to both the IGSM and MESM ensembles. Blue points are unique to the IGSM ensemble and red points are unique to the MESM ensemble.

4   Results

Our results are presented as follows. We first identify the changes in the input forcings used in our historical simulations by comparing the solar and ozone components used in the IGSM runs with those used in the MESM runs. Second, we show how the probability distribution functions change when reducing the number of model diagnostics from three to two through the omission of the upper-air diagnostic. Third, we derive probability distributions using the MESM ensemble and directly compare them to those derived using the IGSM ensemble using the full ensembles and the case where only runs with θs common to both ensembles are used. Finally, we derive the response surfaces for transient climate response and thermosteric sea level rise for MESM and compare them to the corresponding surfaces from IGSM.

To identify changes in the forcing time series used to drive the model, we compare the input forcings for the two components for which we have changed datasets. In Figure 2, we show the old and new solar forcing time series. We see that the biggest
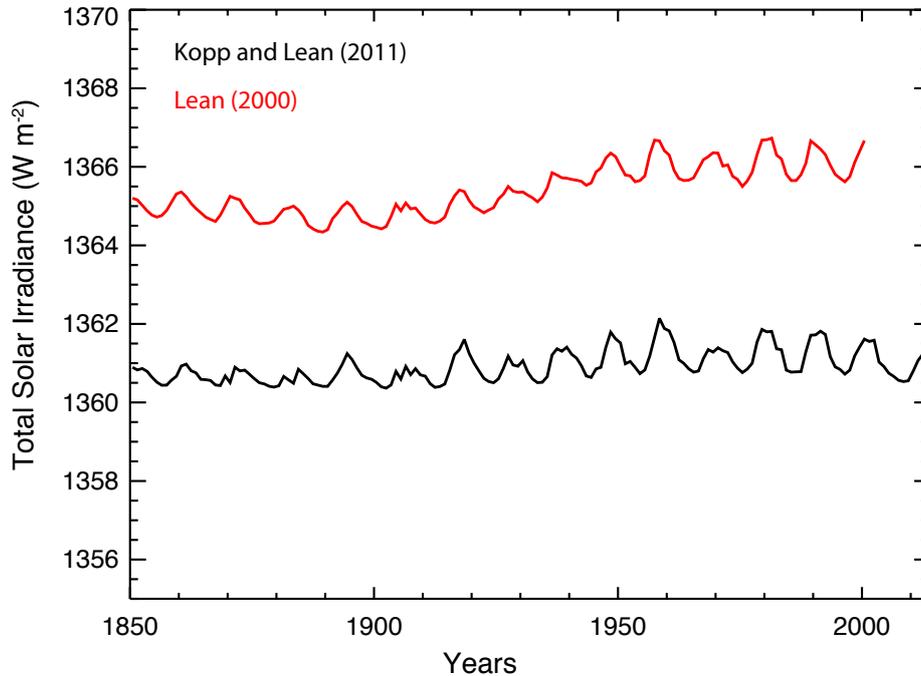
Figure 2. Annual mean total solar irradiance. The bias between the Lean (2000) and Kopp and Lean (2011) datasets leads to a reduction in radiative forcing in the new forcing suite.

difference observed in the solar irradiance time series is a bias towards lower values when using the Kopp and Lean (2011) data. The bias is relatively constant at approximately $4.5\,\mathrm{Wm^{-2}}$ until 1920, then increases towards $5.0\,\mathrm{Wm^{-2}}$ moving forward in time. The mean bias is accounted for in the Q-flux adjustment in the mixed-layer ocean model which specifies the vertically-integrated horizontal heat transport in the mixed layer required to maintain historical sea surface temperatures (Sokolov et al.,

5 2005). However, because the Q-flux is calculated offline from control simulations, the pattern is fixed throughout the run. Any time-varying change to an input forcing cannot be accounted for in the Q-flux calculation. Thus, the growth of the low bias means that the solar forcing weakens with time beginning in 1920 in the new suite of forcings.

   We observe that the ozone concentrations estimated from the AC&C/SPARC dataset differ in both space and time when compared to the previous concentrations used with IGSM (Figure 3). One clear difference is that the AC&C/SPARC dataset

10 introduces more temporal variability in stratospheric ozone concentrations (which we approximate as pressure levels above 200 mb) prior to 1950. Post-1950, AC&C/SPARC tends to have lower ozone concentrations in the stratosphere and slightly greater concentrations in the troposphere (levels below 200 mb). However, similar to with the solar forcing, we are concerned with the temporal change in the forcing imposed by the ozone concentrations, rather than the relative magnitude of the concentrations across datasets. Beginning in 1900, tropospheric ozone concentrations increase less rapidly in the AC&C/SPARC dataset when

