Response to Reviewer #1

P1, L10: "absent an increase" is odd wording.
P1, L12: This statemetn seems redundant to line 4-5 P1, L14-15: What causes these shifts?
P1, L15-16: So if the land surface model has limited effect on temperature evolution, is it updates to the forcings that cause the differences in climate sensitivity estimates? It's not entirely clear what points the authors are trying to convey here. I suggest tightening up the abstract to highlight the significance.

Response: We attribute the observed shifts in the parameter distributions to the changes in model forcings. The land surface model impacts other components of MESM (i.e., carbon fluxes), but in the climate component used here, it has little impact.

Changes: Per these comments, we have revised our abstract to make the summary clearer. An explicit statement has been added that addresses the reason for changes to the distributions.

P7, last paragraph: The authors raise interesting, but somewhat contradictory, points. They state that reducing the number of diagnostics from 3 to 2 has little impact on model parameter estimates, but then go on to state that CS estimates are lower when using 2 diagnostics. Why are the results insensitive to the upper-air diagnostic? Also, the constraint on Kv is not clear. Is there any update since what was shown their previous work (e.g. Libardoni and Forest 2011)? I suggest adding more details to these points to help the reader.

Response: The main reason for omitting the upper-air diagnostic is the significant correlation between the upper-air temperature pattern and the surface temperature pattern as a result of the lapse rate and water vapor feedbacks. Each of these diagnostics reject similar regions of the parameter space for being inconsistent with the observed climate record, thus potentially double counting the same temperature response signal. Removing the upper-air diagnostic removes the risk of bias due to treating it as a statistically independent diagnostic.

There has not been any additional work on constraining Kv between our previous work and this manuscript. Currently, a second publication is in review (Libardoni et al., 2018, ASCMO) that investigates how including additional data in the model diagnostics improves the model parameter estimates. We show there that including additional data improves the model diagnostics and leads to better constraint on Kv. We chose not to incorporate any changes to the model diagnostics in this manuscript to provide a clean comparison of changes resulting from changing only the model version.

Changes: We have cleared up these points by adding clarifying remarks into the manuscript. In Section 3, we include a discussion of why multiple diagnostics are preferable and why independent diagnostics of model performance are important. Further, we provide a reference study that addresses the correlation of the surface and upper-air diagnostics. This point is highlighted further in Section 4 when discussing the changes in the PDFs resulting from the reduction in the number of diagnostics. Contradictory language regarding the size and significance of the changes when moving from three to two diagnostics has been removed for clarity.

P8, L10: Can you show a plot of the ECS pdf for IGSM and MESM for comparison?

Response: For each of the distributions derived from the individual surface temperature datasets, we plotted the marginal PDFs for the full IGSM and MESM ensembles. In all five cases, the same changes are observed: higher climate sensitivity, nearly unchanged ocean diffusivity, and weaker negative aerosol forcing.

Changes: We have added this figure (Figure 5) and supporting text in Section 4.

P9, L3-4: How do these new estimates of net aerosol forcing compare with other recent estimates?

Response: For EMICs like MESM, the net aerosol forcing is a model-specific parameter, making a clear comparison between studies and direct observations challenging. For example, the aerosol forcing pattern may account for different model forcings and be defined for different time periods. For example, while Andronova and Schlesinger (2001) scale the natural and anthropogenic aerosol direct and indirect forcings by adjusting the amplitude in 1990, the aerosol parameter in Knutti et al. (2002) is scaled in 2000 and represents the indirect aerosol effect and any other forcing not explicitly represented in the model. With these differences in mind, estimates of aerosol forcing from energy balance models and EMICs fall in the ranges -1.3 to -0.54 Wm^{-2} (Andronova and Schlesinger, 2001), -1.2 to 0 Wm^{-2} (Knutti et al., 2002), -1.53 to -0.33 Wm^{-2} (Kriegler, 2005), -0.83 to -0.19 Wm^{-2} (Libardoni and Forest, 2011), and -1.7 to -0.4 Wm^{-2} (Skeie et al., 2014).

Changes: No major changes have been made to the manuscript in response to this comment. We have intentionally left the comparison of our parameter estimates with other groups for our other studies. This is done to place the emphasis of this work on setting the baseline for how the change in the forcings and model impact the parameter and TCR estimates.

P10: L14: I'm a little unclear how ocean diffusivity fits in with the analysis. Why did the old ensemble cut of high values of Kv? It is also relatively insensitive to the model updates compared to aerosol forcing and equilibrium climate sensitivity. Why is this? I recommend the authors streamline the results and discussion sections to include a summary of key points about each model parameter, the constraints and model sensitivities, and physical reasoning for the differences.

Response: Kv fits into the analysis because all three model parameters are estimated jointly, with the marginal PDFs calculated by integrating the joint PDF over the other two model parameters. Thus, changes due to the model and forcings can impact any of the three marginal distributions. As we point out in the edited manuscript, physical explanations for the changes in the ECS and aerosol distributions are more accessible than an explanation for Kv. However, because all three parameters are estimated together, changes in the other two parameters can impact our Kv estimates.

In both ensembles, values of Kv outside the of range of values sampled are assigned zero probability. This meant assigning zero probability for regions greater than 5 cms$^{-1/2}$ for the IGSM ensemble and 8 cms$^{-1/2}$ for the MESM ensemble. From the full MESM ensemble, we find non-zero, although small, probabilities of Kv between 5 and 8 cms$^{-1/2}$. By accounting for the extra mass in the tail regions for the MESM ensembles, the Kv quantiles are pulled towards higher values.

Changes: We have added text to the manuscript addressing the points above. Beginning on Page 11, Line 20 of the revised manuscript, we provide discussions of each model parameter as suggested by the reviewer. As part of the discussions, we give physical explanations to support the changes we observe in the parameter distributions. We further strengthen the discussion by including the benefits and challenges of estimating the parameters together. In particular, we highlight that the joint distribution allows us to identify correlations amongst the parameters, but also makes the attributing the changes in a single parameter to one cause less straightforward.

An explicit explanation for the cut-off of high Kv values is given beginning on Page 11, Line 7 of the revised manuscript. The insensitivity of the Kv distribution is addressed in the paragraph devoted to the parameter (Page 12, Line 6).

P12, L7: Why choose a third-order polynomial here? Is there sensitivity in the fits to the functional form? Would you expect similar results in terms of model differences using a 2nd order polynomial?

Response: The third-order polynomial was chosen for consistency with previous work to provide the most direct comparison possible between the surfaces derived for IGSM and MESM. In offline tests, we derived additional surfaces for first-, second-, and fourth-order polynomial fits and compared them to the TCR and SLR values calculated directly from the transient simulations. The first-order approximation leads to an unsatisfactory fit with gradients of TCR and SLR in the Kv direction that are too weak. The second-order fit produces curvature in TCR and SLR contours that are inconsistent with those calculated directly from the transient simulations. In particular, the 1.5 ^{o}C contour for TCR using the second-order fit suggested that for a single Kv value, two different ECS values could be used. Further, the second-order fit shows that sea level rise greater than 14 cm is possible within the sampled domain, whereas none of the transient simulations had SLR that high. The third- and fourth-order fits both showed good agreement with the simulated results but were not without their flaws. The third-order fit showed some error in the 1.5 ^{o}C TCR contour, where the fourth-order fit led to regions of SLR greater than 14 cm within the domain. Improving this fit is a potential avenue of future research.

Changes: We mention the reason for choosing the third-order fit and that other fits were explored in the revised manuscript (Page 14, Line 11).

P12, L24-25: The authors state that the shift towards higher transient climate response is driven by higher climate sensitivity in MESM, but there is not enough explanation in my opinion as to why there is a larger CS in MESM compared to previous versions, how they compare (e.g. posterior distributions), and to what extent the updated forcings play a role.

