Geophysical Research Letters

RESEARCH LETTER
10.1029/2019GL082442

Key Points:

- Estimates are highly sensitive when using a single source to approximate internal variability
- Probabilistic estimates are too narrow if the internal variability comes from a single model
- Estimates tend to converge when enough samples of internal variability are used

Supporting Information:

- Supporting Information S1

Correspondence to:
C. E. Forest,
ceforest@psu.edu

Citation:
Libardoni, A. G., Forest, C. E., Sokolov, A. P., & Monier, E. (2019). Underestimating internal variability leads to narrow estimates of climate system properties. *Geophysical Research Letters*, *46*, 10,000–10,007. https://doi.org/10.1029/2019GL082442

Received 14 FEB 2019
Accepted 31 JUL 2019
Accepted article online 19 AUG 2019
Published online 31 AUG 2019

Underestimating Internal Variability Leads to Narrow Estimates of Climate System Properties

Alex G. Libardoni[1,2], Chris E. Forest[1,3], Andrei P. Sokolov[4], and Erwan Monier[4,5]

[1]Department of Meteorology and Atmospheric Science, Pennsylvania State University, University Park, PA, USA, [2]Now at Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA, [3]Earth and Environmental Systems Institute, Pennsylvania State University, University Park, PA, USA, [4]Joint Program on the Science and Policy of Global Change, Massachusetts Institute of Technology, Cambridge, MA, USA, [5]Now at Department of Land, Air, and Water Resources, University of California, Davis, CA, USA

Abstract Probabilistic estimates of climate system properties often rely on the comparison of model simulations to observed temperature records and an estimate of the internal climate variability. In this study, we investigate the sensitivity of probability distributions for climate system properties in the Massachusetts Institute of Technology Earth System Model to the internal variability estimate. In particular, we derive probability distributions using the internal variability extracted from 25 different Coupled Model Intercomparison Project Phase 5 models. We further test the sensitivity by pooling variability estimates from models with similar characteristics. We find the distributions to be highly sensitive when estimating the internal variability from a single model. When merging the variability estimates across multiple models, the distributions tend to converge to a wider distribution for all properties. This suggests that using a single model to approximate the internal climate variability produces distributions that are too narrow and do not fully represent the uncertainty in the climate system property estimates.

1. Introduction

Despite improvements to climate models, observational systems, and methodologies, the estimated likely range of equilibrium climate sensitivity has remained relatively unchanged from the Charney Report (Charney, 1979) to the most recent Intergovernmental Panel on Climate Change Fifth Assessment Report (Collins et al., 2013). In both reports, the central estimate for climate sensitivity is 3 °C with a likely range of 1.5 to 4.5 °C. A recent review suggests that although a climate sensitivity below 2 °C may be inconsistent with the current physical understanding of climate system feedbacks, the 1.5 to 4.5 °C range remains relatively unchanged (Knutti et al., 2017). The lack of convergence toward a single value does not imply a lack of knowledge, however, but rather the inherent difficulty of the problem. Estimates of climate sensitivity remain uncertain for a number of reasons, some of which are outlined in Hegerl et al. (2000) and remain relevant today. The first reason is uncertainty in observations. Many studies estimate climate sensitivity using historical observations of climate change. Due to instrumental errors, coverage gaps, and inhomogeneity in measurement practices, there is irreducible error in the time series of climate variables. The second reason is systematic/structural errors in models. Many studies derive estimates of climate sensitivity by comparing model output to observations (e.g., Forest et al., 2008; Knutti et al., 2003; Libardoni & Forest, 2013; Olson et al., 2013). All models are approximations of reality and thus do not exactly match the climate system. The last reason is chaotic or internal variability in the climate system. In the absence of external forcing, fluctuations in the atmosphere and ocean still occur due to processes and feedbacks active in the climate system. This unforced internal variability is embedded in any climate signal, and due to its chaotic nature, represents an irreducible uncertainty.

