The Detection and Attribution of Climate Change Using an Ensemble of Opportunity

DÁITHÍ A. STONE

Department of Physics, and Department of Zoology, University of Oxford, Oxford, United Kingdom

MYLES R. ALLEN

Department of Physics, University of Oxford, Oxford, United Kingdom

FRANK SELTEN AND MICHAEL KLIPHUIS

Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands

PETER A. STOTT

Hadley Centre for Climate Prediction and Research, Reading, United Kingdom

(Manuscript received 13 June 2005, in final form 24 April 2006)

ABSTRACT

The detection and attribution of climate change in the observed record play a central role in synthesizing knowledge of the climate system. Unfortunately, the traditional method for detecting and attributing changes due to multiple forcings requires large numbers of general circulation model (GCM) simulations incorporating different initial conditions and forcing scenarios, and these have only been performed with a small number of GCMs. This paper presents an extension to the fingerprinting technique that permits the inclusion of GCMs in the multisignal analysis of surface temperature even when the required families of ensembles have not been generated. This is achieved by fitting a series of energy balance models (EBMs) to the GCM output in order to estimate the temporal response patterns to the various forcings.

This methodology is applied to the very large Challenge ensemble of 62 simulations of historical climate conducted with the NCAR Community Climate System Model version 1.4 (CCSM1.4) GCM, as well as some simulations from other GCMs. Considerable uncertainty exists in the estimates of the parameters in fitted EBMs. Nevertheless, temporal response patterns from these EBMs are more reliable and the combined EBM time series closely mimics the GCM in the context of transient forcing. In particular, detection and attribution results from this technique appear self-consistent and consistent with results from other methods provided that all major forcings are included in the analysis.

Using this technique on the Challenge ensemble, the estimated responses to changes in greenhouse gases, tropospheric sulfate aerosols, and stratospheric volcanic aerosols are all detected in the observed record, and the responses to the greenhouse gases and tropospheric sulfate aerosols are both consistent with the observed record without a scaling of the amplitude being required. The result is that the temperature difference of the 1996–2005 decade relative to the 1940–49 decade can be attributed to greenhouse gas emissions, with a partially offsetting cooling from sulfate emissions and little contribution from natural sources.

The results support the viability of the new methodology as an extension to current analysis tools for the detection and attribution of climate change, which will allow the inclusion of many more GCMs. Shortcomings remain, however, and so it should not be considered a replacement to traditional techniques.

1. Introduction

The detection and attribution of observed climate change play a central role in climate change research, both because of its role in connecting the many other research branches in the field (Houghton et al. 2001; International Ad Hoc Detection and Attribution Group 2005) and because of its ultimate implications for individual stakeholders (Allen 2003; Allen and Lord 2004). This field came into its own with the first experiment with a fully coupled general circulation model (GCM) to include ensembles of simulations representing the transient response to separate forcing sources

Corresponding author address: Dáithí A. Stone, AOPP, Department of Physics, University of Oxford, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, United Kingdom.
E-mail: stoned@atm.ox.ac.uk

(Tett et al. 1999; Stott et al. 2000). Such ensembles permit spatiotemporal comparisons with observed climate change that simultaneously serve to strengthen and constrain our confidence in our observations, our understanding of sources of past forcing, and our understanding of climate processes encapsulated in the dynamical models. Furthermore, they also provide a test of the cause–effect relationship required by stakeholders in the climate change issue.

This methodology for detection and attribution depends on the generation of ensembles of climate model simulations following multiple forcing scenarios. The ensembles are necessary in order to accurately extract the underlying climate response to particular external forcings from the natural internal variability of the climate system. However, such families of ensembles require large computational resources and it may be difficult or even unfeasible to run so many simulations of a given GCM. Currently, the necessary set of ensembles has been performed with only a few GCMs (Stott et al. 2006). However, the generation of the single ensemble of simulations including a relatively comprehensive set of external forcings is more feasible and common. Here we develop a procedure that allows application of the standard detection and attribution methodology when only such an ensemble forced with a single large set of forcings is available. The development of this procedure will allow the eventual inclusion of many more GCMs into the detection and attribution framework and thus a more robust characterization of the importance of the various external forcings on past and future climate change.

2. Model and data

We use output from the Challenge Project conducted in the summer of 2003 by the Dutch Meteorological Institute (KNMI) using machines at the Academic Computing Centre in Amsterdam (SARA; Selten et al. 2003). This project consists of a 62-member initial condition ensemble of simulations of the National Center for Atmospheric Research (NCAR) Community Climate System Model version 1.4 (CCSM1.4) covering the 1940–2080 period. The CCSM1.4 is a fully coupled GCM of the atmosphere, ocean, sea ice, and land surface (Boville et al. 2001), making this the largest initial condition ensemble of transient climate simulations with a coupled GCM at present. Each simulation was initialized by a small random perturbation in the temperature field of the atmosphere in an initial state obtained from an earlier transient simulation. This ensemble is designed to provide a large dataset on changes in extreme events and so all members are forced with the identical historical scenario of prescribed forcings through to 2000 following C. M. Ammann et al. (2006, personal communication). These comprise changes in the mass mixing ratios of tropospheric greenhouse gases (GHGs), changes in sulfate emissions [resulting in tropospheric sulfate aerosols (SUL) through an interactive sulfate simulation (Rasch et al. 2000)], changes in the optical depth from stratospheric volcanic aerosols (VOL), and changes in solar radiation (SOL) according to Hoyt and Schatten (1993). For this study we estimate the global GHG forcing (Fig. 1a) from the mass mixing ratios of the various gases inputted to the GCM simulations (Dai et al. 2001) according to the formulas in Table 6.2 of Ramaswamy et al. (2001), while the SUL forcing is taken as the global column-integrated burden simulated in the GCM's interactive sulfate model scaled to the 1940 and 1990 combined direct and indirect forcing estimates of Boucher and Pham (2002). Because the CCSM1.4 only includes the direct effect of sulfate aerosols in its calculations, we may expect to find that the GCM underestimates the SUL response. The optical depth of the stratospheric aerosols of C. M. Ammann et al. (2006, personal communication) is multiplied by -20 to represent the VOL forcing (Wigley et al. 2005), while the solar forcing of Hoyt and Schatten (1993) is multiplied by 0.175 to account for the planetary albedo and the geometrical nature of the SOL forcing.

Deseasonalized monthly surface air temperature (SAT) anomalies from the GCM simulations are interpolated onto the HadCRUT2v dataset of $5° \times 5°$ gridded monthly mean observed SAT anomalies of Jones and Moberg (2003) and Rayner et al. (2003) according to the availability of observations. Annual global means are calculated for both datasets and used in the subsequent analysis. These time series are plotted in Fig. 1b. Considering the availability of observations through to the end of 2005 we include these extra years in the analysis. The GHG forcing after 2000 in these simulations follows a business-as-usual scenario similar to the Intergovernmental Panel on Climate Change (IPCC) Special Report on Emissions Scenarios (SRES) A1 scenario (Dai et al. 2001) and close to what has actually occurred. On the other hand the SUL, VOL, and SOL forcings are held constant at year 2000 values. This is reasonably appropriate for the currently fairly stable SUL emissions and for VOL due to the lack of any large volcanic eruptions since 2000, but it does miss some of the latest solar cycle and so may lead to a slight underestimate of the SOL detection.