15 compared to the IGSM dataset. Differences in stratospheric ozone concentrations remain relatively constant until 1950, but then
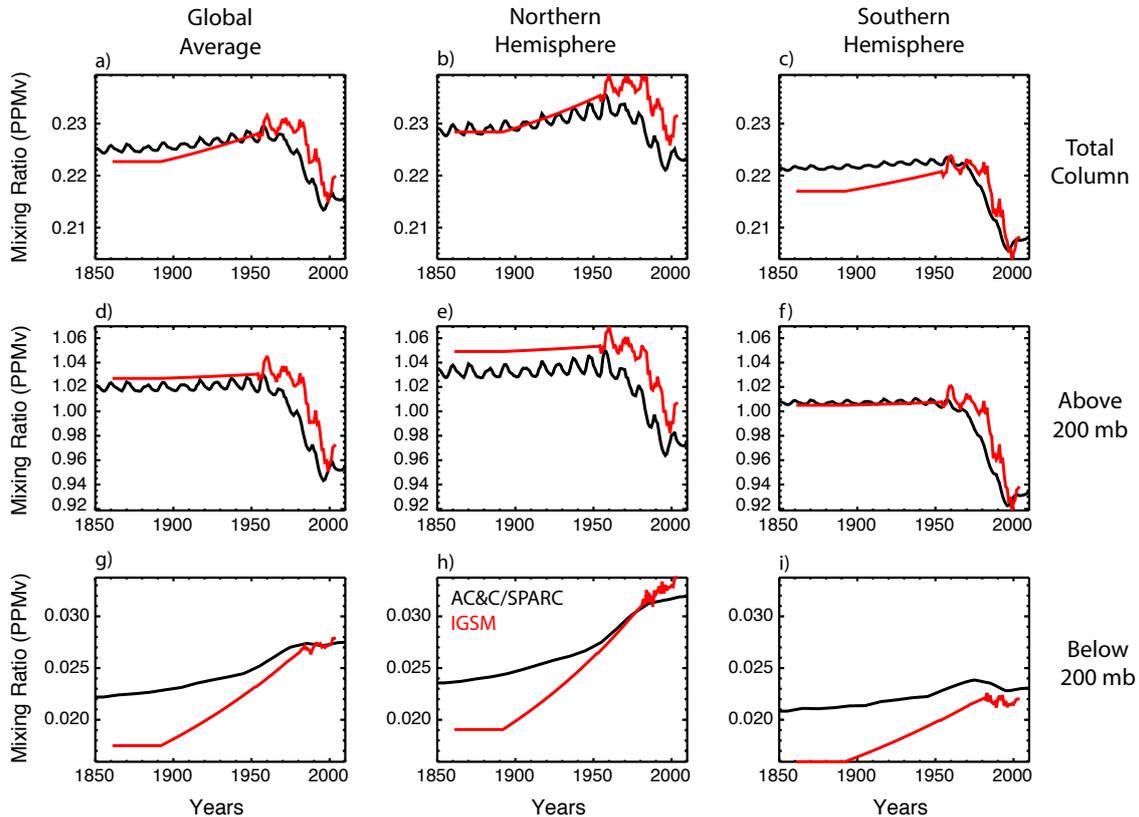
Figure 3. Ozone concentration in the old IGSM time series (red) and the Cionni et al. (2011) AC&C/SPARC concentrations (black). (a-c) Annual mean ozone mixing ratio in the total column in the global average (a), northern hemisphere (b), southern hemisphere (c). (d-f) As in (a-c) but for the average above 200 mb. (g-i) As in (a-c) but for the average below 200 mb.

decrease at a slower rate in the AC&C/SPARC time series. These patterns are generally consistent in the global and hemispheric means. When considered separately, increased tropospheric ozone concentrations tend to increase radiative forcing (Stevenson et al., 2013) and decreased stratospheric concentrations tend to increase radiative forcing (Conley et al., 2013). Thus, the less rapid increase in tropospheric ozone concentration and less rapid decrease in stratospheric ozone concentration in the
5 AC&C/SPARC dataset both contribute to a weaker radiative forcing over the historical period in the new suite of forcings.

With the input forcings documented, we focus on deriving probability distributions for the model parameters. We first test the impact of omitting the upper-air diagnostic. Starting from the distributions calculated in Libardoni and Forest (2011), we derive new distributions based only on the surface temperature and ocean heat content diagnostics presented in Section 3. We show that reducing the number of diagnostics from three to two has little impact on the parameter estimates (Table 1). We only
10 present comparisons for ECS and F_{aer} because distributions of K_v were poorly constrained in Libardoni and Forest (2011) and no uncertainty bounds were given. In general, ECS estimates tend to be slightly lower when using only two diagnostics and

Table 1. 90-percent confidence intervals for ECS and F_{aer}. Distributions that include the upper-air diagnostic are from Libardoni and Forest (2011) and distributions with two diagnostics exclude the upper-air diagnostic.

Surface Temperature Dataset	# Diagnostics	ECS (°C)		F_{aer} (Wm^{-2})	
		5%	95%	5%	95%
HadCRUT2[1]	3	2.0	5.3	-0.19	-0.70
	2	1.9	5.2	-0.19	-0.71
HadCRUT3[2]	3	1.9	5.1	-0.22	-0.74
	2	1.7	5.0	-0.38	-0.79
NCDC[3]	3	1.8	4.7	-0.37	-0.78
	2	1.6	4.8	-0.38	-0.79
GISTEMP250[4]	3	1.3	3.6	-0.32	-0.83
	2	1.1	4.0	-0.35	-0.83
GISTEMP1200[5]	3	1.2	3.4	-0.33	-0.80
	2	1.0	3.7	-0.35	-0.83

[1]Hadley Centre Climatic Research Unit Temperature version 2 (Jones and Moberg, 2003)

[2]Hadley Centre Climatic Research Unit Temperature version 3 (Brohan et al., 2006)

[3]National Climatic Data Center merged land-ocean dataset (Smith et al., 2008)

[4]GISS Surface Temperature Analysis with 250 km smoothing (Hansen et al., 2010)

[5]GISS Surface Temperature Analysis with 1200 km smoothing (Hansen et al., 2010)

aerosol estimates are nearly unchanged. Further, the relationships between the distributions with respect to surface dataset are unchanged. Because the changes using only two diagnostics are minimal and do not change any conclusions from the original study, we justify the removal of the upper-air diagnostic.