Here, we show how estimates of climate sensitivity are affected by different estimates of internal variability. In previous work to evaluate model performance (Libardoni et al., 2018a, 2018b), a goodness-of-fit statistic is used that weighs model-to-observation residuals in temperature diagnostics by the internal variability of the climate system (see equation (1)). In many methods (Aldrin et al., 2012; Olson et al., 2013; Sansó & Forest, 2009), internal variability is estimated from preindustrial control runs of atmosphere-ocean general circulation models (AOGCMs). In these runs, forcing patterns are fixed over long simulations and the coupled atmosphere and ocean system interacts in the absence of changes in external forcings. Thus far, we have only used one model to estimate the internal variability. Over 20 AOGCMs submitted preindustrial controls runs to the Coupled Model Intercomparison Project Phase 5 (CMIP5; Taylor et al., 2012). Due to structural differences, the internal variability is not the same across all models and a single model does not span the full range of variability. We show here that this has a significant impact on the calculation of probability distributions for climate sensitivity and other climate system properties. At best, the uncertainty in the resulting estimates are similar when variability is estimated from a single model compared to when multiple models are used, but the uncertainty is generally underestimated when just a single model is used to estimate the variability.

We present the study as follows. In section 2, we identify the models that are used and how estimates of the internal variability are derived from long control runs. Section 3 presents probability distributions derived from each of the variability estimates and proposes a method for combining estimates across multiple models. We conclude the study in section 4.

2. Methods

In this work, we use the Massachusetts Institute of Technology Earth System Model (MESM; Sokolov et al., 2018) to simulate many possible climate states. Because MESM was developed from model components of Earth System Models, the climate system properties are determined by feedbacks within MESM rather than constant parameters like in most intermediate complexity models. To simulate different climate states, we vary model parameters to adjust the feedback strengths (e.g., the cloud feedback on climate sensitivity), which in turn depend on the model state vector. Using the 1,800-member ensemble of MESM described in Libardoni et al. (2018a), we derive estimates of the joint probability distribution function (PDF) of effective climate sensitivity (ECS), effective ocean diffusivity (K_v), and the net anthropogenic aerosol forcing (F_{aer}). Each model simulation adjusts the climate state to set the climate system properties, $\theta = (\text{ECS}, K_v, F_{aer})$, and simulates past climate when driven by historical forcings. To average over the internal variability in MESM, each model simulation is taken as the ensemble average of a four-member initial condition ensemble. We evaluate the likelihood of a given model run by comparing the model output to time series of observed climate change. In particular, we use the decadal mean time series of surface temperature anomalies in four equal-area zonal bands from 1941 to 2010 with respect to a 1906–1995 climatology and the linear trend in the 0- to 2,000-m global mean ocean heat content change between 1991 and 2010 as diagnostics.

We note here that the dates used in the ocean heat content diagnostic represent a change from our previous work (Libardoni et al., 2018a, 2018b), where data from 1955 to 2010 are used to estimate the linear trend. We have implemented this change for two reasons. First, including only recent data uses the observed accelerated warming of the ocean (Gleckler et al., 2016) as a constraint on model performance. Second, improvements to the ocean monitoring system have decreased the uncertainty in estimates of global mean ocean heat content in recent years (Levitus et al., 2012).

For each diagnostic, we calculate a goodness-of-fit statistic, r^2, as the weighted sum-of-squared residuals between model output and observations. Mathematically, r^2 is expressed as

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

where $\mathbf{x}(\theta)$ is the vector of model output for a set of model parameters, $\mathbf{y}$ is the vector of observed data, and $\mathbf{C}_N^{-1}$ is the inverse of the noise-covariance matrix. In our method, $\mathbf{C}_N^{-1}$ represents the temperature patterns we would expect in the model diagnostics in the absence of external forcing and observational noise. The observations and model output are projected onto the internal variability patterns, with the weight assigned to each element of the residual vector inversely proportional to the size of the deviations expected in the unforced climate. Here, the internal variability patterns are determined through an eigendecomposition of $\mathbf{C}_N^{-1}$ and the weights are proportional to the inverse square root of the corresponding eigenvalues. To avoid projecting onto patterns with infinite variance (i.e., those with small eigenvalues), we retain only the leading 21 eigenvectors. We note that the definition of r^2 presented here is different than the coefficient of determination for evaluating the 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. A more thorough discussion of the goodness-of-fit statistic calculation and its properties is provided in Forest et al. (2001) and Libardoni et al. (2018b). From this statistic, the likelihood of a given model run representing past climate change is calculated and used to derive a joint PDF for the model parameters (Lewis, 2013; Libardoni & Forest, 2011).

We estimate the unforced variability patterns from the preindustrial control runs of AOGCMs submitted to the CMIP5 archive. We choose to use preindustrial control runs for two reasons. First, all forcings are fixed at preindustrial levels, allowing the climate system to evolve without influence from external forcings. Second, control runs tend to be long, allowing for many samples of the model diagnostics to be extracted. From these samples, we are able to calculate a larger statistical sample of the internal variability to better estimate the covariance matrix.