The lack of ensembles forced with subsets of these external forcings means that this large set of simulations cannot be used in a traditional multisignal detec-

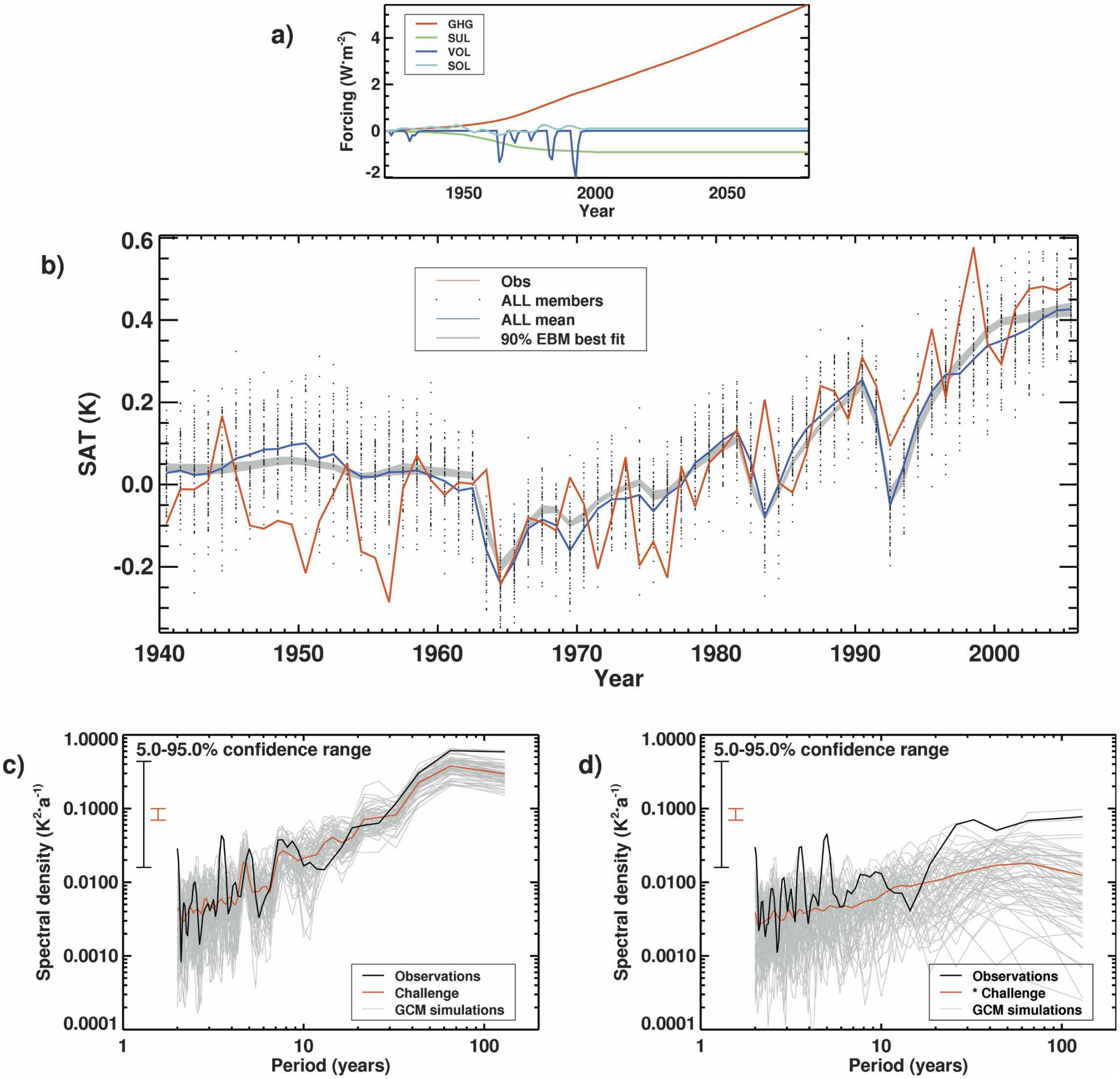

FIG. 1. (a) Global annual mean radiative forcing time series used for the Challenge Project ensemble of climate simulations over the 1920–2080 period, shown as anomalies from the 1940 values. (b) Plot of annual global mean SAT from 1940 through 2005. Values are anomalies from the 1961–90 mean. The 90% confidence interval on the EBM fit is shown in gray shading. "ALL" members refers to the members of the Challenge ensemble (which include all of the forcings). (c) The spectra of annual global mean SAT anomalies from the Challenge simulations and the observations. The spectra from the individual GCM simulations are plotted in gray, while the average spectrum averaged across all ensemble spectra is in red. At the 10% level there is no significant difference between the observed and simulated variability on time scales of 10 yr and longer when an *F* test is performed on the spectrum integrated over these time scales. (d) Same as in (c), but with the GCM ensemble mean response removed from both the simulations and the observations before estimation of the spectra. The observed variability is significantly larger than the simulated variability on time scales of 10 yr and longer.

tion and attribution study. With its large size this ensemble therefore represents an ideal dataset on which to test our new detection and attribution methodology.

For consistency tests of the technique applied here, we also use multiple ensembles of simulations with different forcing scenarios from three other GCMs. The ensembles from the National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory's GFDL-R30 comprise three simulations with GHG forcing, three with GHG and SUL forcing, three with GHG, SUL, and SOL forcing, and three with all four forcings (Broccoli et al. 2003). The ensembles from the NCAR Parallel Climate Model (PCM) comprise 5 simulations with GHG forcing, 4 with SUL forcing, 12 with GHG and SUL forcing (some also include changing stratospheric ozone forcing), 5 with SOL forcing, 4 with VOL forcing, 4 with SOL and VOL forcing, and 4 with all of these forcings (Meehl et al. 2003). The ensembles from the Met Office's (UKMO) Third Hadley Centre Coupled Ocean–Atmosphere GCM (HadCM3) comprise four simulations with GHG forcing, four with GHG and SUL forcing (and also changing stratospheric ozone), four with VOL and SOL forcing, and four with both these anthropogenic and natural forcings (Stott et al. 2000; Tett et al. 2002). With these multiple ensembles we can compare the results of this new methodology applied here with results from traditional detection and attribution studies.

3. Method

For traditional multisignal detection and attribution methods using GCM simulations we need information about the response of the climate system to each of the forcings individually. This is clearly not available from a set of simulations that only includes the combined forcing scenario. However, the global mean SAT response of GCMs tends to closely follow that of a simple tuned energy balance model (EBM) (McAvaney et al. 2001). Thus, in an ensemble forced with a number of external forcings it may be possible to deduce the GCM response to individual forcings by fitting a series of EBMs to the total GCM response, with each EBM representing the response to a different forcing. Each EBM has a different set of parameters because there is no a priori reason to expect, for instance, that the ocean heat uptake corresponding to the SUL forcing, which is concentrated over Northern Hemisphere land, should be identical to that for the more global GHG forcing.