Response: Through the points made to previous comments and the changes made to the manuscript, we believe that this has been more clearly addressed. Looking at the response surfaces, for any Kv value, an increase in ECS leads to larger TCR. Thus, given a constant Kv distribution, shifts towards higher ECS result in a shift towards higher TCR. With the relatively small changes in the Kv distribution from the subsampled MESM ensemble (see Table 2 of the manuscript), we find the assumption of constant Kv distribution needed for this argument justifiable.

Changes: As mentioned in the responses/changes to the comments above, we have added a discussion for each parameter that explains how changes to the model forcings can lead to the shifts observed in the marginal distributions. Furthermore, the addition of Figure 5 provides a direct comparison of the posterior distributions for each parameter derived from IGSM and MESM.

P12-13: The conclusions provide a nice summary of the paper's key points. I suggest expanding the results section to include more in-depth discussion along these lines.

Response: As noted above, we have expanded the results section to provide more in-depth discussions of the reasons for the changes in the parameter and TCR estimates.

Changes: Specific changes to the results section (Section 4) are given in responses to earlier comments.

Response to Reviewer #2

1) I miss a description of the basic components and parameterizations of the model in the method section 3. I miss a section that describes model spin-up and the setup for the different model simulations, including external forcing factors. Further, it is not evident from the description why the model is called "Earth System Model". For example, are biogeochemical cycles included? Does dynamic vegetation affect albedo? Is it an ESM or rather an Earth System Model of Intermediate Complexity? I also miss a brief description of the metric used to compare model and data and how they are used to derive probability distribution. It is not sufficient to refer the reader to the literature (Libardoni and Forest 2011).

Response: The MIT Earth System Model is an integrated model with sub-models for the atmosphere, ocean, land surface, atmospheric chemistry, ocean biogeochemistry, and the terrestrial ecosystem. When all of these sub-models are turned on, the model is set up as an Earth system model. However, under that set up, the model is too computationally expensive to be used for probabilistic studies of the model parameters like what is presented in this study. Turning off all components of the model except the atmospheric, ocean, and land surface models simplifies the model to an EMIC that can be used for probabilistic estimates of the model parameters investigated in this work.

Changes: A more detailed presentation of the EMIC (climate component of MESM) has been added to Section 2. In that discussion, we describe the model components of the EMIC, the input forcings, and the model parameters. In the discussion of the model parameters, we describe how each of the three are adjusted and how the model is being modified to make the changes.

In Section 3, we have included a summary of the methods used to derive the probability distributions. We present the goodness-of-fit statistic used to evaluate the model. This statistic is the weighted sum-of-square residual between the model output and observed climate record for a given diagnostic. A reference to the likelihood function is provided and we explain how the joint distribution is calculated from the goodness-of-fit statistic.

2) Section 3: The authors vary three parameters – ocean diffusivity, an aerosol forcing scaling, and the strength of the cloud feedback determining ECS and constrain the models with two parameters.

2a) There is little information in the method section what these parameters specifically influence. The aerosol forcing scaling is unclear. Does this mean that all aerosol forcings are lumped together and scaled with a constant time invariant factor? How are different uncertainties applying to different aerosol classes (e.g. sulfate versus soot) considered or not and what is the justification for this approach. Please discuss caveats related to your assumption of a scaling factor.

Response: In the description of the model parameters that was added to Section 2, we describe what each of the parameters influence. For completeness, we summarize them again here. ECS is modified by adjusting the strength of the net cloud feedback in the model. More specifically, a number of simulations where CO_2 concentrations have been doubled and the system brought to equilibrium have been run for different values of the cloud adjustment. These are used to provide a lookup table which gives the cloud adjustment needed for a specific ECS. Ocean diffusivity is defined by a latitude-dependent pattern based off of tritium mixing into the deep ocean. Kv represents the global mean value and specific diffusivity values are calculated by scaling the spatial pattern by the same value at all latitudes to achieve the desired global mean value.

The forcing due to all aerosols except sulfate are held constant during historical simulations and the sulfate aerosol is parameterized through adjustments to the surface albedo based on changes in the historical emissions of SO_2. The historical emissions have both spatial and temporal components, with the aerosol parameter setting the amplitude of the pattern in the 1980s. Adjusting the forcing in this manner is not without its drawbacks. As the only adjustable forcing component in the model, this forcing pattern also represents an estimate of all other forcings not included in the model. Thus, this is not a pure estimate of the aerosol forcing.

Changes: We have added a description of the model parameters, what they represent, and how they are adjusted into Section 2.

2b) Effective ocean diffusivity is a very loose term. Is this diapycnal, vertical or horizontal diffusivity or does the parameter refer to the diffusivity associated with Gent-McWilliams parameterization? The subscript v of Kv points to vertical diffusivity. I would hope that this parameter reflects diapycnal diffusivity as diapycnal diffusivity co- governs ocean overturning strength and thus surface-to-deep heat transport. In any case, I am puzzled about the range sampled. Diapycnal diffusivity in coarse resolution, dynamic ocean models is typically of order 0.1 10-4 m2 s-1. Here diffusivity is varied in steps of 1 10-4 m2 s-1 and a very wide range up to 64 10-4 m2 s-1 is used. The upper value is even much larger than applied in classical box-diffusion models (1-2 10-4 m2 s-1); in box-diffusion models the entire vertical transport (mixing, advection, convection) is parameterized by diffusion only. What is the justification for this large sampling range? As a minor point, please use SI units for diffusivity. Further, I though Gent-McWilliams parameterization is included in the MIT model. If yes, why is the Gent-McWilliams diffusivity not varied or is this parameter linked with the "effective diffusivity"?

Response: In the ocean model, horizontal heat transport is prescribed by the Q-flux calculation and the vertical mixing of heat into the deep ocean is prescribed by the spatial diffusivity pattern and scaled by Kv as discussed above. As Kv represents the mixing of heat into the deep ocean by all processes, it is greater than diapycnal diffusion values found in the sub-grid scale parameterizations of dynamic ocean models.

A wide range of Kv values was sampled to simulate many possible climate states, including those with very strong vertical ocean mixing. Similarly, wide ranges were also chosen for climate sensitivity and the aerosol forcing. For the most part, runs with extreme values of any parameter were rejected for being inconsistent with the model diagnostics. In the case of Kv, this supports the claim that such high values should not have been sampled to begin with. The penalty paid for this over sampling of the parameter ranges is a misallocation of computing resources.

Changes: We have added text to the manuscript in Section 2 to address these concerns. We have clarified that a mixed-layer ocean model is used, that Kv represents the mixing due to all processes, and how the mixing is spatially distributed.

2c) ECS is typically used to abbreviate Equilibrium Climate Sensitivity. Here, an effective climate sensitivity is introduced and termed ECS. What represents this effective climate sensitivity?

Response: We mistakenly expressed ECS as effective climate sensitivity, when it is, in fact, equilibrium climate sensitivity. The lookup table for ECS is derived from runs brought to equilibrium, so that any equilibrium climate sensitivity can be obtained through the proper adjustment of the cloud feedback.

Changes: All references to effective climate sensitivity have been changed to equilibrium climate sensitivity.

3) Section 3: I question somewhat the application of only two observational metrics to constrain ECS, TCR, and sea level rise. Namely, pattern of surface air temperature change and "linear" ocean heat uptake are used as constraints by the authors. In my opinion, there is a lack of observational constraints to probe the timescales of deep ocean overturning (e.g. 14C). Thus it appears not surprising that the diffusivity parameter remains not well constrained. There is also a lack of metrics to probe the spatial pattern of heat uptake. This is particularly important as the thermal expansion coefficient varies by almost an order of magnitude in the ocean. Thus it matters, where the heat is taken up to estimate sea level rise. As another focus of the study is on TCR, it would also be nice to invoke additional metrics on thermocline ventilation as for example available by observation-derived fields of CFCs and bomb-produced 14C.

Response: Given the mixed-layer ocean model that is coupled to the atmosphere, we are somewhat limited to the diagnostics that can be used to evaluate the ocean system. As further explained above, the vertical mixing pattern is prescribed with latitudinal dependence, but also fixed throughout the run. The vertically-integrated horizontal heat transport is also prescribed based on offline Q-flux calculation. With these patterns fixed, incorporating ocean diagnostics with spatial dependence is not feasible at this time.