We next evaluate the impacts that changing the model from IGSM to MESM and updating the forcing suite have on the
5   parameter distributions by comparing model output from each ensemble member against the temperature diagnostics discussed in Section 3. Following the methods outlined in Libardoni and Forest (2011), we calculate goodness-of-fit statistics across all runs for each diagnostic and convert them to a joint probability distribution function for the model parameters. Marginal probability distributions for individual parameters are then calculated by integrating the joint distribution over the other two parameters. We present the new distributions in Figure 4 and observe significant differences between distributions derived using
10   IGSM and those derived using MESM with the updated forcings (Table 2). Across all datasets, climate sensitivity distributions shift towards higher values and the uncertainty bounds encompass a wider range. When considering the 90-percent confidence intervals from across the distributions derived from each surface dataset, we find climate sensitivity now lies between 1.3 and 5.7 °C, as opposed to the estimated interval of 1.2 to 5.3 °C from Libardoni and Forest (2011). While the uncertainty bounds are still wide compared to other parameters, we observe that K_v is now better constrained with MESM. The distributions of
15   K_v derived using the GISTEMP datasets are still unconstrained with upper tails extending to the edge of the parameter domain,
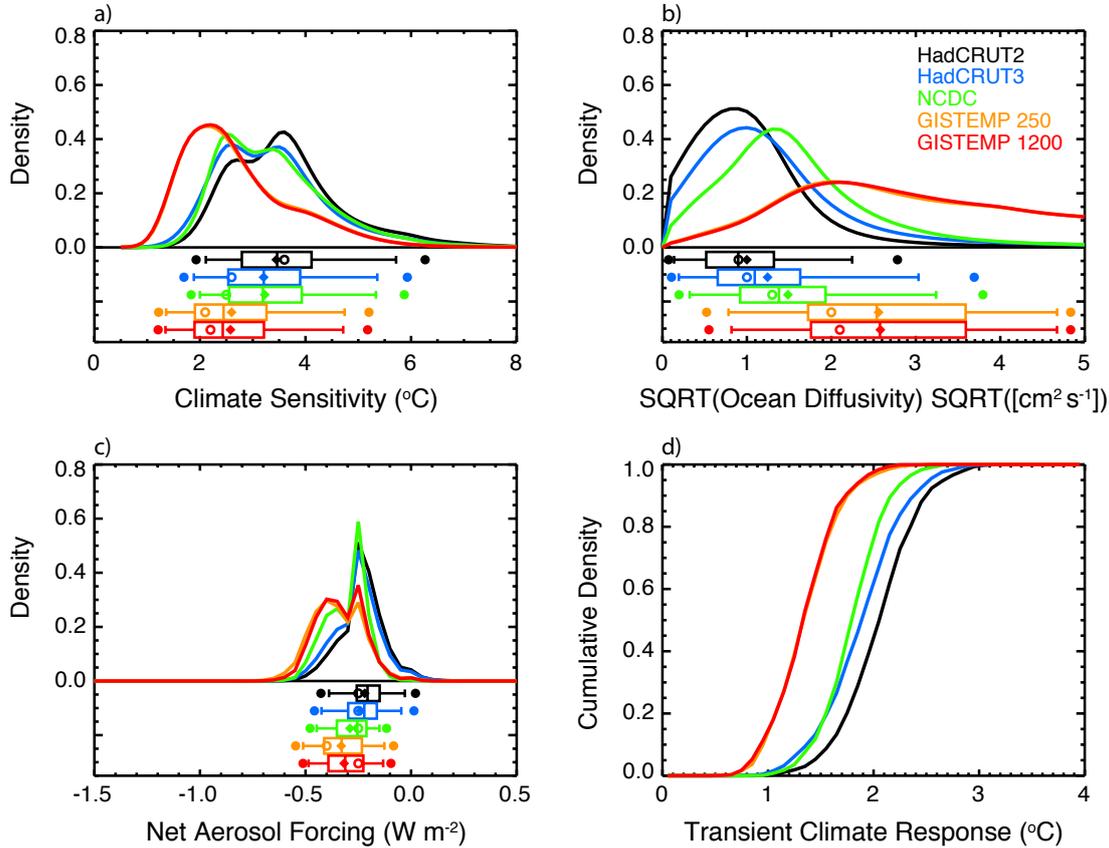
Figure 4. Marginal probability distribution functions and TCR cumulative distribution functions derived from MESM simulations using the HadCRUT2, HadCRUT3, NCDC, GISTEMP 250, and GISTEMP 1200 surface temperature datasets as observations. (a) ECS, (b) K_v, and (c) F_{aer}. Whisker plots indicate boundaries for the 2.5-97.5 (dots), 5-95 (vertical bar ends), 25-75 (box ends), and 50 (vertical bar in box) percentiles. Distribution means are represented by diamonds and modes are represented by open circles. (d) TCR CDFs are derived from 1000 member Latin Hypercube samples from the joint parameter distributions and the TCR(ECS, $\sqrt{K_v}$) functional fit.

but all other datasets now show an upper bound well within the ensemble range. We also observe a marked shift in the aerosol estimates. When MESM is used with the updated forcing suite, there is a sizable shift towards weaker aerosol forcing across all datasets. Whereas past estimates put net aerosol forcing between -0.83 and -0.19 Wm^{-2}, our new estimate of aerosol forcing is between -0.53 and -0.03 Wm^{-2}.

