A Bayesian Assessment of Climate Change Using Multimodel Ensembles. Part I: Global Mean Surface Temperature

SEUNG-KI MIN AND ANDREAS HENSE

Meteorologisches Institut, Universität Bonn, Bonn, Germany

(Manuscript received 25 April 2005, in final form 2 November 2005)

ABSTRACT

A Bayesian approach is applied to the observed global surface air temperature (SAT) changes using multimodel ensembles (MMEs) of the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) simulations and single-model ensembles (SMEs) with the ECHO-G coupled climate model. A Bayesian decision method is used as a tool for classifying observations into given scenarios (or hypotheses). The prior probability of the scenarios, which represents a degree of subjective belief in the scenarios, is changed into the posterior probability through the likelihood where observations enter, and the posterior is used as a decision function. In the identical prior case the Bayes factor (or likelihood ratio) becomes a decision function and provides observational evidence for each scenario against a predefined reference scenario. Four scenarios are used to explain observed SAT changes: "CTL" (control or no change), "Nat" (natural forcing induced change), "GHG" (greenhouse gas–induced change), and "All" (natural plus anthropogenic forcing–induced change). Observed and simulated global mean SATs are decomposed into temporal components of overall mean, linear trend, and decadal variabilities through Legendre series expansions, coefficients of which are used as detection variables. Parameters (means and covariance matrices) needed to define the four scenarios are estimated from SMEs or MMEs. Taking the CTL scenario as reference one, application results for global mean SAT changes for the whole twentieth century (1900–99) show "decisive" evidence (logarithm of Bayes factor >5) for the All scenario only. While "strong" evidence (log of Bayes factor >2.5) for both the Nat and All scenarios are found in SAT changes for the first half (1900–49), there is decisive evidence for the All scenario for SAT changes in the second half (1950–99), supporting previous results. It is demonstrated that the Bayesian decision results for global mean SATs are largely insensitive to both intermodel uncertainties and prior probabilities.

1. Introduction

Recently, the International Ad Hoc Detection and Attribution Group (IDAG) reviewed the advances in the studies on climate change detection and attribution (International Ad Hoc Detection and Attribution Group 2005, hereafter IDAG05). They summarized anthropogenic signals found in different observational datasets: global and regional surface air temperature (SAT), vertical temperature, tropopause height, sea level pressure, and ocean heat content (IDAG05, and references therein). This summary strengthens the previous conclusion by the Third Assessment Report (TAR) of the Intergovernmental Panel on Climate Change (IPCC) that "most of the observed warming over the last 50 years is likely to have been due to the increase in greenhouse gas concentrations" (from the Summary for Policymakers in Houghton et al. 2001). IDAG05 also showed increasing efforts to apply Bayesian statistical approaches and pointed out their advantages to integrate information from multiple evidences as well as prior information and to provide probabilistic output for decision making and a better quantification of the evidence for attribution. Importantly, they enable the users to consider errors in the signal (or model response) as well as natural climate variability. The latter is the only source of uncertainties in conventional statistics although observational and model uncertainties could be assessed from sensitivity tests or intercomparisons (Barnett et al. 1999; Hegerl et al. 2000, 2001).

The uncertainties arising from intermodel differences could not be assessed reasonably due to lack of enough

Corresponding author address: Seung-Ki Min, Meteorologisches Institut, Universität Bonn, Auf dem Hügel 20, 53121 Bonn, Germany.
E-mail: skmin@uni-bonn.de

TABLE 1. Climate change scenarios for Bayesian decision analysis. PD: present-day control run; PI: preindustrial control run (= PIcntrl); Nat: natural forcing run; 20C: 20C3M run with natural plus anthropogenic forcing; and MME: multimodel ensembles of IPCC AR4 simulations

Number	Abbreviation	Forcing	Relevant model simulations used for "mean" estimation
1	CTL	Natural internal variability (control)	ECHO-G_PD, MME_PI
2	GHG	Greenhouse gases	ECHO-G_GHG
3	Nat	Natural forcing (solar and volcanic activities)	ECHO-G_Nat
4	All	Natural plus anthropogenic forcing (GHGs and sulfate aerosols)	ECHO-G_20C, MME_20C

samples of model simulations. Recently, a dataset from multimodel ensemble (MME) simulations became available through the international project coordinated by the IPCC Working Group 1. More than 20 international modeling groups and institutes participated in this project. They carried out ensemble simulations with their atmosphere–ocean general circulation models (AOGCMs) based on IPCC scenarios for the Fourth Assessment Report (AR4).

Recent Bayesian studies for climate change detection and attribution are all based on Bayes factors or likelihood ratios (Min et al. 2004, hereafter M04, 2005a; Schnur and Hasselmann 2005; Lee et al. 2005). The Bayes factors represent the observational evidence for the given scenario (e.g., greenhouse gas–forced climate change) against a reference scenario (e.g., no climate change or control scenario). Schnur and Hasselmann (2005) devised an optimal filtering technique for a Bayesian approach, which maximizes the impact of the observations on the prior likelihood of detection or the Bayes factors, and detected anthropogenic signals of greenhouse gases (G) or greenhouse gases plus sulfate aerosols (GS) in global SAT patterns. Lee et al. (2005) applied the Bayesian extension of conventional optimal fingerprinting of Berliner et al. (2000) by evaluating detection and attribution hypotheses using the Bayes factors and detected GS signals in global SAT fields. M04 developed a Bayesian decision method for climate change signal analysis where, for given scenarios and observations, prior information is changed into posterior through likelihood, and the posterior acts as a decision function. Applying the Bayesian decision method, Min et al. (2005a) detected G signals in East Asian SAT fields.

In this study, we apply the Bayesian decision method of M04 to the data from single-model ensembles (SMEs) and MMEs to test the sensitivity of Bayesian climate change assessment to intermodel uncertainties. As a first step, time series of global mean SATs are used as detection variables. The temporal variability is incorporated by decomposition of the time series using Legendre series expansions. This allows for an effective handling of overall warming (*scale*) and a warming trend (*trend*) (see below). Four scenarios are defined to test how well they can explain the observed SAT changes over the twentieth century: "CTL," "GHG," "Nat," and "All" (see Table 1 for detailed description of the scenarios). The parameters for the four scenarios are estimated from SMEs with the ECHO-G coupled climate model (Legutke and Voss 1999; Min et al. 2005b,c, 2006) or MMEs using the IPCC AR4 models. The sensitivity of Bayesian decision results to intermodel uncertainties is examined by comparing the results from the SMEs with the MMEs. We start with the case of identical priors for the scenarios where the Bayes factor can be used as a decision function. Next a generalized Bayesian decision is done with varying priors of the given scenarios, in which the posterior probability is analyzed for observed signal classification.

Considering limitations to the detection and attribution of climate change that is based solely on the global mean regional and seasonal extensions using space–time SAT data vectors will be analyzed in Part II of this paper (Min and Hense 2006, manuscript submitted to *J. Climate*). A space–time data vector can be readily constructed, for example, by combining Legendre coefficients of regional mean SATs for two or three subregions that constitute a continental-scale region. The block averages will reduce the spatial degree of freedom while the Legendre expansions concern temporal dimensions.

In the next section the Bayesian decision method and Legendre series expansions are described. Observations and model simulations for ECHO-G and IPCC AR4 models are explained briefly in section 3. In section 4, detection variables from observations and model simulations are compared for different time scales using Legendre expansion coefficients. In section 5, the Bayesian decision results for global mean SAT changes are shown for the whole twentieth century and two subperiods of 1900–49 and 1950–99 focusing on the effect of intermodel uncertainties and prior probabilities. Conclusions are given with some discussions in the final section.