As an aside, developing additional model diagnostics to constrain estimates of the model parameters, TCR, and sea level rise is a task that should be undertaken and is of interest to the authors. Care should be taken to ensure that these metrics are independent of each other or that steps be taken to account for the correlation between metrics. However, developing such metrics is beyond the scope of this work.

Changes: We have added a discussion to Section 3 that explains why we chose the model diagnostics that we did (Page 5, Line 24). We include references to other work that helps justify our choices.

4) Page 5 to page 7, results, The description of the difference in input forcing is useful, but in my opinion misplaced. Solar and ozone forcings are model drivers (or forcings) and distinct from a particular model version. These forcings should be described in the method section where the simulations and the applied external forcings are to be described.

Response: While we recognize that the presentation of the model forcings may be better placed in the methods section, we believe that keeping them in the results section is justifiable. The interpretation of the new forcings and their direct application to the model parameters are in themselves a finding in this study. Much of the reasoning for the shifts in the parameter estimates centers around these changes in the model forcings and are essential to the explanation of the results. In our opinion, keeping them together is appropriate.

Changes: No major changes have been made to the location of this discussion. However, we have further clarified that only time variant changes to the forcings impact the historical simulations (Page 7, Line 4).

5) P6, line 3ff; Q-flux adjustment: Does this mean that the authors apply temperature flux correction to their model? This should be explained in the method section.

Response: We have addressed the Q-flux adjustment, which represents the vertically-integrated heat flux, earlier in this response.

Changes: An explanation of the Q-flux adjustment has been added to the manuscript and discusses how it is related to horizontal heat transport in the ocean (see Section 2).

6) Section 4: I miss a figure comparing the modelled pattern of the median (or mean or best-guess version) with the observed pattern of surface air temperature change and similar for the global ocean heat uptake and its spatial pattern (and may be for upper air temperature) to illustrate how well the model is able to capture the observations.

Response: A figure comparing the model output to the observed surface pattern used in our diagnostic does not yield a clean comparison. As a result of weighting the model-to-observation residuals by the noise covariance matrix, the temperature patterns are rotated into a coordinate space defined by a set of orthogonal basis functions defined by the internal variability estimate. Thus, any attempt to compare the model output and observations in the unrotated space does not give a fair representation of an individual model run's fit to the observed record. A fairer assessment of the model fit to the observations is obtained by comparing the global mean temperature time series.

Given the fixed mixing pattern used in the ocean model, the spatial pattern of heat uptake does not vary between the model simulations. Only the magnitude changes, making a comparison between the model and observations for individual runs redundant.

Changes: We have included a figure where the global mean surface temperature of each of the 1800 model runs is shown, along with the observed time series for each of the five datasets used in this study. We have also highlighted the model runs where the parameter settings most closely match the median values from the marginal distributions derived from each of the surface datasets. All anomalies are calculated based off of the 1906-1995 climatology used in the surface diagnostic.

Similar to the global mean surface temperature results, we also include a figure to show the spread in the ocean heat content linear trends calculated from our ensemble. We plot a histogram of the calculated trends from each individual run, while also showing the observed trend and highlighting the runs with parameter settings closest to the distribution medians.

A discussion of these results begins on Page 12, Line 11. In this discussion, we explain the model's ability to match the observations and reasons for mismatches between the model output and the historical record.

7) Page 12, line 7: How well does the polynomial fit represent the model results?

Response: In general, the polynomial fit represents the model results quite well, but is not without error. In our response to Reviewer #1, we discussed using first-, second-, and fourth-order fits, as well as some of the errors associated with the third-order fit.

Changes: We mention the reason for choosing the third-order fit and that other fits were explored in the revised manuscript (Page 14, Line 11).

8) Page 12, line 14: Why is the PDF for the TCR not directly estimated from the 372-member ensemble? Does the fitting add additional uncertainties to the procedure of estimating TCR?

Response: It is possible to directly estimate the PDF for TCR from the 372-member ensemble. Doing such would represent estimating TCR from a joint distribution where all values of ECS and Kv are equally likely to occur. In other terms, the ECS-Kv two-dimensional PDF would be uniform for all pairs within their respective domains. We have shown in this study that ECS and Kv are not uniformly distributed and that some pairs are more likely to occur than others. Drawing from this more realistic distribution yields a probability-weighted sampling of parameter pairs from which to estimate TCR.

Using the polynomial fit adds additional uncertainty to the procedure of estimating TCR by introducing interpolation error. As described in the response to Reviewer #1, the polynomial fit is not an exact match to the model results, and any error in the estimation propagates as an error in the TCR distribution. However, running the transient simulation for each ECS-Kv draw from the Latin Hypercube Sample is infeasible, so the fit is required to estimate TCR for the pairs where there is no corresponding run.

Changes: No major changes have been made to the manuscript.

9) Discussion and conclusion: While the authors suggest that their approach should serve as a template for other groups, they fail to mention that similar, and sometime much more comprehensive approaches of parameter calibration, have been undertaken by other groups. They also fail to compare their estimate of TCR and ECS with published estimate and to put their findings in the context of the wider literature. See for example, Collins et al., IPCC, 2013 for the most recent assessment of TCR and ECS values by IPCC. Of course there are recent updates of these estimates and there are also many other studies that determine model parameters such as vertical ocean diffusivity. Examples that come immediately in my mind are Holden et al., Clim. Dyn., 2010, Richardson; Nat. Clim.Change, 2016, Schmittner et al., GBC, 2009, Steinacher et al., Science, 2013 or Steinacher and Joos, Biogeosciences 2016. It is the task of the authors to identify the recent literature to provide a relevant discussion.

Response: In both the abstract and the penultimate paragraph of the introduction, we state that the point of the study is to assess how the changes in the model can impact the distributions. The paper is not intended to discuss how the results compare with recent estimates of ECS or TCR distributions or specific methodologies for estimating probability distributions. We think the introduction's text reflects this and is included here.

" 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 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 studies in order to separate their impacts from changes due to the model update alone”

The future work will provide the appropriate discussion of other studies as suggested by the reviewer while this work only documents the impact of changes in the model framework.

We are aware that other approaches exist and have avoided stating that our parameter estimation methodology is better. We do think this approach can serve as a template for testing how new versions of models can directly impact parameter estimates and that such tests should be documented in a similar fashion.

Changes: Throughout the manuscript, we have made it clearer that we are only conducting the baseline test of the model in this study. Examples include Page 3, Line 3 and Page 12, Line 11 of the revised manuscript. These areas defer discussion of results that don’t compare the IGSM estimates to the MESM estimates to future/concurrent work.

P1, Line 22: typo: sensitivity

Changes: We have fixed this typo in the manuscript.

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~~ and updated. To provide transparency in the model development, we perform a baseline evaluation ~~of the newest version~~ by comparing model behavior and properties in the newest version to the previous model version. In particular, ~~the impacts~~ changes resulting from updates to the land surface 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, the rate of ocean heat uptake, and the net anthropogenic aerosol forcing are systematically varied. By comparing model output to observed patterns of surface temperature changes ~~,~~ and 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 climate simulations where all model forcings are ~~held fixed , absent an increase in~~ fixed and carbon dioxide concentrations are increased at the rate of 1% per year. From these runs, we derive ~~a response surface~~ response surfaces 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 response to the new forcing suite. The~~ when using the new model and that the climate response surfaces are relatively unchanged between model versions~~, indicating that the updated~~ . Because the response surfaces are independent of the changes to the model forcings and similar between model versions with different land surface models, we suggest that the change in land surface model has limited impact on the temperature evolution in the model. Thus, we attribute the shifts in parameter estimates to the updated model forcings.

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 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~~ sensitivity (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 Earth ~~System Models of Intermediate Complexity~~ 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~~ over a range of climate sensitivity values in addition to adjusting other relevant factors, such as the ~~ocean diffusivity~~ rate of ocean heat uptake 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 ~~then presented~~ given 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 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 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~~ TCR has been shown to depend 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~~ climate component of the Massachusetts 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 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 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 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~~ studies in order to separate their impacts from ~~those due to changes to the model framework~~ changes due to the model update 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~~the rate of ocean heat uptake.