We start with an EBM for forcing *i:*

$$c_i \frac{\partial T_i(t, z)}{\partial t} = F_i(t) - \lambda_i T_i(t, z) - k_i \frac{\partial^2 T_i(t, z)}{\partial z^2}. \qquad (1)$$

Here $T_i(t, z)$ is the time series, as a function of depth in the mixed layer, of the global annual mean temperature response to the evolving forcing F_i shown in Fig. 1a. No boundary conditions are imposed for the bottom of the mixed layer. Here c_i, $(1/\lambda_i)$, and k_i are the heat capacity of the ocean mixed layer, the climate sensitivity, and the parameter for vertical diffusion in the ocean mixed layer for forcing *i* and are tuned to reproduce the mean response of the GCM.

Supposing that the temperature responses to individual forcings add linearly to give the response to the sum of the forcings, we can add the results of the EBMs to get the total temperature response

$$T(t, z) = \sum_{i=1}^{m} T_i(t, z). \qquad (2)$$

This assumption of linear additivity, implicit in the standard detection methodology, appears to hold in GCM output (Gillett et al. 2004). Our aim in fitting the parameters for the EBM is to minimize the squared difference between the total annual mean EBM SAT time series, which we denote as the vector $\mathbf{T} = T(t, 0)$, and the mean response $\mathbf{T}_{\mathrm{GCM}}$ from the ensemble of GCM simulations over the 1940–2005 period. The EBMs are spun up with 20 yr of varying forcings before the start of the comparison in 1940. To tune the parameters in this study we use a downhill simplex method, an iterative geometric method for finding the minimum in a multidimensional function (Nelder and Mead 1965). Uncertainty in this fit arises from the finite GCM ensemble size and the accuracy of the parameter-fitting algorithm. While the downhill simplex method is fairly robust, we are trying to locate the global minimum in a 12-dimensional space using 66 temporal data points, so the parameter fits may end up being somewhat uncertain. The arising distribution of plausible EBM parameter sets and EBM output is estimated using a bootstrap resampling procedure in which 62 simulations are randomly selected with replacement from the full set of 62 GCM simulations. This is performed 100 times, with the mean response of the selected simulations used as input to the analysis to estimate another plausible set of EBM parameters.

Now that we have estimated the responses of the GCM to individual forcings, we can proceed with the standard detection and attribution methodology (Allen and Tett 1999). Under this, we express the observed temperature response pattern $\mathbf{T}_{\mathrm{obs}}$ as a linear sum of the simulated responses determined for each forcing ($\mathbf{T}_i$) plus a residual ($\boldsymbol{\upsilon}_0$):

$$\mathbf{T}_{\mathrm{obs}} = \sum_{i=1}^{m} \mathbf{T}_i \beta_i + \boldsymbol{\upsilon}_0. \qquad (3)$$

Here β_i is the scaling factor corresponding to the response to forcing *i* that is to be estimated in the regression. This relational model depends on the GCM to properly reproduce the temporal pattern of the response and for the EBM to properly reproduce the GCM response, but it explicitly corrects for errors in the amplitude of this response pattern. The regression is performed on the full global annual time series. This is a limitation of this study because spatial information can be important for detecting climate response signals (S. A. Crooks et al. 2006, unpublished manuscript, hereafter referred to as CR06; Stott et al. 2006). Incorporating spatial information will be a future development of this methodology, but for now we want to keep the EBM approximation as simple and appropriate as possible. No optimization or data reduction is used as this is of limited applicability to the temporal data analyzed here.

A control simulation is needed both to estimate the covariance of the residual term, $\boldsymbol{\upsilon}_0$, and to estimate the uncertainty of the β_i scaling parameters. The internal variability is estimated from the 62 ensemble simulations with the transient ensemble mean response removed. The resulting anomaly time series are multiplied by $\sqrt{(62/61)}$ to account for bias in the removal of the mean response. We use these pseudocontrol simulations because they provide more data (4092 yr) than exists in any available control simulation. Half (31) of these pseudocontrol simulations are used for estimating the covariance of the noise term while the other half are used to obtain an independent estimate of the β_i distributions.

The regression formulation used here, referred to as ordinary least squares (OLS), does not include any error in the estimate of the simulated responses; this error is taken into account through the bootstrap resampling. In summary, the attribution results presented here account for sampling uncertainty from both the observations (during the regression step) and the simulations (through the Monte Carlo sampling). The inclusion of the latter source of uncertainty depends implicitly upon the ability of the EBM to represent it. While deficiencies in the EBM fits may partly represent uncertainty arising from structural differences between the GCM and the real world, this uncertainty should not be considered to be included in this analysis. This study also does not account for uncertainty in the source of the forcings nor in how those sources (e.g., emissions) are converted into radiative forcings.

4. Results

a. Comparison of variability

Figure 1c shows the power density spectra of annual global mean SAT anomalies over the 1940 though 2005 period. All spectra have been estimated using a Hanning filter of width 65 yr. The lower power at the 130-yr time scale in the GCM simulations than in the observations reflects the smaller long-term trend visible in Fig. 1b. The significance tests on the spectra take account of the filtering, but they assume a stationary process. It should be noted that this assumption is almost definitely invalid due both to the spatiotemporal nature of the observational masking and the changing external forcings. Considering that these two factors tend to operate on longer time scales, we would expect this assumption to be weakest at longer time scales; the discrepancy between the estimated confidence range in the observed spectrum and the spread of the simulated spectra at longer time scales is indicative of this. At time scales of 10 yr and longer, that is, those relevant for long-term climate change, there is no significant difference, at the 10% level, between the spectra of observed SAT and the ensemble mean spectrum from the GCM simulations. These time scales are of interest as they are the most relevant to anthropogenically forced climate change. Of course this test is of only limited use because it does not distinguish between the internally generated and externally driven components of the variability. The issue is that the GCM may be producing similar variability for the wrong reason, for example, by overestimating the externally forced response and underestimating the internally generated variability.

A first step in improving this comparison is to remove the GCM's estimate of the total externally forced response. If we assume the GCM is producing the same responses as the real world is, then we can remove the ensemble mean response of the GCM simulations from all of the time series. The resulting spectra are plotted in Fig. 1d. The spectra of the GCM simulations have been scaled by a factor of (62/61) to account for the reduction in variance due to the removal of the mean response. Because a large part of the externally forced variability has been removed, there is now a much closer agreement between the estimate of the confidence range of the observed spectrum and the spread of the simulated spectra at long time scales. We now find that the decadal variability in the observations is significantly larger than in the GCM simulations. This suggests that the GCM overly damps long-term variability. However, it could equally reflect an underestimate or overestimate in the GCM's response to external forcing and thus that we have not properly removed the total forced response from the observed time series. To elaborate on this comparison we need to produce a more confident estimate of the forced response, which will require us to first determine the GCM responses to