2. Methodology

a. Bayesian decision method

The Bayesian decision method by M04 is briefly explained here (see M04 for more details). Given a set of N possible scenarios (m_i, $i = 1, \ldots, N$; $N = 4$ in the present case) and the observational data ($\mathbf{d}$), an appropriate question on climate change detection and attribution will be, "How probable is the scenario m_i given the observation $\mathbf{d}$?" This can be expressed as a conditional probability $\mathrm{P}(m_i|\mathbf{d})$, which is the posterior probability of the scenario given the observation. Using Bayes' rule, this can be evaluated from the prior probability $\mathrm{P}(m_i)$, which represents a subjective belief in the scenario, and likelihood function $l(\mathbf{d}|m_i)$, which characterizes the observational probability given the scenario

$$\mathrm{P}(m_i|\mathbf{d}) = \frac{l(\mathbf{d}|m_i)\mathrm{P}(m_i)}{\sum_{j=1}^{N} l(\mathbf{d}|m_j)\mathrm{P}(m_j)}. \quad (1)$$

Assuming multivariate Gaussian distributions for the detection variables of the scenario m_i and the observations $\mathbf{d}$, the likelihood function can be expressed as

$$l(\mathbf{d}|m_i) = \frac{1}{\sqrt{(2\pi)^q}} \sqrt{\frac{\det \mathbf{A}_i^{-1}}{\det \boldsymbol{\Sigma}_i \det \boldsymbol{\Sigma}_0}} \exp\left(-\frac{1}{2}\Lambda_i\right), \quad (2)$$

where q is the dimension of the data vector $\mathbf{d}$, and $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_i$ are the covariance matrices of the observation $\mathbf{d}$ and the scenario m_i, respectively. Here, $\mathbf{A}_i$ is a linear combination of these covariance matrices, $\mathbf{A}_i = \boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_0^{-1}$, and Λ_i is a generalized distance measure between the observation and the scenario, $\Lambda_i = (\mathbf{d} - \mathbf{A}_i^{-1}\mathbf{b})^{\mathrm{T}} \boldsymbol{\Sigma}_0^{-1} (\mathbf{d} - \mathbf{A}_i^{-1}\mathbf{b}) + (\mathbf{A}_i^{-1}\mathbf{b} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{A}_i^{-1}\mathbf{b} - \boldsymbol{\mu}_i)$, where $\boldsymbol{\mu}_i$ is mean of the scenario m_i and $\mathbf{b} = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_0^{-1}\mathbf{d}$ (for more details see M04). In this study, Legendre expansion coefficients (see below) for global mean SATs from the Climate Research Unit (CRU) observation and model simulations under relevant forcing factors (see Table 1) are used to estimate necessary parameters of means ($\mathbf{d}$ and $\boldsymbol{\mu}_i$) and variabilities ($\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_i$) with changing time scales retained (q) through Legendre coefficients. On the basis of the result that the models used here can simulate the internal variability quite well on the decadal and longer time scales (see subsection 2b and Fig. 2), we assume that observed internal variability is identical to that of the CTL scenario ($\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$).

According to the Bayesian decision theory, the posterior probability can be used as a decision function (Duda and Hart 1973; Berger 1985). We decide the scenario with maximum posterior probability by which theoretical decision error becomes a minimum. The theoretical decision error is the overall risk arising from wrong decisions (or actions), and it can be calculated by integrating the expected loss over hypothetically defined observational space (Duda and Hart 1973; M04). Then the decision rule becomes

$$\text{Decide } m_i \text{ if } \mathrm{P}(m_i|\mathbf{d}) > \mathrm{P}(m_j|\mathbf{d}) \text{ for all } j \neq i \ (i, j = 1, \ldots, N). \quad (3)$$

As a tool for quantifying observational evidence in the Bayesian framework, Kass and Raftery (1995) suggested the Bayes factor, the original idea of which was developed by Jeffreys (1935, 1961). The Bayes factor is defined as the ratio of posterior odds to prior odds, and it is identical to the likelihood ratio in the particular setup used here where two scenarios (hypotheses) are single distributions with no free parameters (Kass and Raftery 1995):

$$B_{ir} = \frac{\mathrm{P}(m_i|\mathbf{d})/\mathrm{P}(m_r|\mathbf{d})}{\mathrm{P}(m_i)/\mathrm{P}(m_r)} = \frac{l(\mathbf{d}|m_i)}{l(\mathbf{d}|m_r)}. \quad (4)$$

The Bayes factor B_{ir} of the scenario m_i with respect to the reference scenario m_r represents observational evidence in favor of m_i or against m_r. It should be noted that this Bayes factor is independent of the prior probability while the posterior probability depends on priors. In other words, the Bayes factor corresponds to posterior odds in a special case of identical priors. Kass and Raftery (1995) suggested descriptive scales of Bayes factors (Table 2), which derive from those developed by Jeffreys (1961). If the logarithm of Bayes factors is larger than 1, 2.5, or 5, the observations represent "substantial," "strong," or "decisive" evidence in favor of the scenario m_i or against the reference scenario m_r. The interpretations of Bayes factors in M04, Min et al. (2005a), Schnur and Hasselmann (2005), and Lee et al. (2005) are based on these scales.

TABLE 2. Descriptive scales of Bayes factors after Kass and Raftery (1995).

$\ln B_{ir}$	B_{ir}	Evidence against the reference scenario (m_r)
0–1	1–3	Not worth more than a bare mention
1–2.5	3–12	Substantial
2.5–5	12–150	Strong
>5	>150	Decisive

It is assumed by definition that N possibilities (scenarios) are exhaustive, that is, their prior probabilities add to one. The Bayesian decision is made only for possible scenarios that are predefined. Therefore its re-

sult is definitely dependent upon the scenarios included, meaning that one should set possible scenarios carefully and that one cannot say anything about other scenarios that are excluded.

b. Legendre series expansions

Although we use global mean SAT as detection variables, we deal with a multivariate problem because we treat one time series of global mean SAT as the realization of a multivariate random variable. Therefore we have to summarize the temporal structure by an appropriate compression. The dimension reduction is necessary to avoid singular or near-singular covariance matrices. We propose the use of Legendre polynomials (LPs) $P_n(x)$, which have three advantages: 1) They can handle an overall warming (*scale*) and a linear trend (*trend*) in contrast to Fourier or Wavelet modes. 2) They are data independent in contrast to empirical orthogonal functions (EOFs) or singular spectrum analysis. 3) They are orthogonal in contrast to general polynomials. The LP method has a major advantage over the direct analysis of decadal means in terms of temporal dimension. While the latter defines a fixed time scale, the former covers all time scales up to the smallest resolved one, which allows examining relative roles of each temporal component in the sense of signal detectability. However, we admit that LPs also have some disadvantages, for example, they would explain less variance than EOFs would for a given number of modes, there might be odd behaviors of higher-order modes related to end effects, and there are multiple time scales associated with each LP that differ as one approaches the ends of the time series (i.e., get shorter).

It can be shown that any function $f(x)$ may be expanded in terms of $P_n(x)$ over the interval $[-1, 1]$ (Kaplan 1992, 508–512):

$$f(x) = \sum_{n=0}^{\infty} a_n P_n(x). \tag{5}$$

Using the orthogonality of the LP, the expansion coefficient a_n can be obtained from

$$a_n = \frac{2n+1}{2} \int_{-1}^{1} P_n(x) f(x)\, dx. \tag{6}$$

In this study, $f(x)$ over the interval $[-1, 1]$ corresponds to global mean monthly SAT anomalies (SATAs) over the time interval $[t_1, t_2]$ from either observation or model simulations. After transforming the SAT data, say $g(t)$ $[t_1, t_2]$, into $f(x)$ $[-1, 1]$, substituting $f(x)$ into Eq. (6) produces the coefficient a_n by summing over the data points.

If we expand time series of twentieth-century global mean SATA, the zeroth degree coefficient a_0 of $P_0(x) = 1$ (LP0) handles the overall time average (*scale*). The first degree coefficient a_1 of $P_1(x) = x$ (LP1) is directly proportional to the linear trend (*trend*). Higher-order coefficients characterize the decadal and interannual variability. As examples of Legendre series expansions, Fig. 1 shows reconstructed Legendre components $a_n P_n(x)$ for degrees from 0 to 12 of global mean SATA from observations (HadCRUT2v, Jones and Moberg 2003; see section 3) and different AOGCM simulations. The first 12 degrees of LPs explain about 62% of observed variances in global mean monthly SATs. Observations and model simulations show a global surface warming (*scale*) in the twentieth century by about 0.3°C with respect to 1900–20 means and positive trends (*trend*), both of which are located beyond the range of internal variability estimated from present-day control simulations with the ECHO-G model (denoted as ECHO-G_PD). The scale 0.3°C is different from the well-known warming 0.6°C published by Houghton et al. (2001) because the former is an averaged warming over the whole twentieth century relative to 1900–20 while the latter is a temperature change for 1901–2000 estimated from the linear trend. Observed and simulated SATA variations in higher degrees are located within the simulated internal variabilities except some AOGCMs in LP3 (see below). LP0 values can be arbitrary according to the defined reference period, but the comparison of LP0 coefficients between observations and model simulations is important as a relative distance measure between the observed and simulated warming over the whole period. If one changes the reference period, observed and simulated LP0 coefficients move together and the difference between them is not affected much.