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 Section 5.

2 Model

The ~~coupled atmosphere-ocean-land model~~ climate component 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~~ and is an Earth system model of intermediate complexity. It consists of a zonally-averaged atmosphere, zonally-averaged land model, and a mixed-layer anomaly diffusing ocean model. The mixed-layer ocean model includes specified vertically-integrated horizontal heat transport by the deep ocean, a so-called "Q-flux". This flux has been calculated from a simulation in which sea surface temperatures and the sea-ice distribution were relaxed toward their present-day climatology. Heat mixing into the deep ocean is parameterized by the diffusion of the difference of the temperature at the bottom of the seasonal thermocline from its value in a pre-industrial climate simulation (Hansen et al., 1984; Sokolov and Stone, 1998). Since this diffusion represents the cumulative effect of heat mixing by all physical processes, the values of the diffusion coefficients are significantly larger than those used in the sub-grid scale diffusion parameterizations in ocean global circulation models. The spatial distribution of the diffusion coefficients used in the diffusive model is based on observations of tritium mixing into the deep ocean (Hansen et al., 1988).

The radiation code takes into account major greenhouse gases (H_2O, CO_2, CH_4, N_2O, CFCs, and O_3) and multiple types of aerosols (e.g. SO_2, black and organic carbon). In historical climate simulations, loading for all aerosols except sulfate are kept at their default values. The forcing due to sulfate aerosol is parameterized through changes in surface albedo using historical data on SO_2 emissions. Historical climate simulations are initialized from conditions obtained from a long equilibrium simulation for 1860 conditions.

Three model parameters that impact the climate system response are easily modified in MESM. These parameters are the equilibrium climate sensitivity (ECS), the effective ocean diffusivity (K_v), and the net aerosol scaling factor (F_{aer}). ECS is changed by adjusting the strength of the cloud feedback at different levels in the model (Sokolov, 2006; Sokolov and Monier, 2012). The adjustment required for a specific ECS is obtained from a lookup table derived from model simulations with different feedback strengths where CO_2 concentrations have been doubled and the climate system allowed to reach equilibrium. K_v represents the global mean ocean diffusion coefficient in the mixed-layer ocean model. The global mean diffusivity is adjusted by scaling the spatial diffusivity pattern by the same factor at all locations. A lower global mean diffusivity implies slower mixing of heat into the deep ocean and a higher global mean diffusivity implies faster mixing. The albedo adjustment used for the sulfate aerosol forcing is prescribed by a latitude-dependent pattern that differs over land and ocean (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.

We now highlight two major updates made between the current version of MESM and its predecessor. The first update 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 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 ~~Intergovermental~~ 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 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 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 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 patternis scaled to obtain the desired value. The anthropogenic aerosol forcing 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

~~In this section, we present an outline of~~

In this section, we present an outline of the methodology used to derive the joint probability distribution function (PDF) 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), which we briefly summarize here. To derive the PDFs, we compare output from each model simulation to time series of observed climate change. A given model run is evaluated through the use of a goodness-of-fit statistic

$$r^2 = (\mathbf{x}(\theta) - \mathbf{y})^T \mathbf{C_N^{-1}} (\mathbf{x}(\theta) - \mathbf{y}), \tag{1}$$

where $\mathbf{x}(\theta)$ and $\mathbf{y}$ are n-length vectors of model output for a given set of model parameters and observed data, respectively, and $\mathbf{C}_N^{-1}$ is 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) w two notable changes~~inverse of the noise-covariance matrix. In its simplest form, the r^2 statistic is the weighted sum of squares residual between the model simulation and the observed pattern. The weights applied to the residuals are estimated from the unforced climate variability in a fully coupled, three-dimensional model and represent the observed patterns we would expect in the absence of external forcings. In Libardoni and Forest (2011), surface temperature, upper-air temperature, and global mean ocean heat content patterns were used to evaluate model performance. We note that the definition of r^2 presented here is different than the coefficient of determination for the goodness-of-fit of a linear model. In a linear model, high values of r^2 indicate a good fit to the model. In our weighted sum, low values of r^2 indicate a good fit between the model output and the observations.

The goodness-of-fit statistics for each pattern used to evaluate the model are converted to a PDF using the likelihood function function described in Libardoni and Forest (2011) and modified by Lewis (2013). Through an application of Bayes' Theorem,

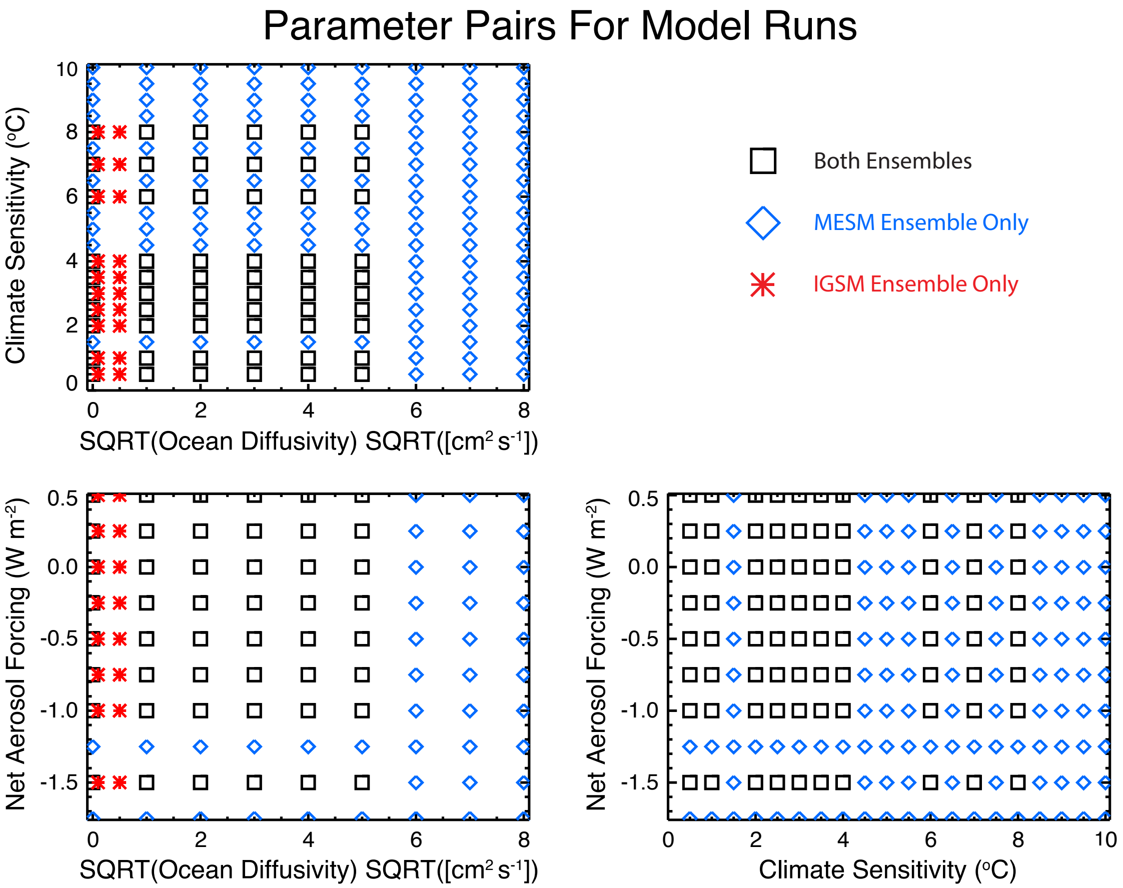

Figure 1. Parameter pairings where the models have been run. Points in black are common to both the IGSM and MESM ensembles. Blue points are unique to the IGSM ensemble and red points are unique to the MESM ensemble.

the individual likelihoods are combined to derive a joint PDF for the three model parameters. As in Libardoni and Forest (2011), we apply an expert prior to ECS and uniform priors to K_v and F_{aer}. Marginal probability distributions for individual parameters are calculated by integrating the joint PDF over the other two parameters.