In recent work (Libardoni et al., 2018a, 2018b), our estimates have used the variability from only one model, the Community Climate System Model, version 4 (CCSM4; Gent et al., 2011), in the derivation of parameter distributions. All models are approximations of reality and thus have internal variability inconsistent with that of the true climate system. We test the sensitivity of the parameter distributions to the internal variability of the climate system by deriving PDFs using internal variability estimates from many different models. To reflect the current state of models across multiple research groups, we choose to use the preindustrial control runs from 25 models from the CMIP5 archive. A list of these models is provided in the supporting information that accompanies this paper.

From each model, we extract multiple estimates of the temperature diagnostics to determine the patterns of unforced variability. To treat the monthly mean surface temperature data from the models as if they were observations, we first regrid the model output from its native grid to the 5 × 5° grid of the observations (Hansen et al., 2010; Morice et al., 2012; Rohde et al., 2013; Vose et al., 2012). In the regridding process, we take the area-weighted average of all grid boxes in the model grid that fall partially or completely within the 5 × 5° grid box of the observations. From the 5° resolution data, we extract 105-year segments to correspond to the 1906–2010 period used in the surface temperature diagnostic. The first segment begins in year one of the simulation and we draw subsequent 105-year segments by shifting the start of the segment forward 4 years until the final year of a segment is the last year of the simulation. We choose a short offset to increase the number of samples available for calculating the covariance matrix.

After the output is regridded and the segments extracted, we mask the model output using the missing data mask from each of the observational data sets. For each control run segment, the output is masked so that the coverage matches that of the observational data set for the length of the time series. We then treat individual segments as if they are observations and calculate the decadal mean temperature anomaly time series. The noise-covariance matrix used in the goodness-of-fit statistic calculation is estimated by calculating the spatial and temporal correlations across all segments from an individual model. Thus, each model has its own $\mathbf{C}_N^{-1}$ and eigendecomposition estimated from its internal variability patterns. In this study, we retain the same number of leading eigenvectors (21) for each model.

We extract noise estimates of the ocean diagnostic following methods similar to those for the surface data. From the gridded data and depth field provided in the model documentation, we calculate the average global mean potential temperature in the 0- to 2,000-m layer. We ignore the small differences between potential temperature and temperature observed at these depths. From the global mean average time series, we extract 20-year segments corresponding to the 1991–2010 ocean diagnostic. We again separate the starting date of the samples by 4 years. Before calculating the trend, we convert the mean temperature in the layer to ocean heat content using the conversion factor 900/0.09 EJ/°C from Levitus et al. (2012). Because the ocean diagnostic is a single point, namely the linear trend in global mean ocean heat content, $\mathbf{C}_N^{-1}$simplifies to the standard deviation in the estimate of the slope.

Once the PDFs are calculated using the variability estimates derived from each CMIP5 model, we extract the median value of each parameter from its marginal distribution. We run simulations of MESM using these parameter sets to further test the distributions a posteriori. Namely, we distinguish models by whether the MESM runs with the median estimates are consistent with observed changes in global mean surface temperature change and/or ocean heat content by comparing long-term trends simulated by the model with the changes observed in the historical records.