We decide the truncation degree of LP by assessing whether the models simulate internal variability reasonably on the time scales that are retained. The model evaluation is done through analyzing power spectra of global mean SATs from model and observations after removing the linear trend. Figure 2 shows a comparison of power spectra between CRU observations and ECHO-G_PD and MME_PI (preindustrial control simulations with multimodels; see Table 3). The Blackman–Tukey method (Blackman and Tukey 1958) is used for performing spectral analysis with a Hamming window. A maximum lag of 25 yr is applied. One should note that there might be the problem of the short record for estimating the power in the 20-yr and longer time scale. Ranges of model spectra represent maximum and minimum values from ensemble members for a given time period. The range of statistical uncertainty

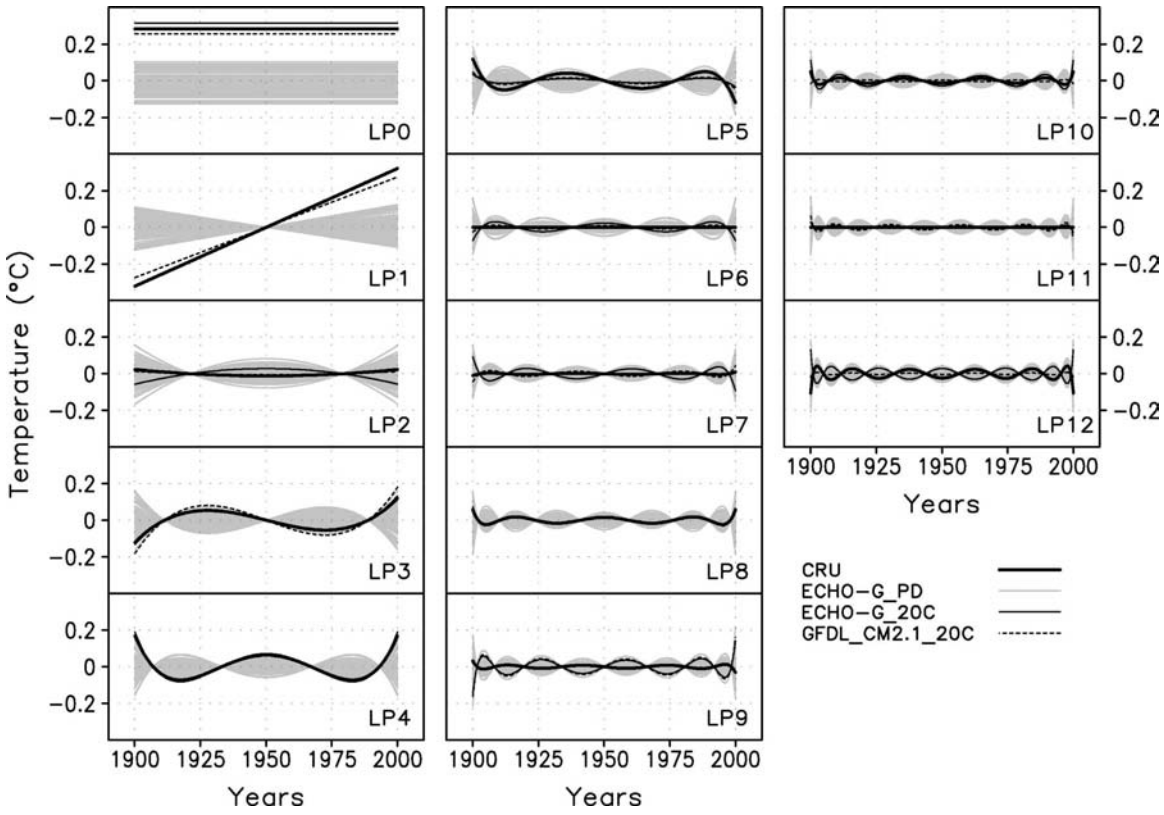

FIG. 1. Reconstructed time series of global mean SATs for 1900–99 from Legendre series expansions for Legendre degrees from 0 to 12: CRU observations (thick solid); selected two 20C3M simulations with different AOGCMs (thin solid and dashed), and 91 samples from ECHO-G 1000-yr present-day control run (ECHO-G_PD; gray). See text for details.

(95% confidence limits) in the spectral estimates is shown as a vertical line, which is identical for the observation and all model ensemble members of the same 100-yr length (i.e., 1200 months). ECHO-G_PD exhibits reasonable variance in the decadal time scales while it has a too-strong 2-yr peak and a weak power in the interannual time scale, which was reported by Min et al. (2005b). It is interesting to see that the uncertainty range from ECHO-G_PD looks similar to that of the statistical error although there are some points (e.g., near 5 yr) where the former is less than the latter. The MME result shows a broader range of modeled variance and the averaged power reveals better consistency with that observed. The range of multimodel spectra is larger than that of the statistical error in all periods. It is shown that the LP12 has a maximum power near 20 yr for the 100-yr period, which would be near 10 yr for 50-yr subperiods (note that LP structure is fixed). Considering that models can do a good job at simulating the internal variability on decadal time scales, we apply the same LP truncation up to the 12th degree for both the whole twentieth century (1900–99) and 50-yr subperiods (1900–49 and 1950–99). Here we only identify a period of maximum power for LP12. The linear fall off of the spectra around the peak indicates that LP expansion is not appropriate to filter out clearly times scales outside of the peak. This is, as discusses above, related with multiple time scales contained in the LP structure, which originates from the shorter-term variability at the end of time series.

3. Data and model simulations

As the observational dataset for SATA we use the variance-adjusted version of the CRU data (HadCRUT2v) for 1900–99 (Jones and Moberg 2003). For the model dataset there are two sources. One comes from SME simulations with the ECHO-G model. These consist of an existing 1000-yr present-day control run (ECHO-G_PD; Min et al. 2005b,c) and newly performed historical simulations for 1860–2000 under three different external forcing factors (we take 1900–99 only for analysis). ECHO-G is a coupled climate model, which uses the ECHAM4 at a T30 resolution with 19 levels as the atmospheric component (Roeckner et al. 1996). The oceanic component utilizes the Hamburg Ocean