We make two changes to the methodology of Libardoni and Forest (2011) to derive PDFs using MESM simulations. First, we run the model for ~~θ~~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 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~~ In general, independent temperature patterns should be used to evaluate model performance because they rule out different regions of the parameter space for being inconsistent with the observed climate record. In particular, Urban and Keller (2009) show that surface temperature and ocean heat content time series provide good constraints on model estimation. Further, Lewis (2013) shows upper-air temperatures to be highly correlated with ~~the surface temperature~~ surface temperature via the lapse rate and water vapor feedbacks. For these reasons, we now omit the upper-air temperature diagnostic. The removal of the upper-air diagnostic ~~(Lewis, 2013). This~~ leaves two 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. Five different joint PDFs are derived by combining the likelihood from the ocean diagnostic with the likelihood derived from each of the individual surface temperature datasets.

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 ~~θ~~θs common to both ensembles are used. Fourth, we evaluate how well the model captures the observations by comparing model output from the MESM ensemble to the observed climate record. 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. When comparing the forcing time series, only differences in the changes relative to 1860 impact the historical simulations. Time invariant differences are accounted for in the offline Q-flux and initial condition calculations, but differences in the changes are not. In Figure 2, we show the old and new solar forcing time series. We see that the biggest 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 Wm^{-2} until 1920, but then increases towards 5.0 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., 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,~~

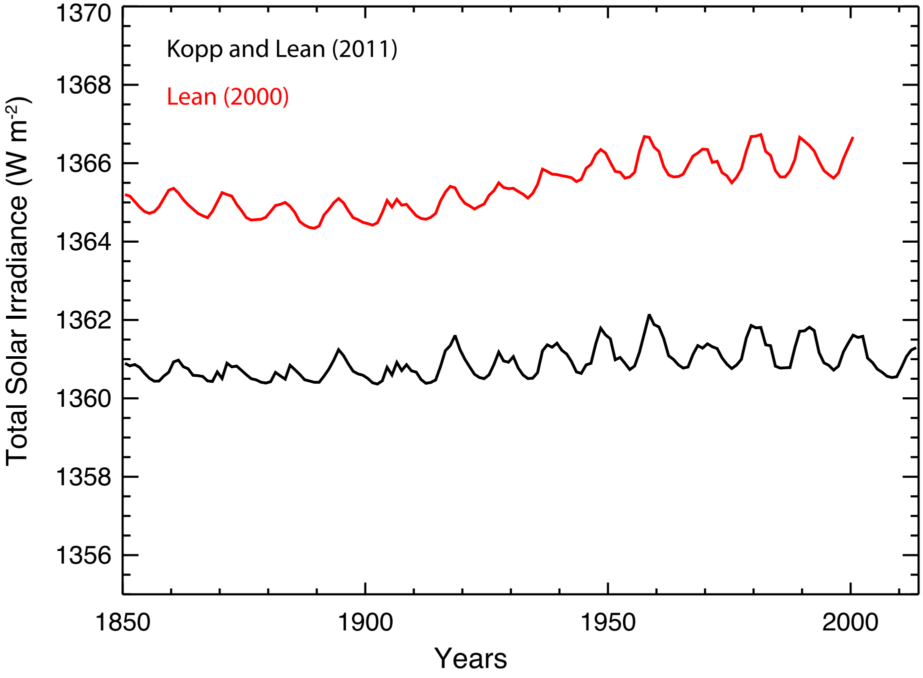

Figure 2. Annual mean total solar irradiance. The bias between the Lean (2000) and Kopp and Lean (2011) datasets leads to a reduction in radiative forcing in the new forcing suite.

~~the growth of the low bias means that~~ growth of this low bias introduces a weakening of 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 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 compared to the IGSM dataset. Differences in stratospheric ozone concentrations remain relatively constant until 1950, but then 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 AC&C/SPARC dataset both contribute to a weaker radiative forcing over the historical period in the new suite of forcings.

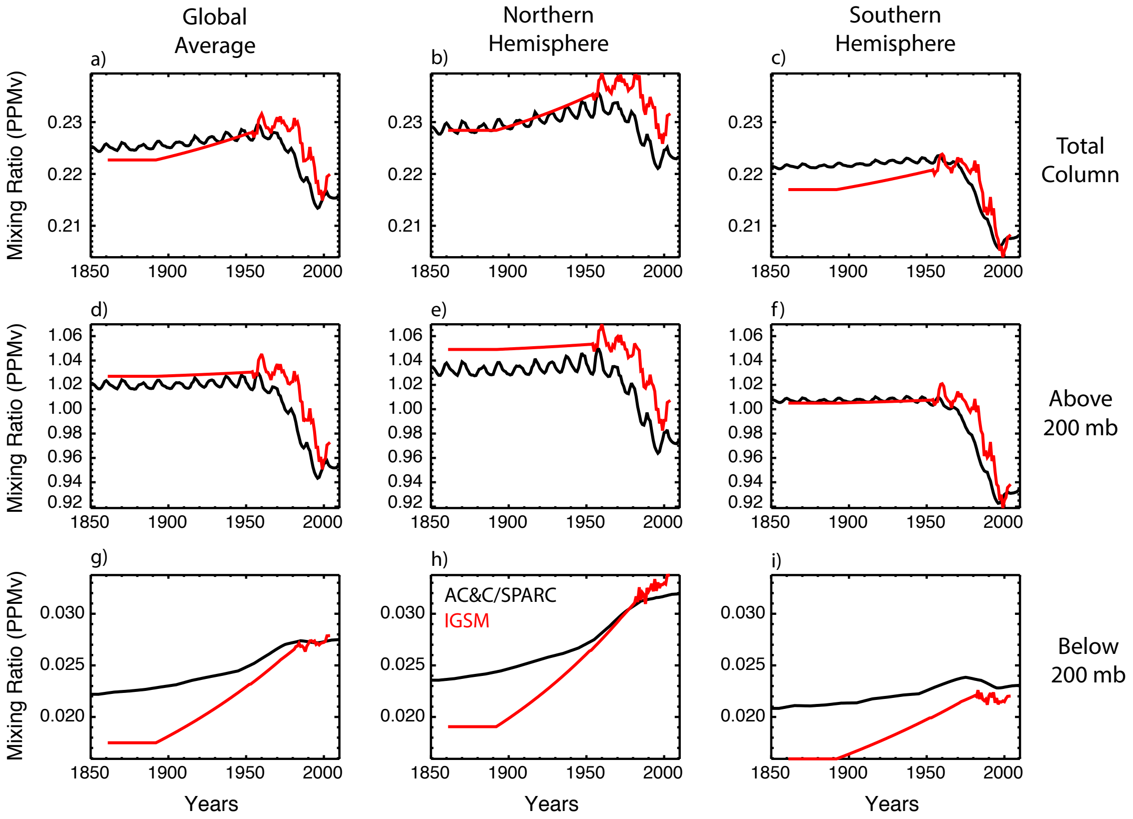

Figure 3. Ozone concentration in the old IGSM time series (red) and the Cionni et al. (2011) AC&C/SPARC concentrations (black). (a-c) Annual mean ozone mixing ratio in the total column in the global average (a), northern hemisphere (b), southern hemisphere (c). (d-f) As in (a-c) but for the average above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

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. As noted in Section 3, the surface and upper-air temperature diagnostics are highly correlated. As a result, they reject similar regions of the parameter space for being inconsistent with the observed climate record. Thus, those regions are rejected twice, while regions inconsistent with the ocean heat content diagnostic are rejected only once. Multiplying the Bayesian likelihood estimate by the same pattern twice leads to a potential bias in the distributions towards regions that are consistent with the surface temperature 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~~ leads to slight changes in the parameter estimates (Table 1). We only 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 aerosol estimates