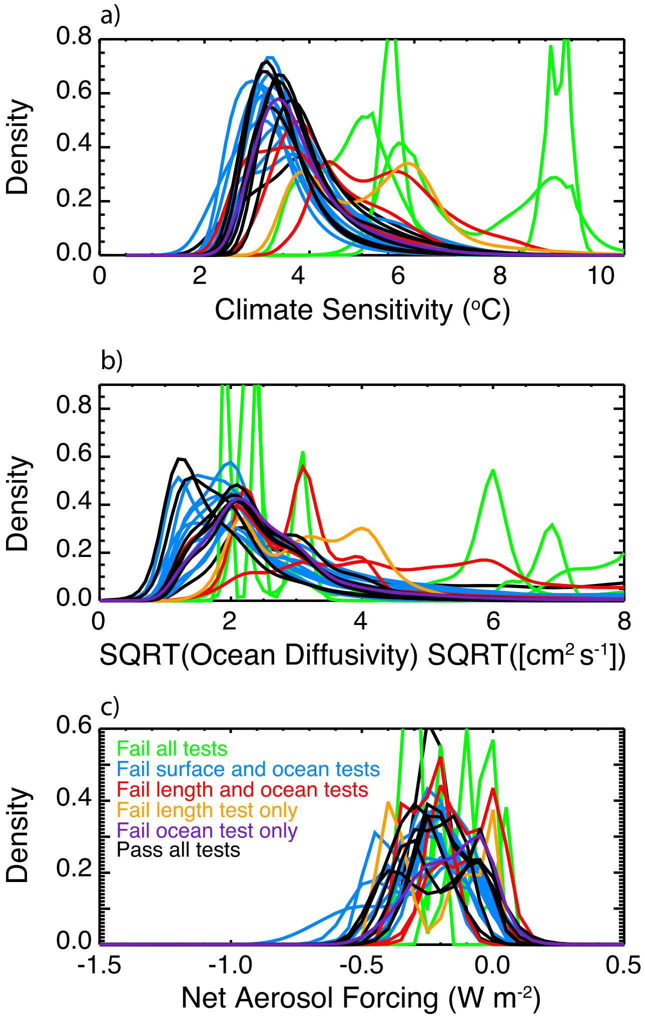

Figure 1. Marginal probability distribution functions for (a) effective climate sensitivity, (b) $\sqrt{K_v}$, and (c) F_{aer} derived using variability estimates from each of the 25 Coupled Model Intercomparison Project Phase 5 models. Distributions are color coded based on how a model passes or fails the selection tests.

3. Results

We present the probability distributions estimated using the variability estimates from each of the 25 different CMIP5 models in Figure 1. As in Libardoni et al. (2018b), we estimate the joint PDF using each of the four surface temperature data sets individually and then merge the estimates by taking the point-by-point average of the four individual joint PDFs. Although not shown here, offline tests show changing to the shorter ocean diagnostic described in section 2 leads to lower ECS and K_v estimates and have a much smaller impact on F_{aer} estimates (see supporting information). The main results and conclusions of this study are independent of these changes, however. We present the median values of these distributions used to run the additional MESM simulations in Figure 2.

We assume that the internal variability of each CMIP5 model is an equally likely representation the climate system and thus assign equal weight to each PDF estimate, choosing not to judge one model as better than any other. For this reason, we have not labeled the individual distributions. However, the distributions are color coded based on three selection methods that we describe next. For method one, we identify models based on their control run length being longer or shorter than 500 years. As the length of the control run increases, the number of estimates of the diagnostics used to calculate the noise-covariance matrix increases. For the second and third methods, we use the MESM simulations with parameters set to the median values from each of the marginal distributions shown in Figures 1 and 2. We calculate the difference between the

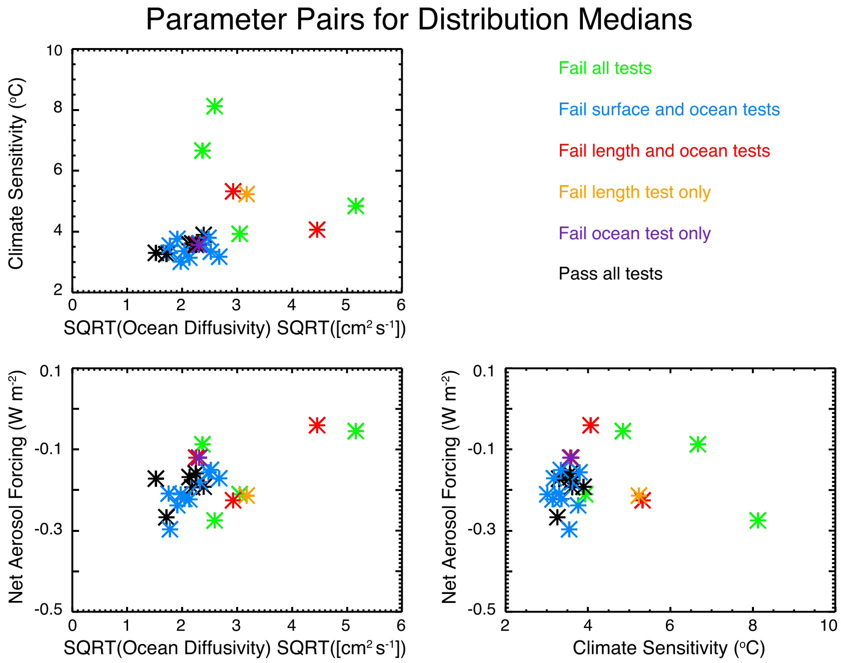

Figure 2. Median parameter values extracted from each distribution in Figure 1 for each combination of parameters. Colors are as in Figure 1.