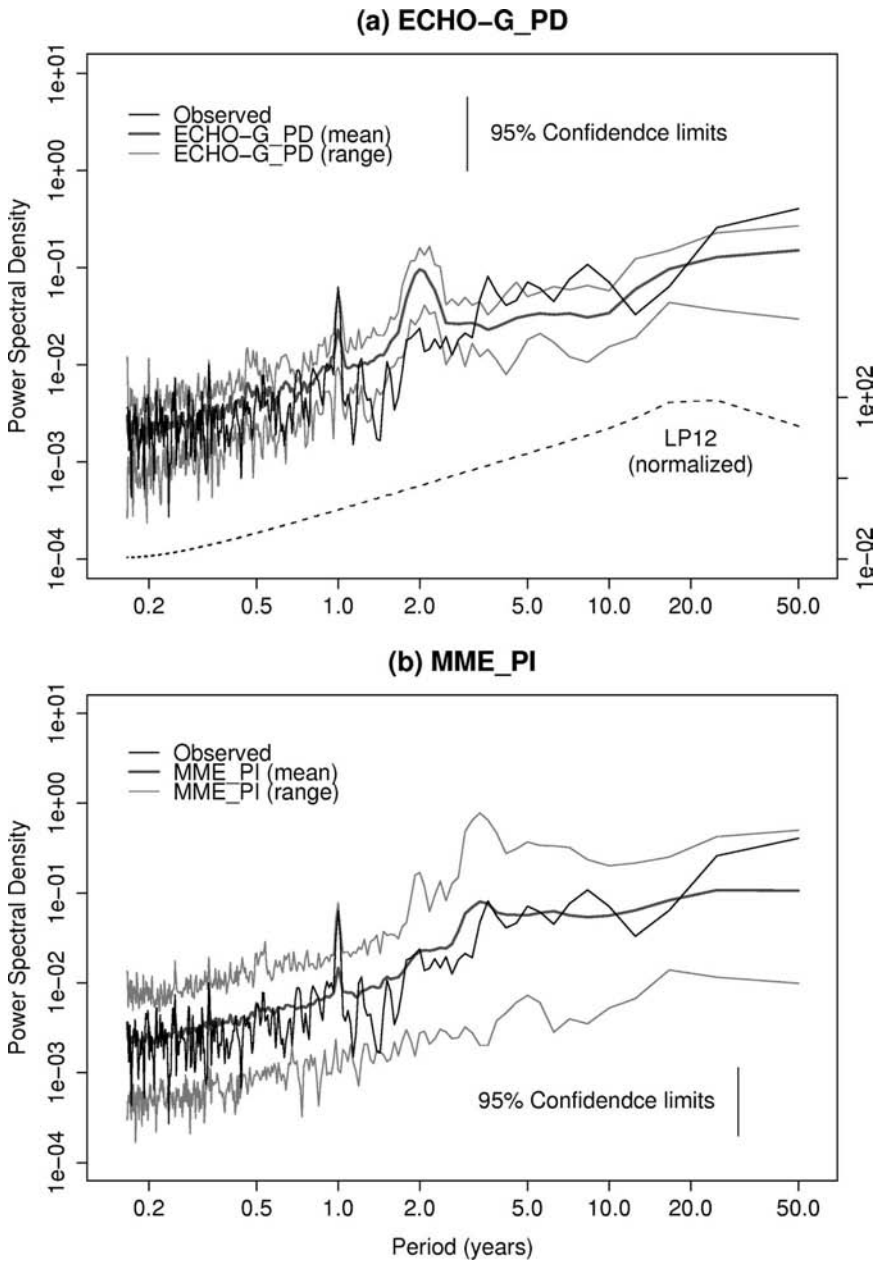

FIG. 2. Power spectra of global mean SATs (1900–99) for (a) ECHO-G_PD and (b) MME_PI compared to CRU observations (thin). Mean (thick) and ranges (gray) of model spectra are obtained from ten and eighty 100-yr nonoverlapping samples from ECHO-G_PD (ECHO-G 1000-yr present-day control run) and MME_PI (multimodel ensemble preindustrial control run; see Table 3), respectively. In (a), normalized power spectra for the 12th degrees of Legendre polynomials (dashed) are drawn together and represent maximum temporal scales near 20 yr. The vertical lines represent the range of 95% confidence in the spectral estimates for the observation and models.

Primitive Equation (HOPE-G) at an equivalent T42 resolution with a meridional refinement in the equatorial region and 20 vertical layers (Legutke and Voss 1999). The model shows a good performance at simulating climatology and natural climate variability, for example, El Niño–Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO; Min et al. 2005b,c). The equilibrium climate sensitivity (surface air temperature change in equilibrium doubling CO_2 minus control with ECHAM4 T30 coupled to slab

TABLE 3. List of AOGCMs used in constructing MME_20C and MME_PI. Ensemble members in MME_20C come from model simulations for 1900–99 including both natural and anthropogenic forcing. Number of 100-yr subsections in MME_PI is obtained by applying a 100-yr-long moving window with a shift of 10 yr while nonoverlapping samples are taken from every 100-yr interval. Atmospheric horizontal resolution of each model is given as number of grid points in longitude (Nx) and latitude (Ny).

Model	Atmospheric resolution (Nx × Ny)	MME_20C (12 models)	MME_PI (22 models)		
		Ensemble members	Total period	No. of 100-yr subsections (nonoverlapping)	Nonuse (climate drift)
BCCR_BCM2.0	192 × 96	—	250	16 (2)	
CCSM3	256 × 128	8 (run1–run7, run9)	230 (run1)	14 (2)	
			500 (run2)	41 (5)	
CGCM3.1 (t47)	96 × 48	—	500	41 (5)	
CGCM3.1(t63)	128 × 64	—	350	26 (3)	
CNRM-CM3	128 × 64	—	390	30 (3)	
CSIRO-Mk3.0	192 × 96	—	379	29 (3)	
ECHAM5/MPI-OM	192 × 96	—	670	58 (6)	
ECHO-G	96 × 48	5	341	25 (3)	
FGOALS-g1.0	128 × 60	—	150 (run1)	1 (1)	1st 50 yr
			150 (run2)	1 (1)	1st 50 yr
			150 (run3)	1 (1)	1st 50 yr
GFDL-CM2.0	144 × 90	3	500	41 (5)	
GFDL-CM2.1	144 × 90	3	500	41 (5)	
GISS-AOM	90 × 60	—	251 (run1)	16 (2)	
			251 (run2)	16 (2)	
GISS-EH	72 × 46	5	400	21 (3)	1st 100 yr
GISS-ER	72 × 46	9	500	41 (5)	
INM-CM3.0	72 × 45	1	330	24 (3)	
IPSL-CM4	96 × 72	—	230	14 (2)	
MIROC3.2(hires)	320 × 160	1	100	1 (1)	
MIROC3.2(medres)	128 × 64	3	500	41 (5)	
MRI_CGCM2.3.2	128 × 64	5	350	26 (3)	
PCM	128 × 64	4	588 (run2)	49 (5)	
UKMO-HadCM3	96 × 73	—	341	25 (3)	
UKMO-HadGEM1	192 × 145	1 (run2)	140	5 (1)	
SUM		48		644 (80)	

ocean) and the transient climate response (TCR; surface air temperature change for years 61–80 of a transient 1% yr^{-1} CO_2 increase minus control, where CO_2 doubles around year 70) of the ECHO-G model are 3.18° and 1.73°C, respectively (cf. Min et al. 2006).

Additionally, we have conducted four simulations with greenhouse gas (GHG) forcing only (ECHO-G_GHG), five with natural forcing only (ECHO-G_Nat), and four combining natural and anthropogenic forcing (ECHO-G_20C). The radiation calculation allows for the inclusion of 19 GHGs including CO_2, CH_4, N_2O, and minor industrial GHGs (Roeckner et al. 1999). In ECHO-G_Nat, solar and volcanic activities are implemented by varying the solar constant following Crowley (2000). Besides GHGs, the anthropogenic forcing runs include sulfate aerosols. The direct and first indirect effects of aerosols are calculated with an interactive sulfur cycle model (Feichter et al. 1997). ECHO-G_20C participates in the "20th Century Climate in Coupled Models" (20C3M) simulations for IPCC AR4 (Table 3). From ECHO-G_GHG, ECHO-G_Nat, and ECHO-G_20C we obtain four, five, and four nonoverlapping samples of 100-yr (1900–99) global mean SATs for GHG, Nat, and All scenarios, respectively. From the ECHO-G_PD control run, 91 time series of 100-yr global mean SATs are sampled for the CTL scenario using a moving window of 100-yr length with a shift of 10 yr. This provides an estimate of internal variability for the analysis period (see below).

The second source for model data is the IPCC AR4 archive (http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php). We extracted monthly mean SATs from 20C3M simulations (MME_20C) and preindustrial control (PIcntrl) simulations (MME_PI) from 12 and 22 models, respectively (see Table 3). Detailed model information can be also obtained from the IPCC AR4 archive. Overall 48 and 80 independent (nonoverlapping) 100-yr global mean SATs could be extracted for MME_20C and MME_PI, respectively. Another sampling for MME_PI is done with overlapping 100-yr moving win-

dows with a 10-yr shift where data from individual models are kept separate. This produces 644 samples of 100-yr global mean SATs (see Table 3 for details). This will be used to estimate parameters (means and covariance matrices) for the scenarios. We treat all model simulations as independent samples. Using different numbers of models, for example, 10, 17, and 22 models, did not change our major results, which means that means and covariance matrices, which are estimated from different combination of models, are rather similar.

For every 100-yr global mean SAT sample from ECHO-G_PD and MME_PI, the anomaly is calculated with respect to the first 20 yr to be compared with the twentieth-century simulations. Taking different reference periods for anomaly calculation (e.g., 1880–1920 or 1961–90) does not change main results. The year 1900 is selected since it is the common period between models. All model data are interpolated linearly onto the observational grid of $5° \times 5°$ and masked with observational coverage on a month-by-month basis prior to analysis.

4. Overall structure of the detection variables

a. Single-model ensembles with ECHO-G

Temporal variations of the twentieth-century global mean SATs from SMEs with the ECHO-G model under different forcing factors are shown in Fig. 3. They are low-pass filtered by retaining Legendre degrees up to 12. SATs from HadCRUT2v observations (thick black) and ECHO-G_PD (gray) are drawn together for comparison with the forced SMEs (thin black, dashed for ensemble mean). The observed warming by the end of twentieth century is about 0.8°C from the 1900–20 mean, and it is characterized by two dominant periods of increasing trend, that is, 1920–45 and 1975 onward. The scale 0.8°C is again different from 0.6°C given by Houghton et al. (2001). The former represents the temperature change near 2000 in the law data whereas the latter is an estimated value from the linear trend (see above). The amplitudes of the observed warming in the two periods are beyond the range of internal variability from ECHO-G_PD. Dissimilar to the observations, the four members of ECHO-G_GHG exhibit a steady warming over the whole twentieth century, produce a too-strong warming by the end of the century (1.2°–1.6°C), and fail in reproducing the observed early warming near the 1940s (Fig. 3a). In contrast, the ECHO-G_Nat simulations capture well the observed early warming with the intraensemble variability being large. Recent observed warming since the 1970s cannot be simulated in these runs, and the temperature changes for the period are located within the range of internal variability (Fig. 3b). The ECHO-G_All simulations (Fig. 3c) reproduce observed temperature variations successfully, indicating that both natural and anthropogenic forcing factors including the aerosol forcing are required to simulate the observed temperature changes. These results are consistent with previous studies from other models with, for example, the Third Hadley Centre Coupled Ocean–Atmosphere GCM (HadCM3; Tett et al. 2002), Geophysical Fluid Dynamics Laboratory R30 model (GFDL_R30_c; Broccoli et al. 2003), and the Parallel Climate Model (PCM; Meehl et al. 2004).