Table 1. 90-percent confidence intervals for climate sensitivity (ECS) and net aerosol forcing (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)

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 and conservatively removes the risk of double counting the surface signal, 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 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~~ marginal distributions for each parameter in Figure 4 and observe significant differences between ~~distributions~~ those derived using 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 K_v derived using the GISTEMP datasets are still unconstrained with upper tails

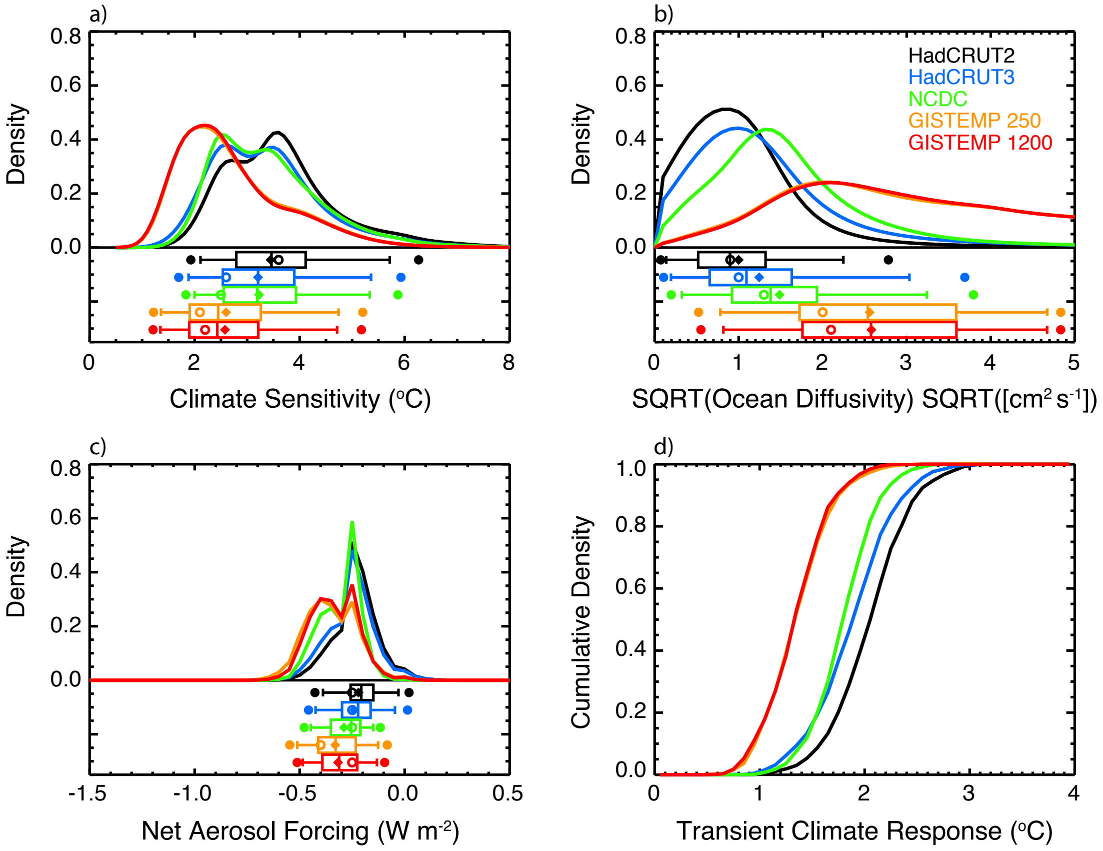

Figure 4. Marginal probability distribution functions and TCR cumulative distribution functions derived from MESM simulations using the HadCRUT2, HadCRUT3, NCDC, GISTEMP 250, and GISTEMP 1200 surface temperature datasets as observations~~.~~: (a) ECS, (b) K_v, ~~and~~ (c) F_{aer}. Whisker plots indicate 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 drawn from the joint parameter distributions and the TCR ~~$(ECS, \sqrt{K_v})$~~ functional fit.

extending to the edge of the parameter domain, 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 the 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}.

~~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 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$_2$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~~ these distributions in Table 2 and see that there is very little sensitivity when the ensembles are subsampled. Across all datasets, the 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~~ we assign a probability of zero to regions of the parameter space that have not been sampled. Thus, for the subsampled MESM ensemble, we assign a probability of zero for ~~K_v~~$\sqrt{K_v}$ ~~sampled in the MESM ensemble~~between 5 and 8 cm s$^{-1/2}$, but the likelihood function does not evaluate to zero in this region when using information from the full ensemble. As a result, the full ensemble does not artificially cut off the distribution ~~for values of~~ at $\sqrt{K_v}$ ~~greater than~~ equal to 5 cm s$^{-1/2}$ ~~. Thus, the~~ and leads to higher upper bounds on the distributions. Knowing this, we can conclude from the similarity between distributions derived from the full and subsampled ensembles that the differences we observe between the ~~old and new~~ IGSM and MESM ensembles are due to the differences between the model and forcing themselves~~and~~, not the increased density of model runs.

To further demonstrate the total effect of changes to the model, forcings, and ensemble design, we compare the marginal distributions derived from the full IGSM and MESM ensembles using each surface temperature dataset (Figure 5). For all five datasets, we observe shifts towards higher climate sensitivity, slightly higher ocean diffusivity, and weaker aerosol forcing, consistent with our previous discussion. Further, we demonstrate that the higher ocean diffusivities using the MESM ensembles are the result of not assigning zero probability for $\sqrt{K_v}$ between 5 and 8 cm s$^{-1/2}$. This is clearly evident in the distributions derived using the GISTEMP datasets (Figure 5b), where the IGSM distributions drop sharply to 0 at $\sqrt{K_v}$ equal to 5 cm s$^{-1/2}$.

~~To estimate~~ Because the parameters are estimated jointly, identifying the causes for specific changes in the marginal distributions are not always straightforward. With this caveat, we now present reasons for the observed changes in the parameter distributions. We begin with F_{aer}. As discussed earlier in this section, changes to both the solar and ozone forcing lead to a reduction in their contribution to the global radiation budget. Additionally, there has been a weakening of non-CO$_2$ greenhouse gas forcing introduced by the new radiation code in MESM. These factors result in a decrease in the net radiative forcing on the planet. With the surface temperature and ocean heat content diagnostics unchanged, the same temperature patterns need to be matched despite the weaker net forcing. One adjustment to the climate system that can help accomplish the matching is to

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 ($^\circ$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

increase the forcing from another term in the energy budget. Of the three model parameters, F_{aer} is the only one that directly changes the radiative forcing, and we thus observe the shift towards less negative aerosol forcing.

An explanation similar to that used for the aerosol distribution can be applied to explaining the observed shifts in the climate sensitivity distribution. In its most basic sense, climate sensitivity is a temperature change per unit forcing. When holding the temperature patterns fixed, the change in temperature is a constant. When explaining the aerosol distribution above, we implicitly fixed the climate sensitivity, requiring the aerosol forcing to be less negative to keep the net forcing constant. However, if we fix F_{aer}, the same temperature change needs to be realized with the weaker forcing due to the changes in the solar and ozone forcings. This implies a higher climate sensitivity is required and explains the shifts we observe in the ECS marginal distribution.

In practice, the model parameters are not independent of each other and can change simultaneously. Many combinations of higher climate sensitivity and weaker aerosol forcing lead to similar agreement with the observed temperature record. This suggests a correlation between these two parameters and highlights a strength of estimating the joint PDF for the model parameters: the identification of relationships between the model parameters. However, these relationships also highlight the challenge in attributing changes in a single parameter to a specific cause.