global mean surface temperature in the last decade of the simulation (2001–2010) and the first 20 years of the simulation (1861–1880) and compare the difference to the observed climate record (Morice et al., 2012). We have identified models where the absolute value of the difference between the temperature change estimated from the model and calculated from the observations is less than 0.05 °C. The overall range of 0.1 °C is slightly wider than the spread in the estimates (0.08 °C) when the temperature change is calculated using different surface temperature data sets for a common period (Hansen et al., 2010; Morice et al., 2012; Rohde et al., 2013; Vose et al., 2012). Similarly, we have identified models by whether the simulated change in global mean ocean heat content matches the observed change. A summary of which test each model passes and fails can be found in the supporting information.

Although presented without uncertainty estimates, it is clear that the PDFs are sensitive to which CMIP5 model is used to estimate the variability (Figure 1). The sensitivities arise because the patterns of internal variability estimated from each CMIP5 model are different, and the elements of the $\mathbf{x}(\theta)-\mathbf{y}$ vector project differently onto these patterns. Thus, a difference between the observations and model output that is assigned little weight with the pattern from one model may be assigned a large weight with the pattern from a second model.

We present the median parameter values for each PDF to further highlight the sensitivity of the estimates to the internal variability and show the θs for the MESM simulations used in the separation methods described above (Figure 2). We find the greatest agreement in the F_{aer} medians. There are no extreme outliers and the estimates tend to center between −0.3 and −0.1 W/m^2. Despite the relatively narrow range, it is wider than those found when considering other factors such as which surface temperature data set is used and the end date of the diagnostic (Libardoni & Forest, 2013; Libardoni et al., 2018b). For most of the models, we find climate sensitivity to be centered between 3.0 and 4.0 °C. However, we do observe variation between the individual estimates and five estimates greater than 4.8 °C and as high as 8 °C. We see a similar spread in the marginal distributions of $\sqrt{K_v}$. While the distributions tend to cluster toward lower values between 1.5 and 3.2 cm/s$^{1/2}$, there are two estimates with $\sqrt{K_v}$ greater than 4.0 cm/s$^{1/2}$.

We note that all of the ECS and $\sqrt{K_v}$ median outliers come from distributions where models with less than 500 years of control run data are used to estimate the internal variability. In these cases, the control run simulation is too short to extract enough samples of the model diagnostics to accurately estimate the noise-covariance matrix. To increase the number of samples used to estimate $\mathbf{C}_N^{-1}$, we pool segments