5   The shifts we observe in the parameter estimates are consistent with the changes in the input forcings. Both the solar and ozone forcing patterns lead to a reduction in their contribution to the global radiation budget and decrease the net radiatiave forcing on the planet. Because the diagnostics do not change, model runs with a weaker external forcing are compared against the same observed temperature patterns. Weaker increases in external forcing require higher climate sensitivity to match the

Table 2. 90-percent confidence intervals and means for climate sensitivity (ECS), ocean diffusivity (K_v), and net aerosol forcing (F_{aer}). Surface temperature datasets are the same as in Table 1.

Surface Temperature Dataset	Model and Runs	ECS (°C)			$\sqrt{K_v}$ (cm s$^{-1/2}$)			F_{aer} (Wm^{-2})		
		5%	95%	Mean	5%	95%	Mean	5%	95%	Mean
HadCRUT2	Full IGSM	1.9	5.2	3.0	0.1	2.1	0.9	-0.19	-0.71	-0.46
	Subsampled IGSM	1.9	5.2	3.0	0.1	2.1	0.9	-0.16	-0.71	-0.45
	Full MESM	2.1	5.7	3.5	0.1	2.3	1.0	-0.03	-0.39	-0.22
	Subsampled MESM	2.1	5.7	3.4	0.1	2.2	1.0	-0.03	-0.39	-0.22
HadCRUT3	Full IGSM	1.7	4.0	2.8	0.2	2.9	1.2	-0.22	-0.75	-0.50
	Subsampled IGSM	1.7	4.0	2.8	0.2	2.9	1.2	-0.20	-0.75	-0.49
	Full MESM	1.9	5.4	3.2	0.2	3.6	1.3	-0.05	-0.43	-0.24
	Subsampled MESM	1.9	5.4	3.2	0.2	3.0	1.2	-0.05	-0.42	-0.24
NCDC	Full IGSM	1.6	4.8	2.7	0.3	3.7	1.6	-0.38	-0.79	-0.59
	Subsampled IGSM	1.6	4.8	2.7	0.3	3.7	1.6	-0.36	-0.79	-0.58
	Full MESM	2.0	5.4	3.2	0.3	3.7	1.6	-0.15	-0.45	-0.29
	Subsampled MESM	2.0	5.3	3.2	0.3	3.2	1.5	-0.15	-0.45	-0.29
GISTEMP 250	Full IGSM	1.1	4.0	2.1	0.7	4.8	2.7	-0.35	-0.86	-0.61
	Subsampled IGSM	1.1	4.0	2.1	0.6	4.8	2.7	-0.35	-0.86	-0.60
	Full MESM	1.3	4.8	2.6	0.8	7.3	3.5	-0.13	-0.53	-0.34
	Subsampled MESM	1.4	4.7	2.6	0.8	4.7	2.6	-0.13	-0.51	-0.33
GISTEMP 1200	Full IGSM	1.0	3.7	1.9	0.8	4.9	3.1	-0.35	-0.83	-0.56
	Subsampled IGSM	1.0	3.7	1.9	0.7	4.9	3.1	-0.35	-0.82	-0.56
	Full MESM	1.3	4.8	2.6	0.8	7.3	3.5	-0.14	-0.49	-0.33
	Subsampled MESM	1.3	4.7	2.6	0.8	4.7	2.6	-0.14	-0.49	-0.32

same warming trend. In the model, the aerosol forcing pattern is a negative term in the global energy budget. It should follow that if, as noted previously, there is weaker net forcing due to the changes in forcing datasets, our estimates of the aerosol amplitude should become less negative. From these arguments, combinations of higher climate sensitivity and weaker aerosol forcing offset the impact of a weaker external forcing. Further, these results are consistent with the weakening of the non-CO$_2$

5 greenhouse gas forcing introduced by the new radiation code in MESM. We observe these shifts in the distributions when comparing the distributions derived from the old and new models (Table 2).

To test whether the differences observed in the parameter estimates were due to the model update, rather than the increased density of model runs, we subsampled each ensemble at the 480 θs where they overlap (see Figure 1). We summarize the distributions in Table 2 and see that there is very little sensitivity when the ensembles are subsampled. Across all datasets, the

10 distributions we derive using the full 640-member IGSM ensemble and those we derive using the 480-member IGSM ensemble are nearly identical for all three parameters. The same is true for the MESM ensemble, except for the distributions we derive for K_v. We consistently estimate a smaller upper bound for K_v in the subsampled MESM ensemble compared to when the full MESM ensemble is used. This arises because the wider range of K_v sampled in the MESM ensemble does not artificially cut off the distribution for values of $\sqrt{K_v}$ greater than 5 cm s$^{-1/2}$. Thus, the differences we observe between the old and new