Unlike the climate sensitivity and aerosol forcing distributions, a clear physical explanation for the observed changes in the K_v distribution is more difficult to identify. One reason for this difficulty is the relative insensitivity of the K_v distribution

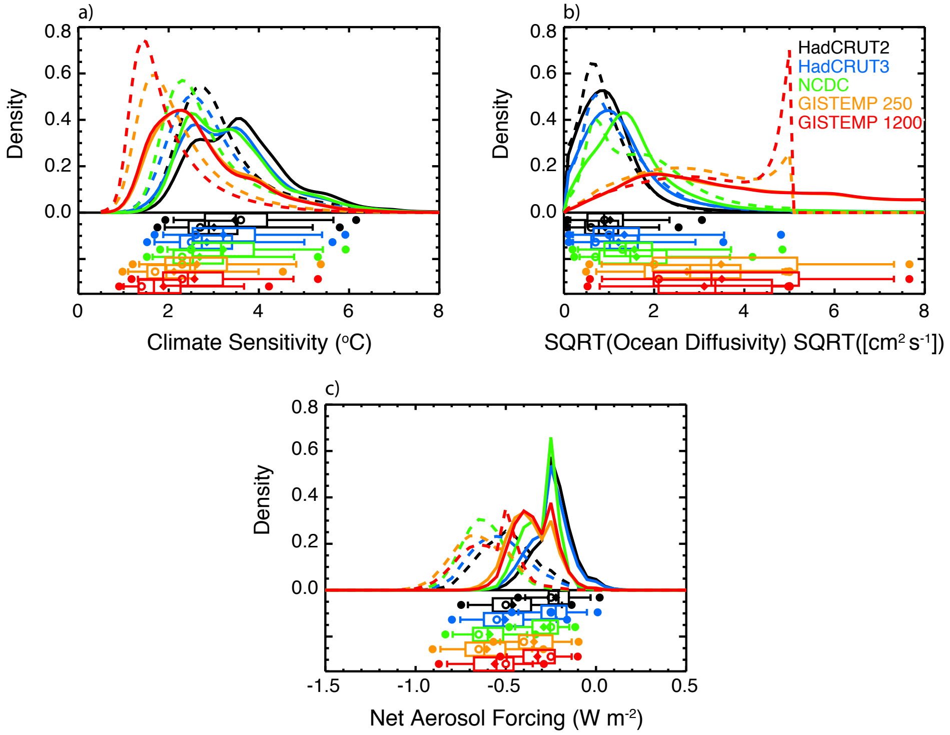

Figure 5. ~~Model response surfaces for~~ Marginal probability distribution functions derived from the full IGSM (~~a~~dashed) ~~TCR~~ and MESM (solid) ensembles using the HadCRUT2, HadCRUT3, NCDC, GISTEMP 250, and GISTEMP 1200 surface temperature datasets as observations: (a) ECS, (b) ~~thermosteric sea level rise~~K_v, (c) F_{aer}. ~~Contours~~ Whisker plots indicate boundaries for the ~~MESM response surfaces are shown~~ 2.5-97.5 (dots), 5-95 (vertical bar ends), 25-75 (box ends), and 50 (vertical bar in ~~black~~ box) percentiles. Distribution means are represented by diamonds and ~~contours for the IGSM surfaces~~ modes are ~~shown in red~~represented by open circles. ~~Differences between~~ For a given dataset, the ~~fits are also shown (c~~ top and ~~d)~~bottom whisker plots correspond to the MESM and IGSM ensembles, respectively.

to the model updates. This suggests that either the ocean response is insensitive to changes in the model forcings or that the diagnostics used in this study are unable to constrain the parameter. The latter is explored in a separate study by the authors (Libardoni et al., 2018).

To evaluate how well the model captures the observed record and demonstrate the wide range of climate states simulated by the MESM ensemble, we compare the model output to the observed climate record (Figures 6 and 7). In Figure 6, we show the global mean surface temperature time series for all ensemble members, along with each of the time series from each of the five observational datasets used in the surface diagnostic. In Figure 7, we compare the linear trend in the 0-3 km global mean ocean heat content estimated from the MESM simulations against the observed estimate. For both the surface and ocean comparisons, we highlight the estimates from the MESM ensemble members which have parameter settings closest to the median values from the full ensemble MESM distributions.

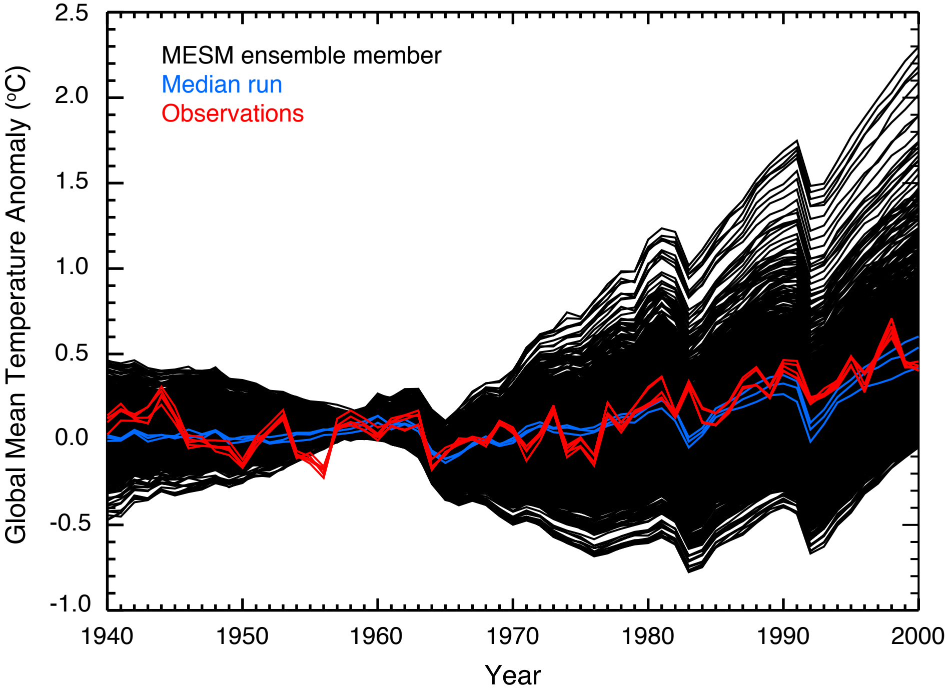

Figure 6. Observed and simulated global mean surface temperature anomalies. The observed time series (red) are derived from each of the five surface temperature datasets used in the surface temperature diagnostic. Also shown are the time series for each MESM simulation (black). Runs with parameter settings closest to the median values from each distribution are highlighted (blue). All anomalies are calculated with respect to the 1906-1995 climatology used in the surface diagnostic.

For both the surface temperature and ocean heat content trends, we have sampled many climate states on the colder and warmer sides of the observed values. We note here that the negative ocean heat content trends are the result of simulations with strong cooling that lie well outside the acceptable range of the parameter space. All simulations with this negative trend have F_{aer} less than or equal to -0.75 Wm^{-2}, a zero-probability region in the MESM ensemble. For the global mean surface temperature time series, the median simulations compare favorably to the observed time series. For the ocean heat content trend, the median simulations tend to overestimate the trend compared to the observed value. Perfect matches should not be expected when comparing the median simulations to the observations, however. Because we derived the distributions using the surface and ocean records, only those runs that agree with both diagnostics are not rejected for being inconsistent with the data. Thus, a model simulation that reproduces the global mean surface temperature perfectly may have too little warming in the deep ocean. Similarly, a model with the perfect ocean heat content trend may not match the surface temperature time series. Small deficiencies in the median runs compared to a single observed record are the result of simultaneously matching the surface and ocean records.