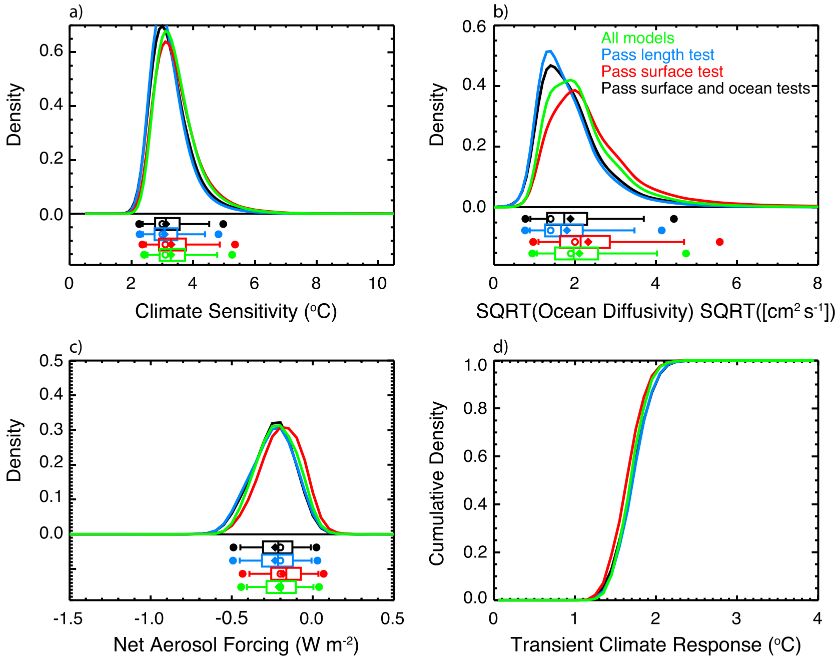

Figure 3. Marginal probability distribution functions for effective climate sensitivity (a), $\sqrt{K_v}$ (b), and F_{aer} (c) and cumulative distribution function for transient climate response (d) resulting from merging the variability estimates across models with similar characteristics. Groupings are for all models (green), all models with length greater than 500 years (blue), all models with median parameters consistent with global mean temperature change (red), and all models with median parameters consistent with global mean temperature change and ocean heat content change (black). Colors for the merged groups are chosen so that the color of the last test passed before failure matches that from Figure 1 (e.g., models that fail all tests are plotted green in Figure 1 and therefore only fit in the "all models" group).

extracted from models with similar length and variability characteristics into a single internal variability estimate. From the increased collection of segments, we estimate a noise-covariance matrix and derive an additional PDF using the estimate as if it came from a single model. Similar to the treatment of the individual models, we retain the leading 21 eigenvectors when determining the patterns of internal variability from the pooled samples.

For each of the separation methods discussed previously—length of the control simulation, MESM simulation with median parameters matching global mean temperature change, and MESM simulation matching global mean ocean heat content change—we group the diagnostic segments from all models that pass the test into a single, merged pool of estimates. In total, four groupings are evaluated: (1) all models, (2) all models with a 500-year or greater control run, (3) all models where the median distribution values lead to a MESM simulation consistent with the observed global mean temperature change, and (4) all models where the median parameter simulation is consistent with both the global mean surface temperature and ocean heat content changes. The models that are included in each grouping are given in the supporting information and the resulting distributions are shown in Figure 3.

We observe a much smaller spread in the parameter estimates when the merged variability estimates are used to calculate the noise-covariance matrix compared to when individual CMIP5 models are used. Climate sensitivity estimates are nearly independent of the criteria used to pool the variability estimates. Across the four groupings, the 5th and 95th percentiles of the distributions vary between 2.4 and 2.5 and 4.4 and 4.9 °C, respectively. The widths of these distributions are similar to the 90% confidence interval of 2.9 to 5.3 °C estimated from the distribution using only the CCSM4 internal variability derived in Libardoni et al. (2018b). We note that similar to the climate sensitivity distributions, the F_{aer} distributions show strong agreement regardless of which tests are used to group variability estimates. Unlike with the climate sensitivity and aerosol distributions, we do not observe the $\sqrt{K_v}$ distributions collapsing toward a single distribution with the different merged variability estimates. We estimate 90% confidence intervals of 0.9–3.7, 0.9–3.5, 1.1–4.7, and 1.1–4.0 cm/s$^{1/2}$ when variability from all models, models longer than 500 years, models that pass the

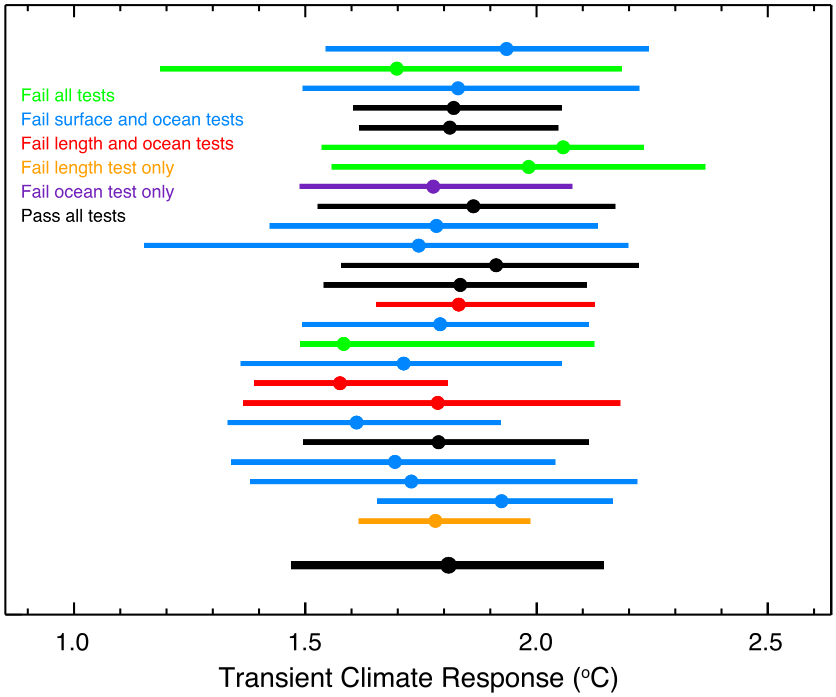

Figure 4. Ninety-percent confidence interval (horizontal line) and median (dots) from the probability distribution functions of transient climate response derived from 1,000-member Latin Hypercube Samples drawn from the joint PDFs derived from the internal variability estimates of individual models. Colors are as in Figures 1 and 2. Also shown is the estimate from the PDF derived using the merged variability segments from all models (bottom bold black line). PDF = probability distribution function.