15 ensembles are due to the differences between the model and forcing themselves and not the increased density of model runs.
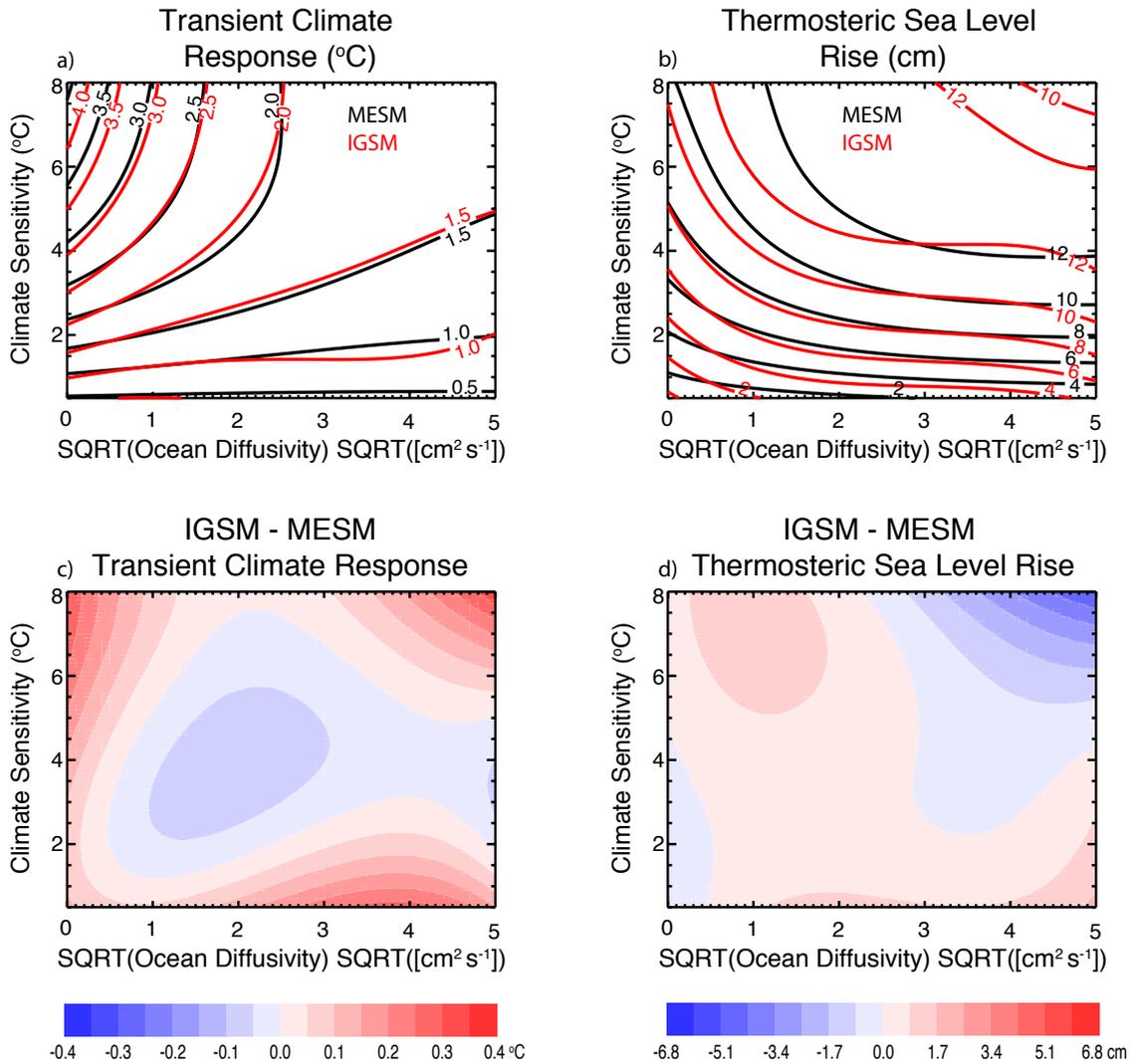
Figure 5. Model response surfaces for (a) TCR and (b) thermosteric sea level rise. Contours for the MESM response surfaces are shown in black and contours for the IGSM surfaces are shown in red. Differences between the fits are also shown (c and d).

To estimate TCR in MESM, we run a 372-member ensemble where all forcings are held fixed and carbon dioxide concentrations are increased by 1% per year. We calculate TCR by estimating the global mean temperature change from the beginning of the simulations to the time of CO_2 doubling. Concentrations double in year 70 and we estimate TCR as the average global mean temperature change in years 60-80 of the simulation. Temperature changes are calculated with respect to a control simulation with the same model parameters and all forcings held fixed. In a similar manner, we also estimate thermosteric sea level rise (SLR) at the time of doubling. Because all forcings except those attributed to CO_2 are fixed, each ECS-$\sqrt{K_v}$ pair yields a single TCR value and a single SLR value, independent of F_{aer}. We fit a third-order polynomial in ECS and $\sqrt{K_v}$ to the TCR and SLR values calculated from each run to derive a functional fit for all parameter pairs within the domain. From these fits, we derive response surfaces for each of the transient properties (Figure 5). For comparison, we also show the fit derived using the IGSM and its corresponding 1% per year runs, in addition to the differences between the two. Outside of the region where ECS is greater than 4 °C and $\sqrt{K_v}$ is less than about 0.5 cm s$^{-1/2}$, and away from the edges of the domain, TCR values from IGSM and MESM agree quite well. There is a similar pattern of agreement in the SLR response surface, with the biggest discrepancies occurring in the high ECS-high $\sqrt{K_v}$ region and near the edges of the parameter domain.