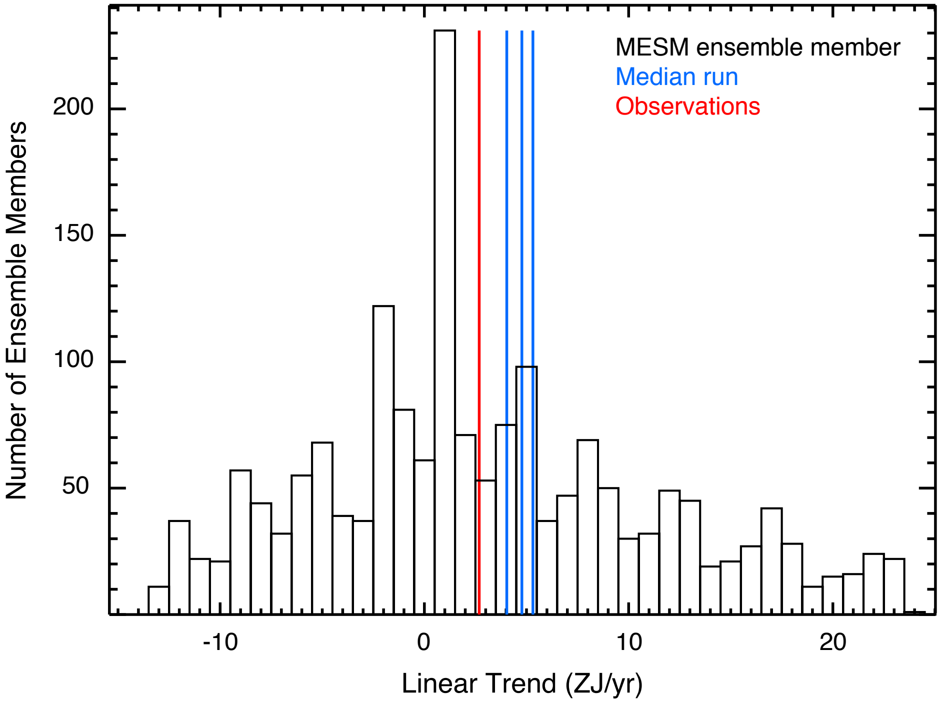

Figure 7. Histogram of linear trends in the 0-3 km global mean ocean heat content estimated from each MESM ensemble member. The observed trend (red) and trends estimated from the MESM simulations with parameter values closest to the medians from each distribution (blue) are shown as vertical lines.

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~~ simulation 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~~ The third-order polynomial fit is chosen to be of the same form as the fits derived for the IGSM model. Further, an investigation of different order fits (not shown) indicated that at least a third-order fit is required to satisfactorily fit the data. From the functional fits, we derive response surfaces for each of the transient properties (Figure 8). 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

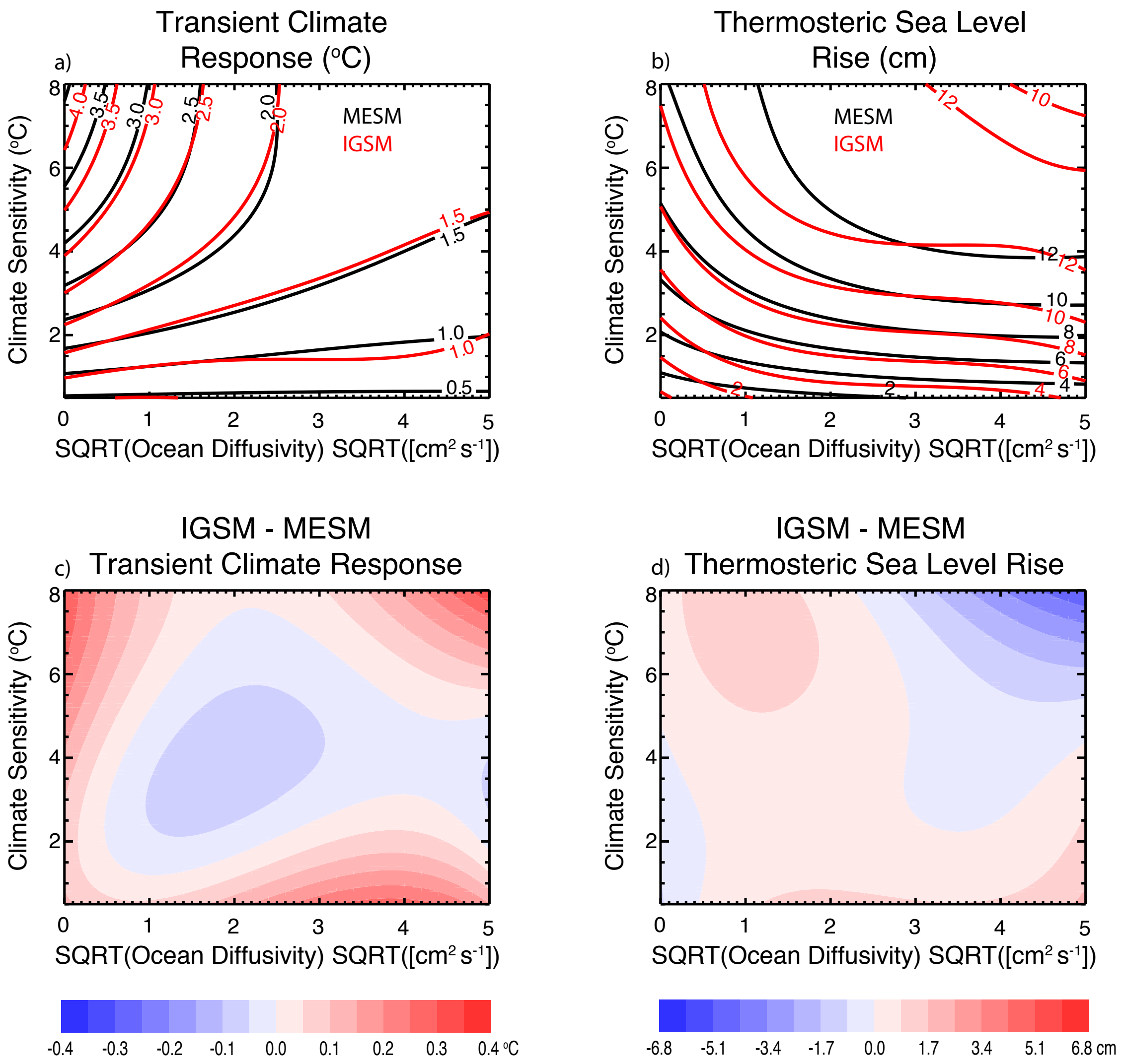

Figure 8. Model response surfaces for (a) TCR and (b) thermosteric sea level rise. Contours for the MESM response surfaces are shown in black and contours for the IGSM surfaces are shown in red. Differences between the fits are also shown (c and d).

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~~ using 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~~ $\sqrt{K_v}$ 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~~ We hope that this study motivates ~~similar investigations moving forward that produce~~ other modeling groups to perform similar investigations that provide documented accountings of model updates~~and that lead to a~~, leading to a more robust understanding of the impacts that the changes have on ~~the parameter calibration~~parameter estimation and model behavior.

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 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 thus shown the total effect of changing both the model and forcings on the parameter distributions, not the effects of the 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~~ samples from the parameter distributions to provide estimates of TCR from the 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.

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.

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 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.

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.: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, J. Geophys. Res., 111, doi:10.1029/2005JD006 548, 2006.

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.

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, 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.

Hansen, J., Lacis, A., Rind, D., Russell, G., Stone, P., Fung, I., Ruedy, R., and Lerner, J.: Climate sensitivity: Analysis of feedback mechanisms, in: Climate Processes and Climate Sensitivity, Geophysical Monograph, edited by Hansen, J. E. and Takahashi, T., vol. 29, pp. 130–163, American Geophysical Union, Washington, D.C., 1984.

Hansen, J., Lacis, A., Rind, D., Russell, G., Stone, P., Fung, I., Ruedy, R., and Stone, P.: Global climate changes as forecast by Goddard Institute for Space Studies three-dimensional model, J. Geophys. Res., 93, 9341–9364, 1988.

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.

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.

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.

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, 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.

Libardoni, A. G., Forest, C. E., Sokolov, A. P., and Monier, E.: Estimates of climate system properties incorporating recent climate change, manuscript submitted for publication, 2018.

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., 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, 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., 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, 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. and Stone, P. H.: A flexible climate model for use in integrated assessments, Clim. Dyn., 14, 291–303, 1998.

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.

Urban, N. M. and Keller, K.: Complementary observational constraints on climate sensitivity, Geophys. Res. Let., 36, doi:10.1029/2008GL036 457, 2009.