Global mean SATA time series of the ECHO-G SMEs can be decomposed into *scale*, *trend*, and decadal shorter-term components using the Legendre series expansions. The decomposition is applied to SATAs for the whole twentieth century (1900–99) and its first (1900–49) and second halves (1950–99). This allows an assessment of the different external forcing factors in different periods and a quantification of simple comparisons between observed and simulated time series described above. Figure 4 (left) shows the distributions of Legendre expansion coefficients for the three forced ECHO-G SMEs for the whole twentieth century. There is clear evidence that the observed coefficients a_0 and a_1 (amplitude for *scale* and *trend*) have positive values lying outside of the uncertainty range of ECHO-G_PD. Observed coefficients for higher degrees are within the range of internal variability (also see Fig. 1). It is found that observed coefficients a_3 and a_4 are significantly different from the mean of ECHO-G_PD at the 5% level assuming normal distributions, while a_2 is not. In terms of conventional univariate statistics, this means that external forcing signals in global mean SATs can be detected only on time scales longer than about 50 yr, which is consistent with a previous study by Stott and Tett (1998).

The ECHO-G_GHG results for 1900–99 show positive coefficients beyond ECHO-G_PD ranges for degrees from zero to two. As seen above, the amplitude for trend is much larger than the observed. In the ECHO-G_Nat simulation, the coefficients for a_0 and a_1 are positive but smaller than the observed one while the coefficients a_4 are close to the observations. Because the observed warming trend for 1920–45 projects strongly on LP4 (Fig. 1), this result indicates that the natural forcing might explain parts of the observed early warming. The ECHO-G_All results exhibit good consistencies with observations for Legendre degrees for 0–4 confirming that observed long-term variations of the twentieth-century global mean SATs may be attributable to both natural and anthropogenic forcing.

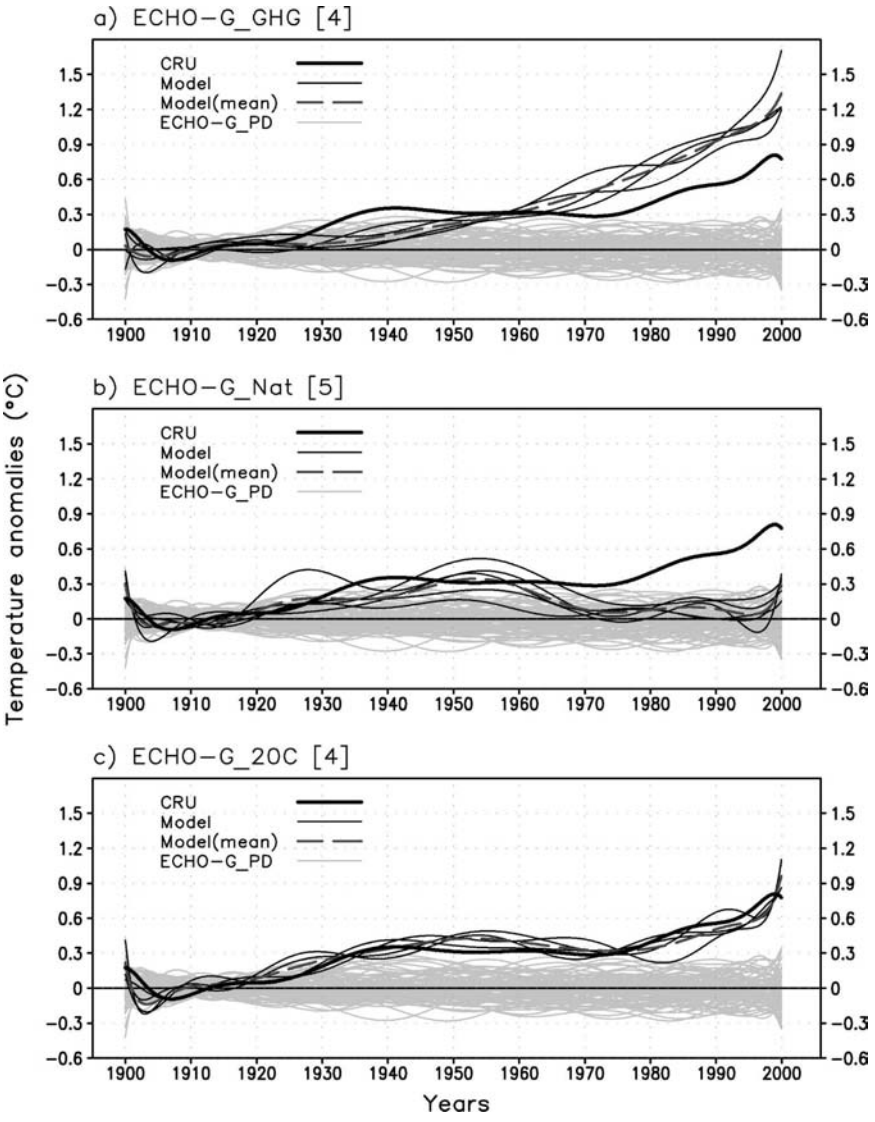

FIG. 3. Time series of reconstructed global mean SATs from the ECHO-G simulations under different forcing factors (thin lines): (a) ECHO-G_GHG, (b) ECHO-G_Nat, and (c) ECHO-G_20C. Number in brackets indicates number of ensemble members. Gray lines show a range of internal variability from ECHO-G_PD (91 samples from 100-yr moving windows with a 10-yr shift), dashed lines are ensemble mean, and thick lines are CRU observations. All values are low-pass-filtered temperature anomalies from 1900–20 means using Legendre series expansions retained at the 12th degree.

The Legendre coefficients for the first half of the twentieth century (1900–49) are presented in Fig. 4 (middle). Observations show a general warming of 0.15°C (a_0) and an increasing trend (a_1), both of which are beyond the internal variability range of the ECHO-G_PD runs. Observed values for higher degrees from LP2 are within the internal variabilities similar to the results for the full twentieth century. The ECHO-G_GHG runs show a similar behavior to the observations, but with weak amplitudes. The ECHO-G_Nat simulations exhibit a good skill only in two members out of five indicating the large intraensemble difference in responses to natural forcing as discussed above. The ECHO-G_All ensemble captures the observed warming for 1900–49 very well in the coefficients a_0 and a_1 with a small intraensemble difference or a good consistency between ensemble members.

Similar results can be found in the Legendre expansion coefficients for the second half (1950–99), where the observed mean warming from 1900–20 mean (a_0) is 0.42°C. While the ECHO-G_GHG simulations display a too-large warming and the ECHO-G_Nat runs do not