We use the response surface to derive probability distributions for TCR. From each of the joint probability distributions derived from the subsampled MESM ensemble, we draw a 1000-member Latin Hypercube Sample (McKay et al., 1979) of model parameters. The subsampled distributions are chosen so that we restrict the domain to that of the IGSM ensemble, allowing for a more direct comparison of the distributions. Otherwise, high K_v values that are within the domain of the functional fit to the MESM runs would be selected, for which there is no fit using the IGSM function. We map each of the ECS-$\sqrt{K_v}$ pairs onto the response surface to provide an estimate of TCR values. Binning the responses in a histogram with bin size = 0.1 °C allows a PDF to be calculated, and the resulting cumulative density functions derived using MESM are displayed in Figure 4d. Comparing TCR distributions for the IGSM and MESM ensembles shows a shift towards higher TCR with the latest results. When comparing the range of 90-percent confidence intervals derived using MESM to those from Libardoni and Forest (2011), we find that TCR estimates increase from 0.87-2.31 °C using IGSM to 0.90-2.72 °C using MESM. We have shown previously that the marginal distributions of $\sqrt{K_v}$ are similar between the two models, indicating that this shift towards higher TCR is driven by the higher ECS estimates derived from MESM.

5 Conclusions

In this study, we have provided an open, transparent means of testing the changes in model response and parameter estimation to changes in the MIT Integrated Global Systems Model modeling framework. Not only does this systematic accounting of the impacts give a reference point moving forward for studies involving MESM, it proposes a template for assessing the impact that changes in other simplified climate models have on the calibration of their own model parameters. It hoped that this study motivates similar investigations moving forward that produce documented accountings of model updates and that lead to a robust understanding of the impacts the changes have on the parameter calibration.

By updating the model and its input forcings, we identify the impact that the switch from the MIT Integrated Global Systems Model to the MIT Earth System Model has on the probability distributions of model parameters. The decreases in radiative forcing due to the change in radiative forcing code, the new solar radiation data, and the new ozone concentrations used to estimate the ozone forcing lead to a net energy deficit when compared to the replaced forcings. This drives an upward

5 shift in our estimates of the 90-percent confidence interval for climate sensitivity from 1.2 to 5.3 °C to 1.3 and 5.7 °C, a better constraint on ocean diffusivity, and a decrease in the 90-percent confidence interval for the net anthropogenic aerosol forcing from between -0.83 and -0.19 Wm^{-2} to between -0.53 and -0.03 Wm^{-2}. One caveat of our analysis is that because we changed the forcings and CLM simultaneously, we cannot fully attribute the parameter shifts to the model forcings alone. We have shown the total effect of changing both the model and forcings on the parameter distributions, not the effects of the

10 changes individually.

Because TCR is independent of the input forcings, the only difference between the IGSM and MESM configurations in the transient simulations is the land surface model. By showing that the transient climate response surfaces derived from the two models differ only slightly, we provide evidence that the switch to CLM3.5 does not greatly impact the temperature evolution in the model. We have drawn Latin Hypercube Samples from the parameter distributions to provide estimates of TCR from the

15 new response surface. Due to the shift towards higher climate sensitivity and slightly weaker ocean diffusivity, we observe an increase in our 90-percent confidence interval of transient climate response from 0.87-2.31 °C to 0.85-2.73 °C. By investigating the impact that the new forcings and a newer version of CLM have on the estimates of model parameters and TCR, we have shown the inherent differences that are present when comparing distributions derived using IGSM and those derived from MESM.

20 *Code and data availability.* The source code of MESM will become publicly available for non-commercial research and educational purposes as soon as a software license that is being prepared by the MIT Technology Licensing Office is complete. For further information contact mesm-request@mit.edu. A working paper describing and evaluating the MESM is available at http://svante.mit.edu/ mesm/publications/MESM-paper.pdf. All data required to reproduce the figures and tables in the main text and scripts to replicate the figures are available in an online archive. Model output is available upon request.

25 *Author contributions.* AGL and APS carried out the MESM simulations. APS wrote the codes for extracting model output. AGL performed the analysis and prepared the original manuscript. AGL and CEF developed the model ensemble and experimental design. AGL, CEF, APS, and EM all contributed to interpreting the analysis and synthesizing the findings.

Competing interests. The authors declare no competing interests.

Acknowledgements. This work was supported by U.S. Department of Energy (DOE), Office of Science under award DE-FG02-94ER61937 and other government, industry and foundation sponsors of the MIT Joint Program on the Science and Policy of Global Change. For a complete list of sponsors and U.S. government funding sources, see http://globalchange.mit.edu/sponsors/. The authors would like to thank the National Climatic Data Center, the Hadley Centre for Climate Prediction and Research, and the NASA Goddard Institute for Space Studies

5   for producing publicly available surface data products and the NOAA National Centers for Environmental Information for providing publicly available ocean heat content data. We would also like to thank the University of Maryland for access to the Evergreen high-performance computing cluster for model simulations.

References

Allen, M. R. and Frame, D. J.: Call off the Quest, Science, 318, 582–583, 2007.

Andrews, D. G. and Allen, M. R.: Diagnosis of climate models in terms of transient climate response and feedback response time, Atm. Sci. Letters, 9, 7–12, 2008.