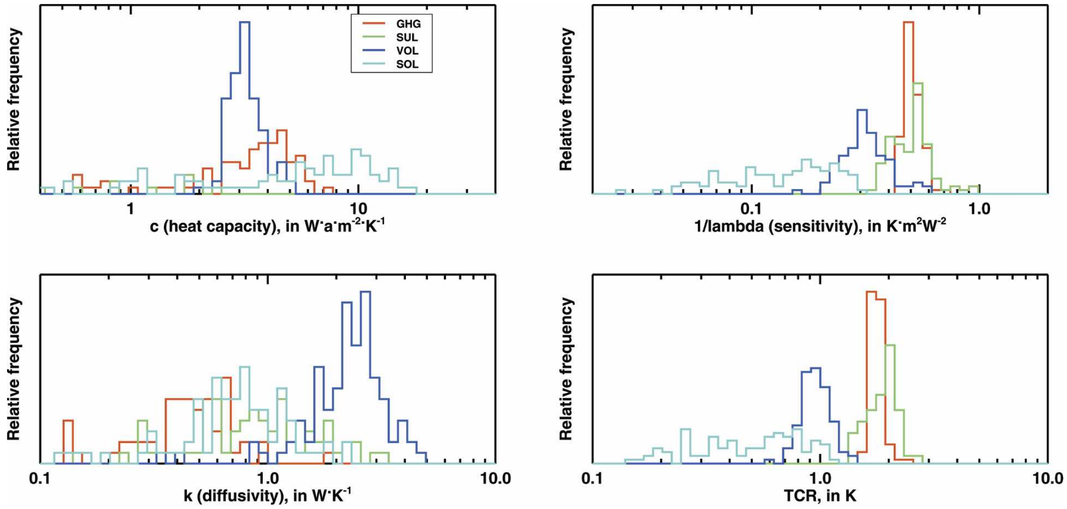

FIG. 2. The histograms of best-fit EBM parameters for each of the forcings: (top left) c_i, the heat capacity of the ocean mixed layer for forcing i; (top right) $(1/\lambda_i)$, the climate sensitivity; and (bottom left) k_i, the parameter for vertical diffusion in the mixed layer. (bottom right) The histogram of the estimated TCR resulting from these parameter sets. The TCR is the temperature increase from the forcing arising from an equivalent 1% yr^{-1} increase in CO$_2$ concentrations up to a doubling after 70 yr. The horizontal scale is identical on all plots. Note that the histograms follow logarithmic bins and so are not normalized on a linear scale.

each of the forcings and then to compare the response patterns with the observed SAT time series.

b. Fitting the EBMs

Figure 2 shows the histograms of the c_i, $(1/\lambda_i)$, and k_i EBM parameters from the bootstrap resampling. The estimates of the heat capacities are not very constrained, except for the VOL estimates, which range by only about a factor of 2. Some differences appear to exist across different forcings, with values for SOL tending to be higher and those for SUL tending to be lower (and generally outside the plotting range). On the other hand, the estimates of the climate sensitivities are generally tighter. While estimates for sensitivities for GHG and SUL tend to have similar values, the VOL estimates tend to be lower and the SOL estimates lower still. The best guess of the equilibrium response to a GHG increase equivalent to a doubling of CO$_2$ is 2.0 K (3.97 W m^{-2} per CO$_2$ doubling times 0.50 K m^2 W^{-1} climate sensitivity), matching the 2.1 K diagnosed by Meehl et al. (2000) (both values are for unmasked data). The estimates of the k_i vertical diffusion parameter are ill constrained across forcings, but values for VOL are generally higher.

A more relevant quantity for characterizing transient climates is the transient climate response (TCR) (Allen et al. 2005). This is the temperature change after an increase in forcing equal to a doubling of CO$_2$ at a rate of increase in emissions of 1% yr^{-1} (over 70 yr). While the TCR is defined for and traditionally applied to GHG forced changes, it can just as easily be applied to the other forcings examined here; even though it becomes more hypothetical in particular for VOL forcing, the same issue also applies to interpreting the climate sensitivity. Histograms of estimates of the TCR are also shown in Fig. 2. Like the climate sensitivity, the TCR is generally more tightly constrained than are the other parameters. While the TCR estimates are similar for GHG and SUL, the estimates are consistently lower for VOL and SOL. The best guess of the TCR for GHG, estimated by integrating the EBM, is 1.8 K compared to the 1.43 K found by Meehl et al. (2000) for this GCM (both values are for unmasked data). Thus, while much uncertainty exists in their structure, the EBM surrogates appear to be fairly appropriate in representing the basic transient response behavior of this GCM.

The 90% confidence intervals in the EBM responses to each forcing are shown in Fig. 3. The spread results entirely from the uncertainty in the EBM parameter sets. These responses add to produce the combined EBM response whose 90% confidence interval is plotted in Fig. 1b. The uncertainty in the combined response is smaller than in the individual responses because of partial degeneracy in the forcing profiles (e.g., between GHG and SUL); because EBM parameters are tuned simultaneously, the uncertainties in individual responses tend to cancel in the combined response. As expected, this combined EBM response

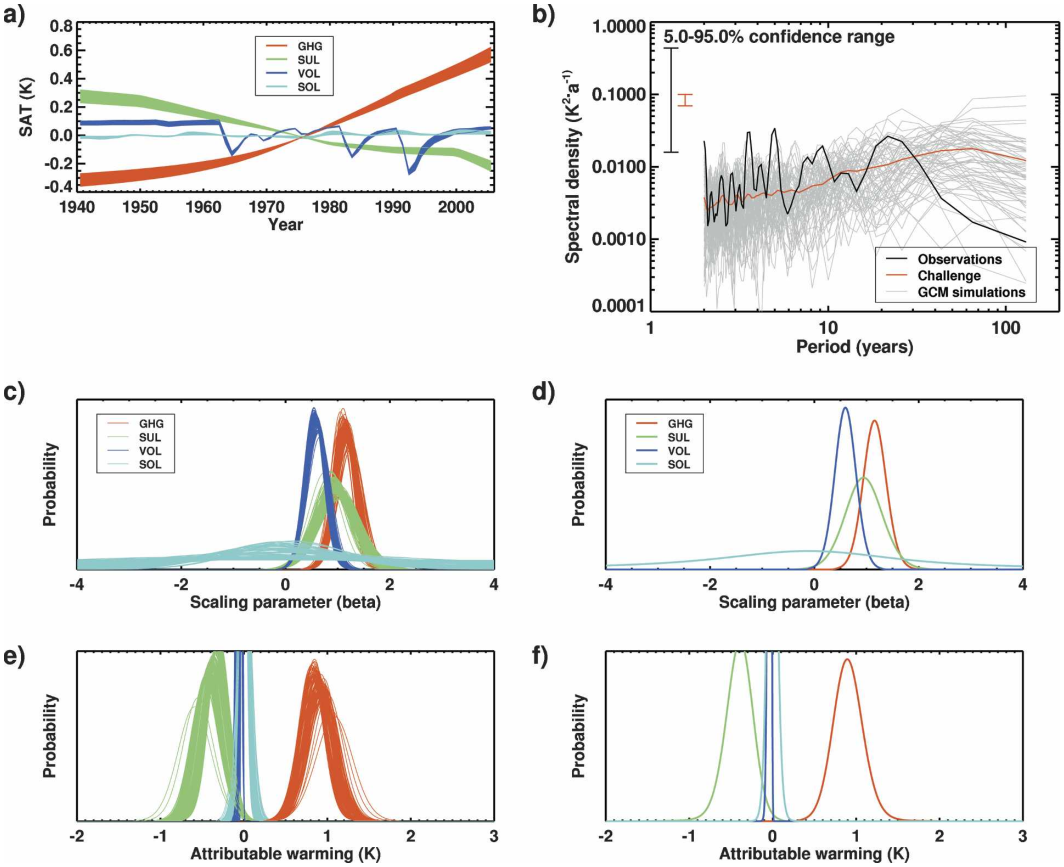

FIG. 3. (a) The 90% confidence range on the best-fit EBMs for each of the individual forcings. Values are shown as anomalies from the 1961–90 average. (b) Same as in Fig. 1d but with the best-guess scaled EBM fit responses removed from the observations before estimation of the spectrum. The observed variability is not inconsistent with the simulated variability on time scales of 10 yr and longer. (c) The distributions of the amplitude scalings (β_i) derived from the regression of observed SAT changes onto the EBM estimates of the GCM response patterns to each of the four forcings included in the simulations. The distributions are shown for each of the Monte Carlo estimates of the EBM fits. (d) Same as in (c) but with the average distributions across Monte Carlo samples. (e) The distributions resulting from (c) of the estimated change in SAT in the 1996–2005 decade relative to the 1940–49 decade attributable to each of the forcings. (f) Same as in (e) but with the average distributions across Monte Carlo samples.