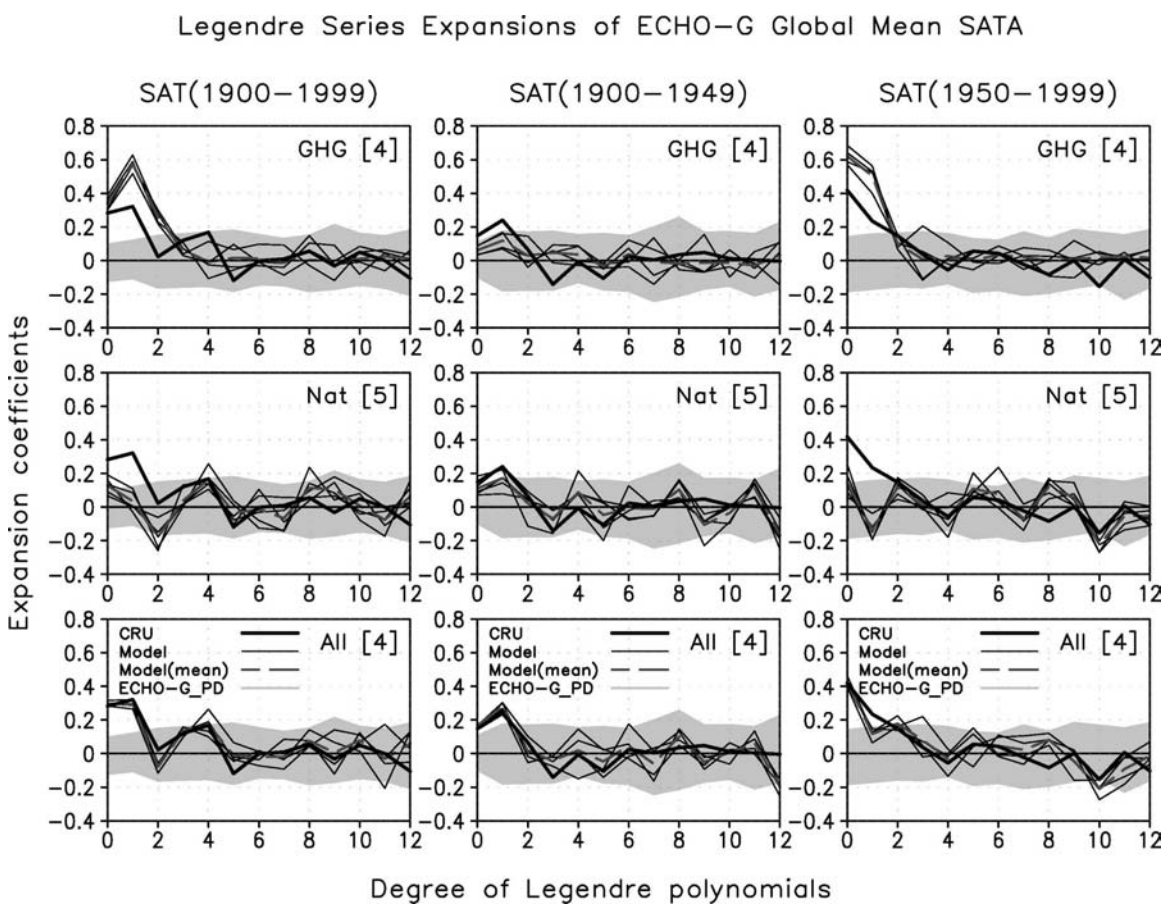

FIG. 4. Legendre expansion coefficients for global mean SATs for the period of (left) 1900–99, (middle) 1900–49, and (right) 1950–99 from ECHO-G_GHG, ECHO-G_Nat, ECHO-G_20C simulations (thin), ensemble means (dashed), and CRU observations (thick lines; see Fig. 3 for low-pass-filtered original time series). Number in brackets shows number of ensemble members. Gray shading represents the range (maximum to minimum) from ECHO-G_PD.

have significant warming, the ECHO-G_All simulations show similar amplitude of warming although the trend (a_1) is slightly smaller than that observed. According to Roeckner at al. (1999), who used the same atmospheric model with the same historical sulfate emissions as here, the direct and first indirect aerosol forcing for 1860–1990 is estimated as -0.35 and -0.91 W m^{-2}, respectively. Subtracting Legendre coefficients of ECHO-G_GHG and ECHO-G_Nat from that of ECHO-G_All provides an estimate of aerosol cooling effect ($a_0 = -0.35$ and $a_1 = -0.26$ for 1950–99). They are very similar to those estimated from three-member ensemble runs with sulfate aerosol forcing only ($a_0 = -0.32$ and $a_1 = -0.24$), indicating linearity of the response to external forcing in the ECHO-G model (e.g., Gillett et al. 2004).

To summarize, these results show that 1) Legendre series expansion provides a useful information for climate change signal analysis, and 2) only the ECHO-G simulations with all (natural and anthropogenic together) forcing can reproduce observed global mean temporal SAT patterns not only for the whole twentieth century but also the first and second halves of the century separately.

b. *Multimodel ensembles from the IPCC AR4 models*

To consider uncertainties from intermodel differences, detection variables are obtained from the multimodel dataset. Figure 5 shows reconstructed global mean SATs for the twentieth century using LP0 to LP12 from MME_20C (48 members) and MME_PI (644 members). The MME_20C dataset is based on 12 models that are available for the natural plus anthropogenic forcing run only while the MME_PI set is based on 22 models (Table 3). All results are relative to 1900–20 means. Some members in MME_20C do a

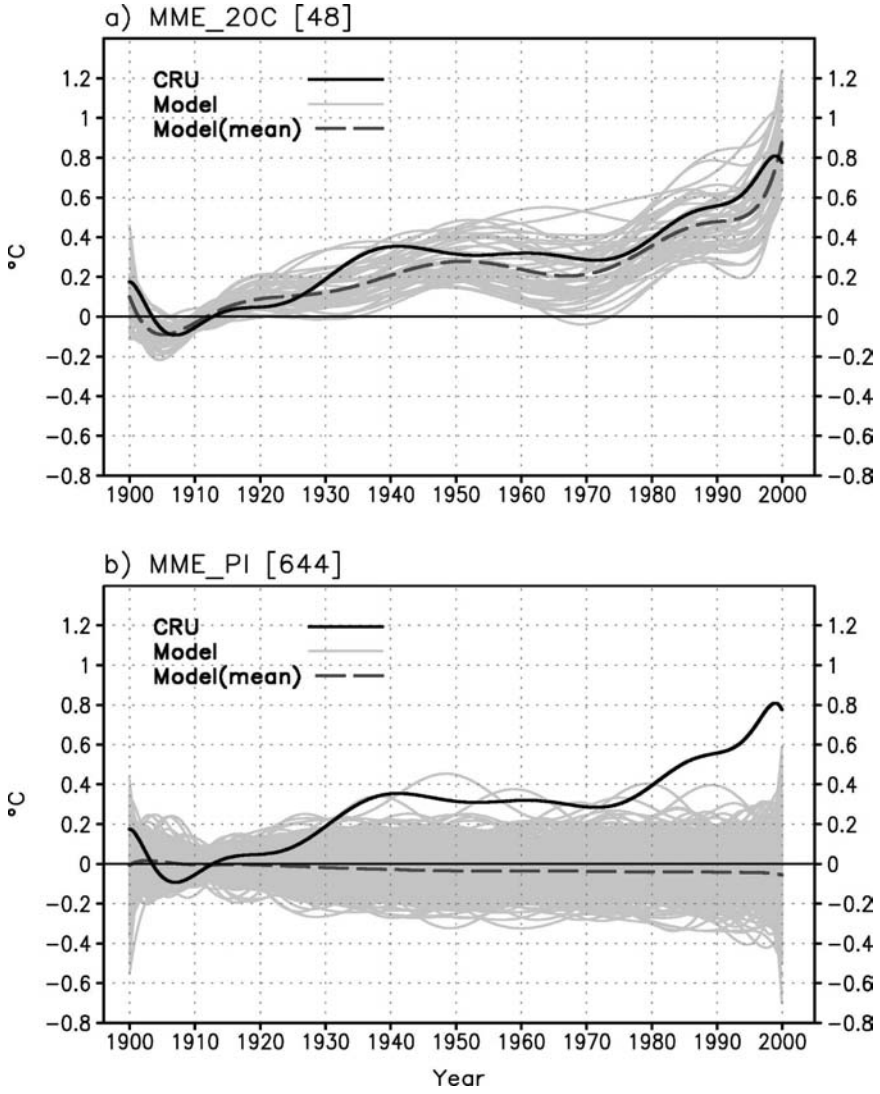

FIG. 5. Reconstructed global mean SATs from (a) MME_20C (12 models 48 members) and (b) MME_PI (22 models 644 members; see Table 3) simulations (gray lines). Dashed lines are ensemble means, and thick lines are reconstructed SATs from CRU observations. MME_PI ensemble members are constructed from 100-yr moving windows with a 10-yr shift. All data are anomalies from 1900–20 means and are low-pass filtered using Legendre series expansions retained at the 12th degree.

good job in simulating observed temperature changes although intermodel difference is a bit large. Ensemble mean of MME_20C shows a slightly lower warming than observations and the difference is largest near 1940. MME_PI also exhibits large uncertainty ranges, but the observed recent warming is still outside of the range.