5 Bony, S., Colman, R., Kattsov, V. M., Allan, R. R., Bretherton, C. S., Dufresne, J.-L., Hall, A., Hallegatte, S., Holland, M. M., Ingram, W., Randall, D. A., Soden, B. J., Tselioudis, G., and Webb, M. J.: How well do we understand and evaluate climate change feedback processes?, J. Clim., 19, 3445–3482, 2006.

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D.: Unceratainty estimates in regional and global observed temperature changes: A new data set from 1850, J. Geophys. Res., 111, doi:10.1029/2005JD006 548, 2006.

10 Cionni, I., Eyring, V., Lamarque, J. F., Randel, W. J., Stevenson, D. S., Wu, F., Bodeker, G. E., Shepherd, T. G., Shindell, D. T., and Waugh, D. W.: Ozone Database in Support of CMIP5 Simulations: Results and Corresponding Radiative Forcing, Atmos. Chem. Phys., 11, 11 267–11 292, 2011.

Conley, A. J., Lamarque, J.-F., Vitt, F., Collins, W. D., and Kiehl, J.: PORT, a CESM Tool for the Diagnosis of Radiative Forcing, Geosci. Model Dev., 6, 469–476, 2013.

15 Forest, C. E., Allen, M. R., Sokolov, A. P., and Stone, P. H.: Constraining Climate Model Properties Using Optimal Fingerprint Detection Methods., Clim. Dyn., 18, 277–295, 2001.

Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R., and Webster, M. D.: Quantifying uncertainties in climate system properties with the use of recent climate observations., Science, 295, 113–117, 2002.

Forest, C. E., Stone, P. H., and Sokolov, A. P.: Estimated PDFs of climate system properties including natural and anthropogenic forcings,
20 Geophys. Res. Let., 33, doi:10.1029/2005GL023 977, 2006.

Forest, C. E., Stone, P. H., and Sokolov, A. P.: Constraining climate model parameters from observed 20th century changes, Tellus, 60A, 911–920, 2008.

Gregory, J. M., Stouffer, R. J., Raper, S. C. B., Stott, P. A., and Rayner, N. A.: An observationally based estimate of the climate sensitivity, J. Clim., 15, 3117–3121, 2002.

25 Hansen, J., Sato, M., Ruedy, R., Kharecha, P., Lacis, A., Miller, R., Nazarenko, L., Lo, K., Schmidt, G., Russell, G., Aleinov, I., Bauer, S., Baum, E., Cairns, B., Canuto, V., Chandler, M., Cheng, Y., Cohen, A., Genio, A. D., Faluvegi, G., Fleming, E., Friend, A., Hall, T., Jackman, C., Jonas, J., Kelley, M., Kiang, N., Koch, D., Labow, G., Lerner, J., Menon, S., Novakov, T., Oinas, V., Perlwitz, J., Perlwitz, J., Rind, D., Romanou, A., Schmunk, R., Shindell, D., Stone, P., Sun, S., Streets, D., Tausnev, N., Thresher, D., Unger, N., Yao, M., and Zhang, S.: Climate simulations for 1880–2003 with GISS modelE, Clim. Dyn., 29, doi:10.1007/s00 382–007–0255–8, 2007.

30 Hansen, J., Ruedy, R., Sato, M., and Lo, K.: Global surface temperature change, Rev. Geophys., 48, doi:10.1029/2010RG000 345, 2010.

Johansson, D. J. A., O'Neill, B. C., Tebaldi, C., and Haggstrom, O.: Equilibrium climate sensitivity in light of observations over the warming hiatus, Nature Clim. Change, 5, 449–453, 2015.

Jones, P. and Moberg, A.: Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001, J. Clim., 16, 206–223, 2003.

35 Klimont, Z., Smith, S. J., and Cofala, J.: The last decade of global anthropogenic sulfur dioxide: 2000–2011 emissions, Environ. Res. Lett., 8, doi:10.1088/1748–9326/8/1/014 003, 2013.

Knutti, R., Stocker, T. F., Joos, F., and Plattner, G.: Probabilistic climate change projections using neural networks, Clim. Dyn., 21, 257–272, 2003.

Kopp, G. and Lean, J. L.: A New, Lower Value of Total Solar Irradiance: Evidence and Climate Significance, Geophys. Res. Let., 38, doi:10.1029/2010GL045 777, 2011.

5 Lean, J. L.: Evolution of the Sun's Spectral Irradiance Since the Maunder Minimum, Geophys. Res. Let., 27, 2425–2428, 2000.

Levitus, S., Antonov, J., and Boyer, T.: Warming of the world ocean, 1955–2003, Geophys. Res. Let., 32, doi:10.1029/2004GL021 592, 2005.

Lewis, N.: An objective Bayesian improved approach for applying optimal fingerprint techniques to climate sensitivity, J. Clim., 26, doi:10.1175/JCLI–D–12–00 473.1, 2013.

Lewis, N. and Curry, J. A.: The implications for climate sensitivity of AR5 forcing and heat uptake estimates, Clim. Dyn., 45,
10   doi:10.1007/s00 382–014–2342–y, 2014.

Libardoni, A. G. and Forest, C. E.: Sensitivity of distributions of climate system properties to the surface temperature dataset, Geophys. Res. Let., 38, doi:10.1029/2011GL049 431, 2011.