closely follows the GCM ensemble mean response. While the GCM ensemble mean does seem to lie outside the EBM envelope more often than may be expected from random variability, the differences are small compared to the main features of the time series.

c. Scaling the model responses to observations

Now that we have an estimate of the GCM responses to each of the forcings we can proceed with the traditional detection methodology outlined in section 3. The univariate distributions of the regression scaling parameters (β_i) are shown in Fig. 3c for each forcing and each Monte Carlo realization. The average distributions across Monte Carlo realizations are shown in Fig. 3d. An obvious feature is that generally for each forcing the major source of uncertainty is in the fingerprinting regression rather than the Monte Carlo estimate of the EBM representation of the GCM, because the spread of best guesses is smaller than the width of the individual distributions (although this is not so clear for SOL). The distributions for GHG, SUL, and VOL indicate that the scaling factors are significantly different from zero at the 5% level, implying that the inclusion of these forcings is necessary in order to adequately represent the observed record. The distributions for GHG and SUL also indicate a scaling factor consistent with a value of one, meaning that an adequate representation of the observed record can be produced by the GCM without any alteration to the amplitude of these response patterns. The VOL scaling is significantly different from one though, indicating that the GCM has a tendency to overestimate the VOL response amplitude, as may be guessed from visual inspection of Fig. 1b. On the other hand, the SOL scaling is poorly constrained and the observations are consistent both with its absence and presence. Considering that the historical evolution of SOL forcing is poorly known, this result may just as easily be indicative of a problem with the forcing scenario used here as of a problem with the GCM or methodology.

d. Comparison of internally generated variability

The residual variability after removal of the scaled response patterns from the observations is significantly higher (based on an F test) than the variability in the first subset of 31 pseudocontrol simulations, produced by subtracting the ensemble mean change from the first 31 transient simulations. We can also examine this residual in the context of the spectra we were examining earlier. The spectra of the residuals, estimated by subtracting the best-guess EBM fit adjusted by the appropriate scalings from the observations, are plotted in Fig. 3b. The variability at interdecadal time scales in the observations does not differ significantly from the pseudocontrol simulations. Inspection of the spectra suggests the main discrepancy is at time scales of 3–6 yr, which are less important for the slowly varying anthropogenic forcings than for the more rapidly varying natural forcings.

e. Attributable warming to present

With estimates of the scaling factors, we can now estimate the amount of SAT change between 2005 and 1940 attributable to the various forcings (Figs. 3e,f). For consistency with other studies and because of the nonlinear nature of the response patterns, we estimate the attributable warming as the difference between the 1996–2005 mean SAT and the 1940–49 mean SAT in the scaled EBM response. GHG forcing dominates over this interval, contributing a significant warming of around 0.6–1.2 K. This is partly countered by a significant cooling from SUL forcing of around a quarter to a third that magnitude. The SOL and VOL forcings contribute little because their values are almost identical in the two decades compared.

f. Projections of global mean climate

If we suppose that the EBM approximation holds beyond the historical climate and into future climates, and that the linear additivity of the forcings also holds, then we can also extend our estimates of attributable warming into the future if we suppose a certain scenario of forcing for the future (Allen et al. 2000; Stott and Kettleborough 2002). The GCM has only been used to determine the temporal response pattern of the climate system to the various forcings, but not the amplitudes of those patterns. Thus, if the basic assumptions of the fingerprinting methodology hold, such scaled estimates of future warming are in fact constrained by the observed climate record, and not by the GCM itself.

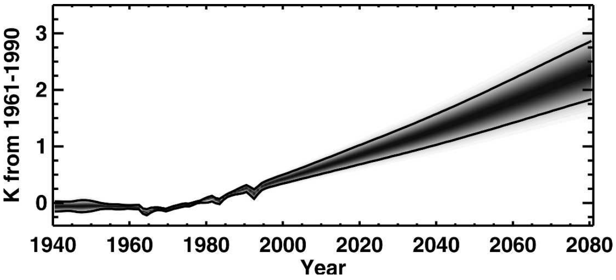

FIG. 4. The PDF of the global annual mean SAT climate from 1940 to present and through 2080 assuming the future business-as-usual emissions scenario, similar to the IPCC SRES A1 scenario, used in the Challenge ensemble. The lines show the 90% confidence range in this evolving PDF, which is the average of 100 Monte Carlo bootstrap sample PDFs and is standardized each year to have a standard peak amplitude for visibility.

The temporal evolution of the estimated probability density function (PDF) on this constrained hindcast and prediction is shown in Fig. 4. For each year, the PDF is estimated as the average across those for each Monte Carlo sample; the amplitudes of the averaged PDFs for each year are scaled to have a standard peak value in order to aid visibility into the future. Over the historical period, the effects of volcanic eruptions are clearly seen as dips. The prediction in the future follows the forcing scenario used in the Challenge ensemble, which resembles the IPCC SRES A1 scenario for the GHG forcing but is constant for all other forcings at year-2000 values (Dai et al. 2001). Uncertainties increase over time such that by 2080 the 90% confidence range is 1.8–2.9 K.

There is a lower uncertainty range in this prediction estimate than for other studies such as Stott and Kettleborough (2002) due to the use of a different forcing scenario (which assumes constant SUL forcing and thus no growing uncertainty from it) and because natural variability has not been added to the distribution here [Stott and Kettleborough (2002) include estimates of natural externally forced variability and of internally generated variability]. Overall, this prediction takes account of sampling uncertainty in both the observations