It would be worth discussing the spread in the simulated SATs, which widens after 1920 in association with possible effects of climate drift and the choice of centering on the properties of LP coefficients and Bayesian decision results. First, concerning climate drift, we excluded a few model simulations, which show noticeable drift, constituting MME_PI (Table 3). In the case of MME_20C, the selection of 12 models, which were integrated under natural and anthropogenic forcing together, left out a few models of larger climate drift (not shown). Even after this procedure, the climate drift would remain (see ensemble mean line in Fig. 2b). The remaining drift would serve as an upper bound by enlarging the internal variability estimated. That is, as one removes climate drift more, the estimated internal vari-

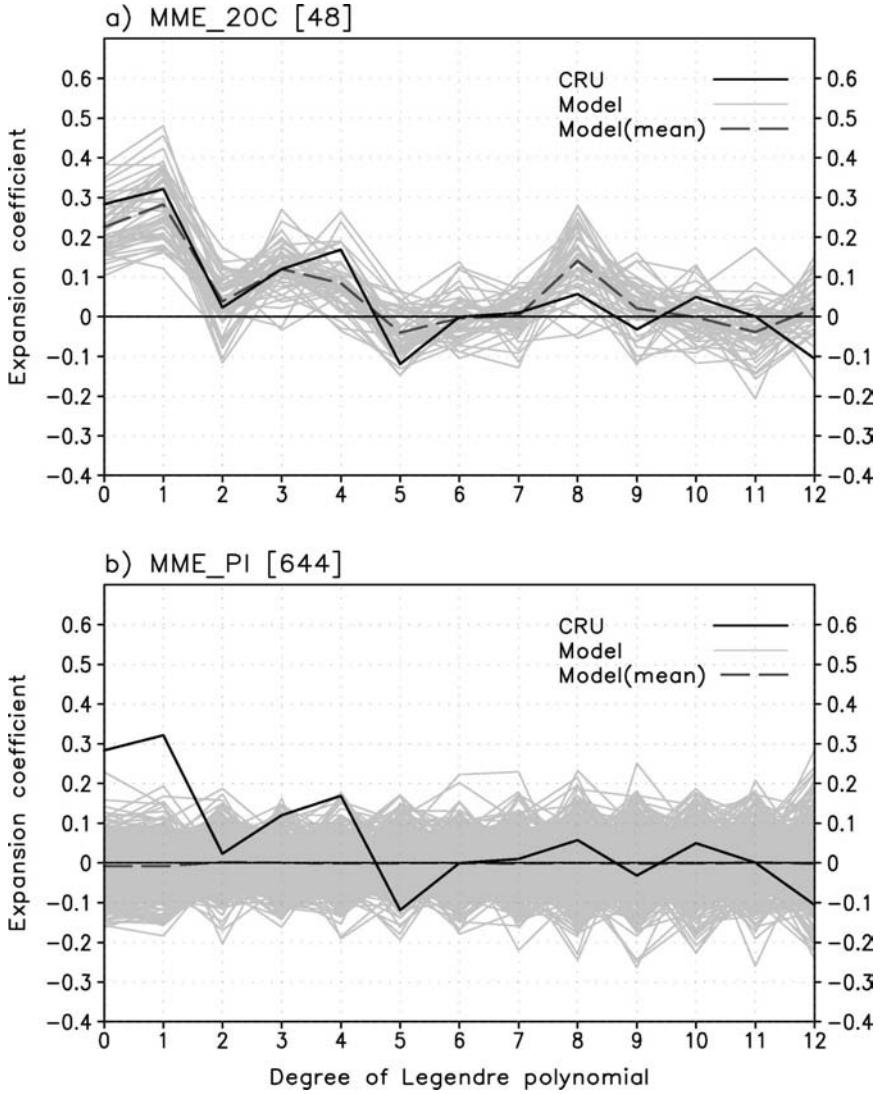

FIG. 6. Legendre expansion coefficients for global mean SATs for the period of 1900–99 from (a) MME_20C (12 models 48 members) and (b) MME_PI (22 models 644 members). Thick black lines are CRU observations, and dashed lines are model ensemble means. See Fig. 5 to compare with low-pass-filtered original time series.

ability will be reduced. Second, the choice of centering affects only the zeroth degree of LP (scale) without giving any changes in the first and higher degrees of LPs. It means that applying different reference periods cannot change any shorter-term component except for the time average. In some comparisons of Bayesian decisions from taking different centering, we found negligible effects (not shown).

Legendre coefficients for the two MMEs are shown in Fig. 6. It clearly shows that observed SAT changes summarized in the LP0 and LP1 coefficients are inside of the uncertainty range of MME_20C and outside of that of MME_PI. MME_20C has coefficients in LP0 and LP4 that are a bit smaller than observations, which can be inferred from SAT patterns in Fig. 5a described above.

5. Bayesian decisions

a. Experiment setup

The Bayesian decision analysis quantifies the comparisons or similarity measures described in the above section with respect to various uncertainty estimates from either SMEs or MMEs. The combinations for seven different Bayesian decision experiments are summarized in Table 4. The difference of the experiments

TABLE 4. Experiment list for the Bayesian decision analysis. Number in brackets represents number of ensemble members with the same abbreviation of model simulations as in Table 1. SINGLE represents an experiment using SMEs while six MULTIs represent those using MMEs (bold). MME_PI [80] are based on independent (nonoverlapping) sampling while MME_PI [644] are sampled from 100-yr moving windows with a 10-yr shift. See text for details.

Experiment name	Model simulations for parameter estimations				
	Mean and variability for CTL	Mean for GHG	Mean for Nat	Mean for All	Variability for GHG, Nat, and All
SINGLE	ECHO-G_PD [91]	ECHO-G_GHG [4]	ECHO-G_Nat [5]	ECHO-G_20C [4]	ECHO-G_PD [91]
MULTI1	ECHO-G_PD [91]	ECHO-G_GHG [4]	ECHO-G_Nat [5]	**MME_20C [48]**	ECHO-G_PD [91]
MULTI2	ECHO-G_PD [91]	ECHO-G_GHG [4]	ECHO-G_Nat [5]	**MME_20C [48]**	**MME_20C [48]**
MULTI3	**MME_PI [80]**	ECHO-G_GHG [4]	ECHO-G_Nat [5]	**MME_20C [48]**	**MME_PI [80]**
MULTI4	**MME_PI [80]**	ECHO-G_GHG [4]	ECHO-G_Nat [5]	**MME_20C [48]**	**MME_20C [48]**
MULTI5	**MME_PI [644]**	ECHO-G_GHG [4]	ECHO-G_Nat [5]	**MME_20C [48]**	**MME_PI [644]**
MULTI6	**MME_PI [644]**	ECHO-G_GHG [4]	ECHO-G_Nat [5]	**MME_20C [48]**	**MME_20C [48]**

comes from various model combinations to define the four scenarios: CTL, GHG, Nat, and All. Practically the entries in Table 4 define those SMEs/MMEs from which the mean and covariance matrices are estimated. The mean estimation for the GHG and Nat scenarios is fixed for all the experiments, because at present we have only ECHO-G_GHG and ECHO-G_Nat with four and five samples at our disposal. The method can be extended to a multimodel approach if a large number of different model simulations becomes available with relevant forcing. The SINGLE experiment in Table 4 uses only data from SMEs (Legendre coefficients in Fig. 4) while MULTI1 to MULTI6 use different combination of multimodel data from MME_20C and MME_PI (Legendre coefficients in Fig. 6).

Two possible effects of MMEs are on the mean and on the variability (covariance matrix). In the MULTI1 we assess the effect of MME_20C on the mean only. MULTI2 has an additional effect of MME_20C on the variability. MULTI3–6 commonly use MME_PI rather than ECHO-G_PD for the mean and variability estimation for the CTL scenario. The differences between MULTI3–6 are based on the number of samples as well as the assumption about the forced variability. While MULTI3–4 take nonoverlapping 80 samples to estimate the mean and variability of MME_PI, MULTI5–6 use overlapping 644 samples (Table 3). Hence the comparison of results from MULTI3–4 with those from MULTI5–6 provides a sensitivity test to the number of samples in defining scenarios. On the other hand, MULTI3 and MULTI5 are different from MULTI4 and MULTI6 in treating the variability for the forced scenario (GHG, Nat, and All scenarios). The former experiments adopt MME_PI while the latter experiments take MME_20C as a best guess for the forced variability. This corresponds to assuming that the unforced internal variability remains unchanged given external forcing (MULTI3 and MULTI5) or it changes notably due to the external forcing (MULTI4 and MULTI6). By comparing results from different experiments, the effect of intermodel uncertainties on the Bayesian detection and attribution is assessed below.