Libardoni, A. G. and Forest, C. E.: Correction to "Sensitivity of distributions of climate system properties to the surface temperature data set", Geophys. Res. Let., 40, doi:10.1002/grl.50 480, 2013.

15 Masters, T.: Observational estimate of climate sensitivity from changes in the rate of ocean heat uptake and comparison to CMIP5 models, Clim. Dyn., pp. doi:10.1007/s00 382–013–1770–4, 2014.

McKay, M. D., Beckman, R. J., and Conover, W. J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245, 1979.

Miller, R. L., Schmidt, G. A., Nazarenko, L. S., Tausnev, N., Bauer, S. E., DelGenio, A. D., Kelley, M., Lo, K. K., Ruedy, R., Shindell, D. T.,
20   Aleinov, I., Bauer, M., Bleck, R., Canuto, V., Chen, Y., Cheng, Y., Clune, T. L., Faluvegi, G., Hansen, J. E., Healy, R. J., Kiang, N. Y., Koch, D., Lacis, A. A., LeGrande, A. N., Lerner, J., Menon, S., Oinas, V., Garcia-Pando, C. P., Perlwitz, J. P., Puma, M. J., Rind, D., Romanou, A., Russell, G. L., Sato, M., Sun, S., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M.-S., and Zhang, J.: CMIP5 historical simulations (1850–2012) with GISS ModelE2, J. Adv. Model. Earth Syst., 6, 441–477, 2014.

Oleson, K. W., Niu, G.-Y., Yang, Z.-L., Lawrence, D. M., Thornton, P. E., Lawrence, P. J., Stockli, R., Dickinson, R. E., Bonan, G. B., Levis,
25   S., Dai, A., and Qian, T.: Improvements to the Community Land Model and their Impact on the Hydrological Cycle, J. Geophys. Res., 113, doi:10.1029/2007JG000 563, 2008.

Olson, R., Sriver, R., Chang, W., Haran, M., Urban, N. M., and Keller, K.: What is the effect of unresolved internal climate variability on climate sensitivity estimates?, J. Geophys. Res.: Atmos., 118, 1–11, 2013.

Otto, A., Otto, F. E. L., Boucher, O., Church, J., Hegerl, G., Forster, P. M., Gillett, N. P., Gregory, J., Johnson, G. C., Knutti, R., Lewis, N.,
30   Lohmann, U., Marotzke, J., Myhre, G., Shindell, D., Stevens, B., and Allen, M. R.: Energy Budget Constraints on Climate Response, Nature Geosci., 6, 415–416, 2013.

Sato, M., Hansen, J. E., McCormick, M. P., and Pollack, J. B.: Stratospheric aerosol optical depths, J. Geophys. Res., 98, 22 987–22 944, 1993.

Smith, S. J., van Aardenne, J., Klimont, Z., Andres, R. J., Volke, A., and Arias, S. D.: Anthropogenic sulfur dioxide emissions: 1850–2005,
35   Atmos. Chem. Phys., 11, 1101–1116, 2011.

Smith, T. M., Reynolds, R. W., Peterson, T. C., and Lawrimore, J.: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006), J. Clim., 21, 2283–2296, 2008.

Sokolov, A., Schlosser, C., Dutkiewicz, S., Paltsev, S., Kicklighter, D., Jacoby, H., Prinn, R., Forest, C., Reilly, J., Wang, C., Felzer, B., Sarofim, M., Scott, J., Stone, P., Melillo, J., and Cohen, J.: The MIT Integrated Global System Model (IGSM) Version 2: Model Description and Baseline Evaluation, Joint Program Report Series, Report 124, 40 pages, 2005.

Sokolov, A., Kicklighter, D., Schlosser, A., Wang, C., Monier, E., Brown-Steiner, B., Prinn, R., Forest, C., Gao, X., Libardoni, A., and Eastham, S.: Description and Evaluation of the MIT Earth System Model (MESM), manuscript submitted for publication, 2018.

Sokolov, A. P.: Does model sensitivity to changes in CO_2 provide a measure of sensitivity to other forcings?, J. Clim., 19, 3294–3306, 2006.

Sokolov, A. P. and Monier, E.: Changing the climate sensitivity of an atmospheric general circulation model through cloud radiative adjustment, J. Clim., 25, 6567–6584, 2012.

Sokolov, A. P., Forest, C. E., and Stone, P. H.: Comparing Oceanic Heat Uptake in AOGCM Transient Climate Change Experiments, J. Clim., 16, 1573–1582, 2003.

Stevenson, D. S., Young, P. J., Lamarque, V. N. J.-F., Shindell, D. T., Voulgarakis, A., Skeie, R. B., Dalsoren, S. B., Myhre, G., Berntsen, T. K., Folberth, G. A., Rumbold, S. T., Collins, W. J., MacKenzie, I. A., Doherty, R. M., Zeng, G., van Noije, T. P. C., Strunk, A., Bergmann, D., Cameron-Smith, P., Plummer, D. A., Strode, S. A., Horowitz, L., Lee, Y. H., Szopa, S., Sudo1, K., Nagashima, T., Josse, B., Cionni, I., Righi, M., Eyring, V., Conley, A., Bowman, K. W., Wild, O., and Archibald, A.: Tropospheric Ozone Changes, Radiative Forcing and Attribution to Emissions in the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP), Atmos. Chem. Phys., 13, 3063–3085, 2013.