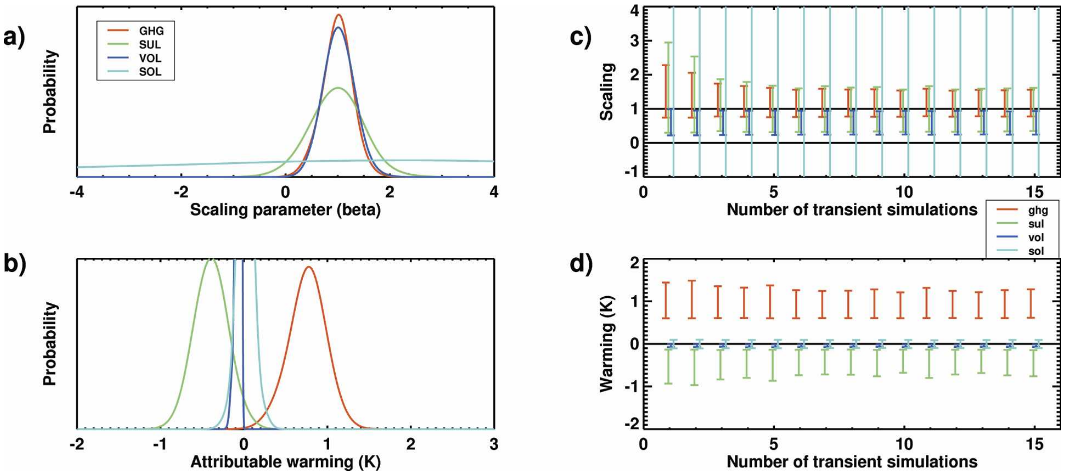

FIG. 5. (a), (b) Same as in Figs. 3d,f but replacing the observations with random GCM simulations. Also shown are the 90% confidence intervals on the (c) β_i scaling parameters and the (d) 1996–2005 vs 1940–49 attributable warming as a function of the number of historical GCM simulations used.

and the GCM simulations. It does not, of course, account for uncertainty in past or future emissions and forcings. While deficiencies in the EBM fitting to the GCM output could represent some uncertainty arising from the structural differences between the GCM and the real world, this prediction should not be considered to comprehensively account for that uncertainty.

5. Sensitivity of results

The technique developed in this paper involves adding another fitting step into the detection and attribution framework, which adds more potential uncertainty and possibly decreases the robustness of the results. In this section we test the robustness of the results to some changes in the input and methodology.

a. Substituting observations with GCM simulations

A first question is whether the method is estimating the correct values for the detection results. This can be tested using a perfect model setup, whereby the observations are replaced by a single random simulation for each of the 100 Monte Carlo samples. Because the GCM is implicitly reproducing the correct response pattern and amplitude in this setup, the scaling parameters should be centered on a value of one. The resulting average probability distribution of scaling values is shown in Fig. 5a. As expected, the scaling parameter distributions are all centered on a value of about one (except possibly for the very broad SOL). As with the observations, the GHG and VOL responses are the most tightly constrained, but in general the width of the distributions is wider than when the comparison is against the observations. This probably arises because we have used multiple simulations in place of the observations throughout the different Monte Carlo samples, rather than repetitively using the same simulation. The inability to constrain the SOL scaling in this perfect model exercise suggests that the SOL response simply has too small a signal-to-noise ratio to be detected. Notably, the residual variability in about 30% of the Monte Carlo samples is inconsistent at the 10% level with the residual variability; this usually involves a smaller residual, highlighting a bias in the methodology.

b. Number of transient simulations

Most applications of this methodology will be with GCM ensembles with far fewer than 62 simulations. With such smaller ensembles, the EBM fits might not be expected to be as accurate due to the larger sampling uncertainty. Furthermore, if the ensembles are quite small (e.g., one member) then the Monte Carlo approach used here to quantify this uncertainty cannot be applied, so it would be useful to use the large ensemble here to characterize this component of the uncertainty. The estimated confidence ranges on the scaling parameters for different ensemble sizes ranging from 1 to 15 members are shown in Figs. 5c and 5d. In this analysis, the given number of simulations was randomly selected from all of the 62 members of the Challenge ensemble with replacement; otherwise the analysis is identical to before. As in the full 62-member ensemble analysis, GHG, SUL, and VOL responses are detected for all ensemble sizes. Beyond ensembles of three, the scaling

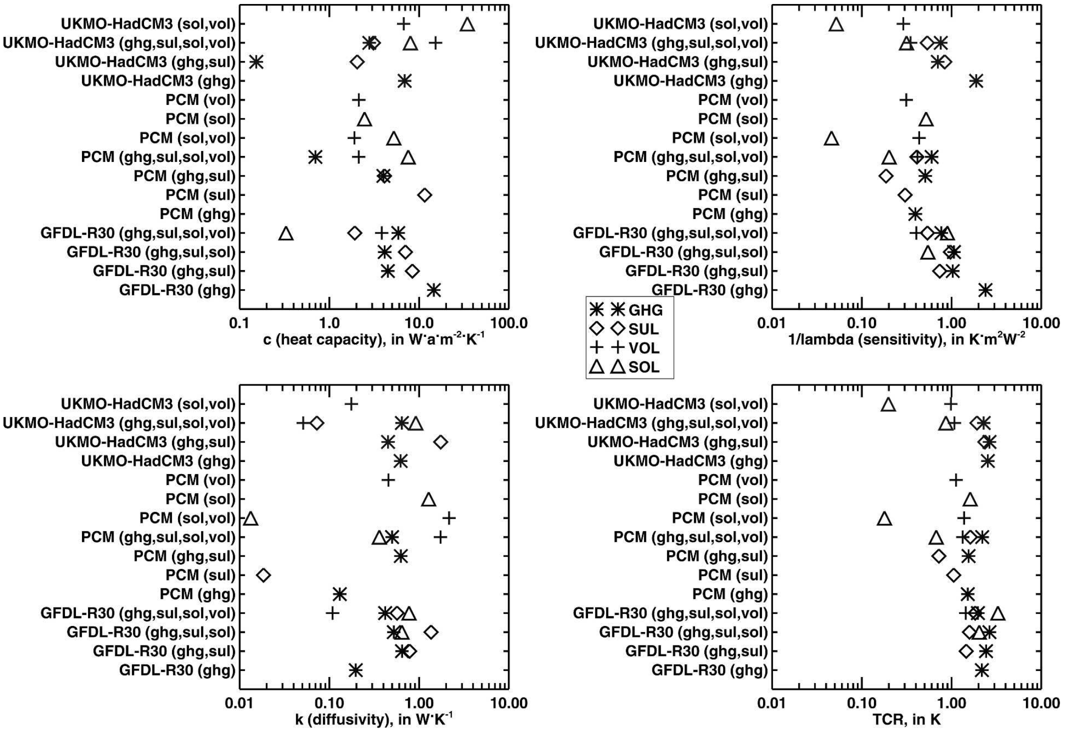

FIG. 6. Similar to Fig. 2, but showing instead the best-guess parameter sets for the EBM fits to various GFDL-R30, PCM1, and UKMO-HadCM3 GCM ensembles. The resulting TCR estimates are also shown.

parameter values are only marginally more uncertain when fewer simulations are used. Attributable warming values show no noticeable changes. This indicates that the main source of uncertainty is not sampling of the GCM but sampling of the observations and/or limitations in the applicability of the one-dimensional EBM surrogate to the GCM. In particular there may be a small limit to the amount of information contained in the globally averaged data used here, and this limit is reached with ensembles of about three members. This is a limitation of the temporal methodology used here versus the standard attribution method, in that tighter constraints than those found here can be obtained using spatiotemporal data from multiple simulations (Stott et al. 2006).

c. Consistency with the standard attribution method

Multiple ensembles of simulations with different forcing scenario combinations exist from the GFDL-R30, PCM1, and UKMO-HadCM3 GCMs. These simulations can therefore be used both as an internal consistency test of the methodology applied here and as an external consistency test with the results from the standard fingerprinting method applied to temporal data only. In the analysis of these GCMs, we use the period starting in 1901 and ending in 1995–2002, depending upon the length of available of simulations. The forcing scenarios used are provided by PCMDI [see Stone et al. (2007) for more information].