surface temperature test, and models that pass both the surface temperature and ocean heat content tests are used to estimate the noise-covariance matrix, respectively. We also note longer upper tails in distributions when merged variability estimates are used compared to the distribution using variability from CCSM4 alone (1.3–3.8 cm/s$^{1/2}$ 90% confidence interval).

To complement our estimates of ECS, we also estimate the transient climate response (TCR). As suggested in Knutti et al. (2017), TCR is a more relevant metric for predicting temperature change over the next few decades. Following the methods of Libardoni et al. (2018a), we draw a 1,000-member Latin hypercube sample (McKay et al., 1979) of the model parameters from each of the joint PDFs derived using a single CMIP5 model to estimate the internal variability. We convert each of the ECS-$\sqrt{K_v}$ pairs to TCR using the functional fit derived in Libardoni et al. (2018a). We show the median and 90% confidence intervals for the TCR distributions derived from the variability estimates from each individual model along with the TCR distribution derived using the merged variability estimate including all models in Figure 4. Similar to the $\sqrt{K_v}$ distributions, the estimates of TCR are sensitive to the model used to estimate the internal variability. When compared to the transient climate response PDF using only CCSM4 (Libardoni et al., 2018b), using the merged variability across all models broadens the 90% confidence interval from 1.5–2.0 to 1.4–2.1 °C.

4. Conclusions

In this study, we show that the internal variability estimate, as used in the likelihood function to weight the residuals between model output and observations, has a strong impact on parameter estimation. By estimating parameters using internal variability from different individual models, we observe that the posterior distributions are sensitive to the control run data. We have explored several criteria to account for these differences. In particular, some models have control runs that are too short to extract enough samples of variability to provide a good estimate of the noise-covariance matrix. Combining the variability from multiple models provides enough samples for a stable covariance estimate and leads to a convergence of the distributions. Because models are approximations of the real world, we cannot assume that any model will exactly simulate the internal variability of the natural world. Model developers make different decisions when choosing dynamical cores, cloud microphysics schemes, and many other components. These choices impact model behavior, leading to structural uncertainty in the estimation of the patterns of internal variability. Drawing samples from a single model cannot account for this structural uncertainty and leads to only one of the many possible representations being used when estimating covariance matrices. Using multiple models to estimate the internal variability draws samples across the different model structures and better represents the uncertainty we have in knowing the variability of the natural world. When sampling across this wider range of model structures, we see a broadening of our estimates of model parameters and transient climate response. In particular, the constraints on all model parameters and transient climate response when using the merged estimate of the internal variability across 25 different CMIP5 models were either the same as or narrower than those when using the CCSM4 model alone. Given this, we suggest using multiple models when estimating the internal variability of the climate system, rather than a longer run of a single model.

Acknowledgments
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/current. This research was partially supported by the National Science Foundation (NSF) through the NSF Cooperative Agreement GEO-1240507. We acknowledge the World Climate Research Program's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups listed in the supporting information included with this paper for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. A. G. L. also recognizes the valuable discussions and suggestions made by K. Davis, K. Keller, and R. Najjar in the development of this work. The model data will be stored and available at http://www.datacommons.psu.edu/ or directly at https://doi.org/10.26208/f3x1-pt27 website.

References

Aldrin, M., Holden, M., Guttorp, P., Skeie, R. B., Myhre, G., & Bernstein, T. K. (2012). Bayesian estimation of climate sensitivity based on a simple climate model fitted to observations of hemispheric temperatures and global ocean heat content. *Environmetrics*, *23*, 253–271.

Charney, J. (1979). *Carbon dioxide: A scientific assessment* (pp. 33). Washington, DC: National Academy of Sciences Press.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., et al. (2013). 2013: Long-term Climate Change: Projections, Commitments and Irreversibility. In T. F. Stocker et al. (Eds.), *Climate Change 2013:The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Forest, C. E., Allen, M. R., Sokolov, A. P., & Stone, P. H. (2001). Constraining climate model properties using optimal fingerprint detection methods. *Climate Dynamics*, *18*, 277–295.