The best-guess estimates of the EBM parameter values from these ensembles are shown in Fig. 6. The ensembles are too small for a Monte Carlo method to be used to estimate distributions. In general, there is a wide spread in parameter values across scenarios with the same GCM. This is not surprising considering the large spread in values in Fig. 2 obtained with a much larger ensemble. Parameters associated with GHG or VOL forcings tend to be more consistent across ensembles than other parameters, but some spread still exists. As in Fig. 2, estimates of the TCR are more constrained across ensembles (Fig. 6).

Figures 7a–c show the best-fit EBM responses to each of the forcings for each of the three models when fit against the ensemble of simulations including all four forcings. The ensemble mean responses from simulations forced only with individual forcings are shown for comparison. Because of the availability of simulations, the VOL and SOL forcings were combined into a single forcing when estimating these EBM responses. Because the EBMs were estimated from a separate ensemble of simulations forced with all four forcings, this plot provides an independent test of the adequacy of the EBM

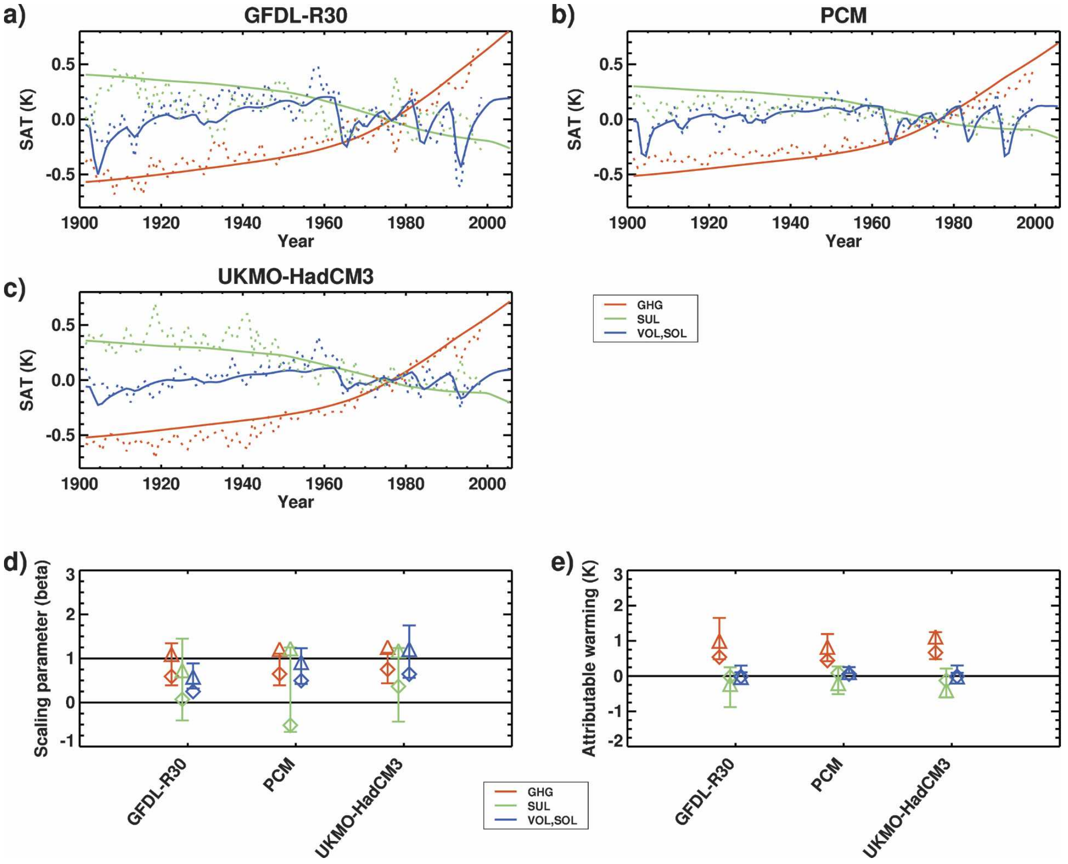

FIG. 7. (a)–(c) Similar to Fig. 3a but instead showing the individual components of the best-guess EBM fits to ensembles of simulations of the three respective GCMs forced with all four forcings (solid lines). Unlike Fig. 3a, VOL and SOL forcings have been treated as a single forcing in these estimates. Dotted lines show the ensemble averages of actual simulations forced with only the indicated forcing. (d), (e) Similar to Figs. 3c,e but instead showing the 90% confidence ranges on the estimated regression scaling parameters and attributable warming for various ensembles of the GFDL-R30, PCM1, and UKMO-HadCM3 GCMs. The bars represent the estimates using the EBM fitting technique, with VOL and SOL forcings treated as a single forcing. Because of the small ensemble sizes for these GCMs, no Monte Carlo sampling is included in these estimates. The diamonds and triangles represent the lower and upper bounds on scaling estimates and attributable warmings obtained using the traditional OLS regression methodology applied to global mean time series of three ensembles forced with different combinations of the three forcings considered.

fits. The GHG and SUL responses appear to be slightly overestimated for PCM1, but otherwise the estimated EBM responses fit the actual responses nicely.

Estimates of the regression scaling parameters for these three GCMs are shown in Fig. 7d. Estimates from the EBM fitting methodology applied to simulations forced with all four forcings are shown, as well as estimates from the traditional regression methodology applied to the GHG, GHG + SUL, and GHG + SUL + VOL + SOL ensembles. GHG and VOL + SOL responses are detected in all three GCMs, while both 0 and 1 are consistent with the SUL scalings. Uncertainties are always smaller for results from the traditional methodology, which is not surprising considering the extra step and use of a single ensemble in the EBM method. Satisfyingly, results from the two methods always entirely or almost entirely overlap.

6. Discussion

This paper introduces an extension to the traditional detection and attribution methodology that allows the multisignal analysis of historical simulations from GCMs even when only a single ensemble of simulations forced with a single forcing combination is available. Tests of this extension indicate some issues; nevertheless the constraints provided by the observations appear to overwhelm any such problems, supporting the robustness of the fingerprinting approach.

The EBM approximation indicates that the GCM used in the Challenge ensemble responds differently to most forcings. While the estimated GCM responses to GHG, SUL, and VOL forcing are detected in the observed record, the response to SOL forcing is not. The observational constraints on the responses indicate that recent climate change has been driven by GHG forcing, with a partial counteracting effect from SUL forcing, and thus provides constraints on future warming.

CR06 also develop a technique for extracting multisignal attribution results from ensembles of a single forcing scenario. Their technique applies a space–time separable approach to extract spatial response patterns, rather than the temporal patterns used here. When applied to identical GCM simulations both methods produce broadly similar results, although CR06 seem more likely to detect SUL and SOL responses but not VOL responses, while the EBM fitting method tends to find the opposite. The difference in the VOL detection appears at least partly due to the use of forcing time series, rather than response time series, to estimate the spatial response pattern in the method of CR06. We plan in future work to incorporate the EBM fits developed here into the method of CR06 in order to develop a more comprehensive attribution methodology.

The methodology used here requires EBMs to be fitted to GCM output and one would like the tuned parameters in the EBMs to be robustly constrained. Unfortunately that does not appear to be the case here. However, the parameter sets tend to lead to EBMs with relatively more robust behavior, such as that characterized by the TCR and by the response pattern. The latter is judged by the ability of the regression to consistently detect and scale the response pattern appropriately. Consequently, detection and attributable warming results appear fairly robust.

The extended methodology developed here should not be considered a replacement for the traditional multisignal detection and attribution procedure. Nevertheless, the tests indicate that the new step is fulfilling its purpose in providing an adequate surrogate for the GCM response behavior. Thus application of this technique will allow the inclusion of many more GCMs into the detection and attribution framework and permit a more robust quantification of the contribution of external forcings to past and future climate change.

Acknowledgments. The authors wish to thank Caspar Ammann, Simon Crooks, Mark Hawkins, and Hugo Lambert for their help with the analysis. We also thank Francis Zwiers and two anonymous reviewers for many helpful comments on earlier versions of the manuscript. We acknowledge the HadCRUT2v dataset provided by the Climate Research Unit (UEA), the GFDL-R30 simulation output provided by the Geophysical Fluid Dynamics Laboratory (NOAA), and the PCM1 simulation output provided by the Climate and Global Dynamics Division (UCAR). D. A. S. and P. A. S. were supported by the U.K. Department for Environment, Food, and Rural Affairs, and D. A. S. was also partially supported by a Wellcome Trust Showcase Award. M. R. A. received partial support from the U.S. NOAA/DoE International Detection and Attribution Group.

REFERENCES

Allen, M., 2003: Liability for climate change. *Nature,* **421,** 891–892.

——, and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dyn.,* **15,** 419–434.

——, and R. Lord, 2004: The blame game. *Nature,* **432,** 551–552.

——, P. A. Stott, J. F. B. Mitchell, R. Schnur, and T. L. Delworth, 2000: Uncertainty in forecasts of anthropogenic climate change. *Nature,* **407,** 617–620.

——, and Coauthors, 2005: Observational constraints on climate sensitivity. *Avoiding Dangerous Climate Change,* J. S. Schellnhuber et al., Eds., Cambridge University Press, 281–289.

Boucher, O., and M. Pham, 2002: History of sulfate aerosol radiative forcings. *Geophys. Res. Lett.,* **29,** 1308, doi:10.1029/2001GL014048.

Boville, B. A., J. T. Kiehl, P. J. Rasch, and F. O. Bryan, 2001: Improvements to the NCAR CSM-1 for transient climate simulations. *J. Climate,* **14,** 164–179.

Broccoli, A. J., K. W. Dixon, T. D. Delworth, T. R. Knutson, R. J. Stouffer, and F. Zeng, 2003: Twentieth-century temperature and precipitation trends in ensemble climate simulations including natural and anthropogenic forcing. *J. Geophys. Res.,* **108,** 4798, doi:10.1029/2003JD003812.

Dai, A., T. M. L. Wigley, B. A. Boville, J. T. Kiehl, and L. E. Buja, 2001: Climate of the twentieth and twenty-first centuries simulated by the NCAR Climate System Model. *J. Climate,* **14,** 485–519.

Gillett, N. P., M. F. Wehner, S. F. B. Tett, and A. J. Weaver, 2004: Testing the linearity of the reponse to combined greenhouse gas and sulfate aerosol forcing. *Geophys. Res. Lett.,* **31,** L14201, doi:10.1029/2004GL020111.

Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, Eds., 2001: *Climate Change 2001: The Scientific Basis.* Cambridge University Press, 881 pp.

Hoyt, D. V., and K. H. Schatten, 1993: A discussion of plausible solar irradiance variations, 1700–1992. *J. Geophys. Res.,* **98,** 18 895–18 906.

International Ad Hoc Detection and Attribution Group, 2005: Detecting and attributing external influences on the climate system: A review of recent advances. *J. Climate,* **18,** 1291–1314.

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

McAvaney, B. J., and Coauthors, 2001: Model evaluation. *Climate Change 2001: The Scientific Basis,* J. T. Houghton et al., Eds., Cambridge University Press, 471–524.

Meehl, G. A., W. D. Collins, B. A. Boville, J. T. Kiehl, T. M. L. Wigley, and J. M. Arblaster, 2000: Response of the NCAR Climate System Model to increased CO_2 and the role of physical processes. *J. Climate,* **13,** 1879–1898.

——, W. M. Washington, T. M. L. Wigley, J. M. Arblaster, and A. Dai, 2003: Solar and greenhouse gas forcing and climate response in the twentieth century. *J. Climate,* **16,** 426–444.

Nelder, J. A., and R. Mead, 1965: A simplex method for function minimization. *Comput. J.,* **7,** 308–313.

Ramaswamy, V., and Coauthors, 2001: Radiative forcing of climate change. *Climate Change 2001: The Scientific Basis,* J. T. Houghton et al., Eds., Cambridge University Press, 349–416.

Rasch, P. J., M. C. Barth, J. T. Kiehl, S. E. Schwartz, and C. M. Benkovitz, 2000: A description of the global sulfur cycle and its controlling processes in the National Center for Atmospheric Research Community Climate Model, Version 3. *J. Geophys. Res.,* **105,** 1367–1385.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice and night marine air temperature since the late nineteenth century. *J. Geophys. Res.,* **108,** 4407, doi:10.1029/2002JD002670.

Selten, F., M. Kliphuis, and H. Dijkstra, 2003: Transient coupled ensemble climate simulations to study changes in the probability of extreme events. *CLIVAR Exchanges,* Vol. 8, No. 4, International CLIVAR Project Office, Southampton, United Kingdom, 11–13.

Stone, D. A., M. R. Allen, and P. A. Stott, 2007: A multimodel update on the detection and attribution of global surface warming. *J. Climate,* **20,** 517–530.

Stott, P. A., and J. A. Kettleborough, 2002: Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature,* **416,** 723–726.

——, S. F. B. Tett, G. S. Jones, M. R. Allen, J. F. B. Mitchell, and G. J. Jenkins, 2000: External control of 20th century temperature by natural and anthropogenic forcings. *Science,* **290,** 2133–2137.

——, J. F. B. Mitchell, J. M. Gregory, B. D. Santer, G. A. Meehl, T. L. Delworth, and M. R. Allen, 2006: Observational constraints on past attributable warming and predictions of future global warming. *J. Climate,* **19,** 3055–3069.

Tett, S. F. B., P. A. Stott, M. R. Allen, W. J. Ingram, and J. F. B. Mitchell, 1999: Causes of twentieth-century temperature change near the Earth's surface. *Nature,* **399,** 569–572.

——, and Coauthors, 2002: Estimation of natural and anthropogenic contributions to twentieth century temperature change. *J. Geophys. Res.,* **107,** 4306, doi:10.1029/2000JD000028.

Wigley, T. M. L., C. Ammann, B. D. Santer, and S. C. B. Raper, 2005: Effect of climate sensitivity on the response to volcanic forcing. *J. Geophys. Res.,* **110,** D09107, doi:10.1029/2004JD005557.