Forest, C. E., Stone, P. H., & Sokolov, A. P. (2008). Constraining climate model parameters from observed 20th century changes. *Tellus*, *60A*(5), 911–920.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., et al. (2011). The Community Climate System Model version 4. *Journal of Climate*, *24*, 4973–4991. https://doi.org/10.1175/2011JCLI4083.1

Gleckler, P. J., Durack, P. J., Stouffer, R. J., Johnson, G. C., & Forest, C. E. (2016). Industrial-era global ocean heat uptake doubles in recent decades. *Nature Climate Change*, *6*, 394–398.

Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, *48*, RG4004. https://doi.org/10.1029/2010RG000345

Hegerl, G. C., Stott, P. A., Allen, M. R., Mitchell, J. F. B., Tett, S. F. B., & Cubasch, U. (2000). Optimal detection and attribution of climate change: Sensitivity of results to climate model differences. *Climate Dynamics*, *16*, 737–754.

Knutti, R., Rugenstein, M. A. A., & Hegerl, G. C. (2017). Beyond equilibrium climate sensitivity. *Nature Geoscience*, *10*, 727–736.

Knutti, R., Stocker, T. F., Joos, F., & Plattner, G.-K. (2003). Probabilistic climate change projections using neural networks. *Climate Dynamics*, *21*, 257–272.

Levitus, S., Antonov, J. I., Boyer, T. P., Baranova, O. K., Garcia, H. E., Locarnini, R. A., et al. (2012). World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. *Geophysical Research Letters*, *39*, L10603. https://doi.org/10.1029/2012GL051106

Lewis, N. (2013). An objective Bayesian improved approach for applying optimal fingerprint techniques to climate sensitivity. *Journal of Climate*, *26*, 7414–7429. https://doi.org/10.1175/JCLI-D-12-00473.1

Libardoni, A. G., & Forest, C. E. (2011). Sensitivity of distributions of climate system properties to the surface temperature dataset. *Geophysical Research Letters*, *38*, L22705. https://doi.org/10.1029/2011GL049431

Libardoni, A. G., & Forest, C. E. (2013). Correction to “Sensitivity of distributions of climate system properties to the surface temperature data set”. *Geophysical Research Letters*, *40*, 2309–2311. https://doi.org/10.1002/grl.50480

Libardoni, A. G., Forest, C. E., Sokolov, A. P., & Monier, E. (2018a). Baseline evaluation of the impact of updates to the MIT Earth System Model on its model parameter estimates. *Geoscientific Model Development*, *11*, 3313–3325.

Libardoni, A. G., Forest, C. E., Sokolov, A. P., & Monier, E. (2018b). Estimates of climate system properties incorporating recent climate change. *Advances in Statistical Climatology, Meteorology and Oceanography*, *4*, 19–36.

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

Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research*, *117*, D08101. https://doi.org/10.1029/2011JD017187

Olson, R., Sriver, R., Chang, W., Haran, M., Urban, N. M., & Keller, K. (2013). What is the effect of unresolved internal climate variability on climate sensitivity estimates? *Journal Geophysical Research: Atmosphere*, *118*, 4348–4358. https://doi.org/10.1002/jgrd.50390

Rohde, R., Muller, R. A., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., et al. (2013). A new estimate of the average Earth surface and land temperature spanning 1753 to 2011. Geoinfor. Geostat: An Overview, 1:1 https://doi.org/10.4172/gigs.1000101

Sansó, B., & Forest, C. (2009). Statistical calibration of climate system properties. *Applied Statistics*, *58*, 485–503.

Sokolov, A., Kicklighter, D., Schlosser, A., Wang, C., Monier, E., Brown-Steiner, B., et al. (2018). Description and evaluation of the MIT Earth System Model (MESM). *Journal of Advances in Modeling Earth Systems*, *10*, 1759–1789. https://doi.org/10.1029/2018MS001277

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experimental design. *Bulletin of the American Meteorological Society*, *93*, 485–498.

Vose, R. S., Arndt, D., Banzon, V. F., Easterling, D. R., B.Gleason, Huang, B., et al. (2012). NOAA's Merged Land-Ocean Surface Temperature analysis. *Bulletin of the American Meteorological Society*, *93*, 1677–1685.