b. Results for 1D and 2D variables: Scale and trend

Bayesian decision processes for one-dimensional (1D) and two-dimensional (2D) detection variables are explained before extending Bayesian analysis to a higher-dimensional case, assuming that the dominant signals of the external forcing enforce low-frequency responses as discussed above. The coefficients for LP0 and LP1 are selected either as two sets of 1D variables ($q = 1$) or as one 2D variable ($q = 2$). Legendre coefficients retained at kth degree ($k \geq 2$) provide higher-dimensional detection variables, which will be analyzed later with $q = k + 1$. Figures 7 and 8 illustrate the Bayesian decision processes for either 1D or 2D variables using SATs for 1900–99. As contours or lines, we include the likelihood probability density functions (PDFs). The decision areas for the four scenarios are indicated by the shadings, and the observational points are marked with crosses. All values are derived from the SINGLE (Fig. 7) and MULTI5 (Fig. 8) experiments. Model simulations described in Table 4 provide samples to estimate the parameters for the likelihood PDF of each scenario in Eq. (2). Means and covariance matrices are calculated from all samples assuming independency between the samples. The likelihood PDF for each scenario is calculated by applying hypothetical observational values (Legendre coefficients) in the range from –1.0 to 1.0 at steps of 0.01. The shaded regions are derived from selecting the most probable scenario given the observation at each point using the decision rule of Eq. (3). For simplicity it is assumed that prior probabilities of all four scenarios are identical, $P(m_1) = P(m_2) = P(m_3) = P(m_4)$. In this case, the Bayes factor or likelihood ratio can act as a decision

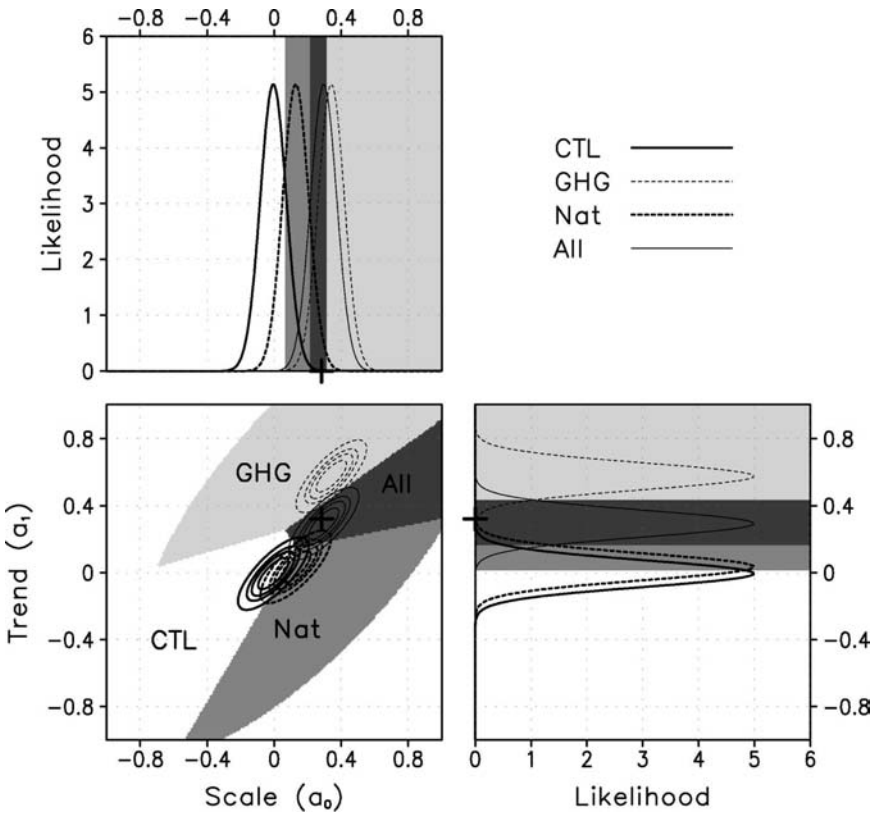

FIG. 7. Distribution of likelihood PDFs (contour lines) for four scenarios concerned in Table 1, locations of observations (plus symbol), and Bayesian decision areas (shadings) for three detection variables: (top) *scale* (using a_0 only), (bottom right) *trend* (using a_1 only), and (bottom left) *sale* and *trend* simultaneously (using a_0 and a_1 together) from the SINGLE experiment described in Table 4. Gray shadings represent decision area for each scenario into which hypothetical values for observations are classified (white for the CTL). Bayesian decisions are made assuming identical priors for the scenarios.

function since it becomes identical to posterior odds [see Eq. (4)]. The shaded regions indicate therefore the scenario of maximal Bayes factor. A generalized Bayesian decision is made below (in Figs. 10 and 11) where we introduce varying prior probability of each scenario. The plots in Figs. 7 and 8 are organized such that the 1D result can be viewed as the marginal result of the 2D case.

In the 1D results of the SINGLE experiment (upper- and bottom-right panels of Fig. 7), the observation in *scale* (a_0) is positioned near the center of likelihood PDFs of the All and GHG scenarios, which are very close to each other, while the observation in *trend* (a_1) is located near the center of the All scenario, which is well separated from the GHG (too big a trend) and the Nat/CTL scenarios (trend close to zero). It is a quantification of the visual presentation of the coefficients a_0 and a_1 in Fig. 4. The 2D results using both *scale* and *trend* together show an even clearer separation between the scenarios than the 1D results (Fig. 7, bottom left). The observations are well positioned near the center of the All scenario area. This is a good example showing an advantage of multidimensional or multipattern approach to get clearer signals. The same plots from the MULTI5 experiment (Fig. 8) represent similar results to the SINGLE ones, except for two notable differences. First, the center (mean) of the All scenario is moved toward zero especially in *scale* (a_0), which is caused by a smaller overall warming simulated in multimodel averages (cf. Fig. 5a with Fig. 3c). The second effect of MMEs is that, due to strengthened correlation structure between a_0 and a_1, the decision area for the All scenario in the 2D result of MULTI5 is reduced compared to that in SINGLE. This makes the decision in favor of the All scenario more selective.

The overall comparison demonstrates that, for the 1D and 2D study of the variables *scale* and *trend*, the introduction of the larger intermodel uncertainties does

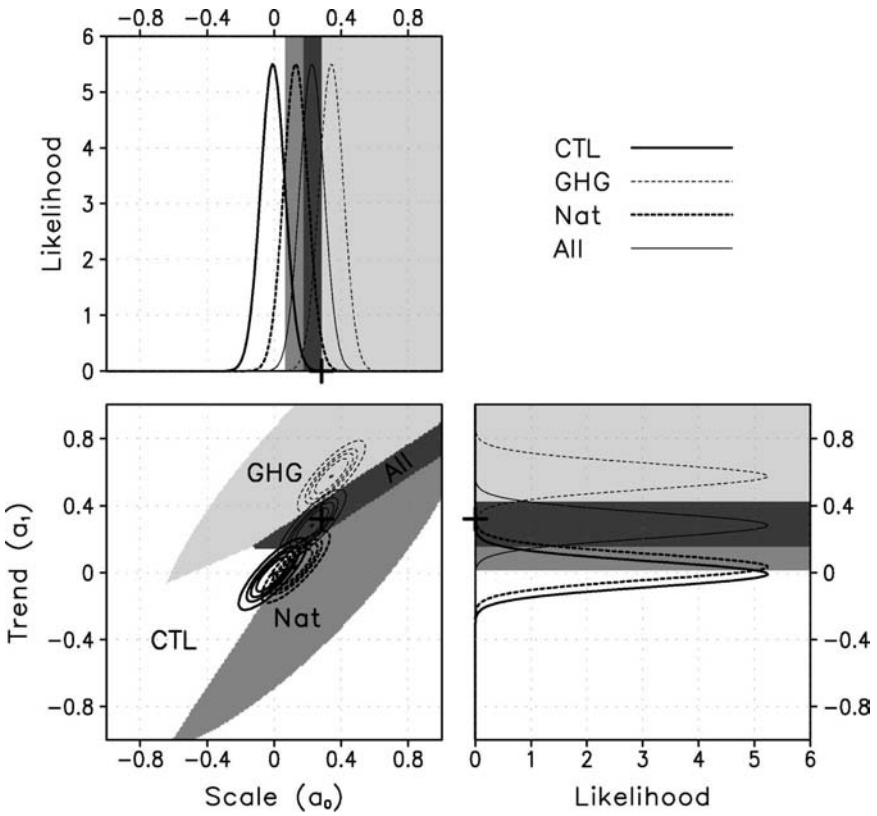

FIG. 8. Same as in Fig. 7, except for the MULTI5 experiment.

not change the main results of Bayesian decisions. Observed global mean SAT changes in the twentieth century are classified into the All scenario. That is, the observations can only be attributed to the all-forcing case. The results for 1900–49 and 1950–99 show the same conclusions (figures not shown), which is also consistent with the visual comparisons of the time series described in section 4. Effects of using MMEs appear as movement of likelihood PDFs of the All scenario toward zero for *scale* (a_0) and *trend* (a_1) of 1900–49 and a better consistency between the All scenario and observations for *trend* (a_1) of 1950–99, which are associated with relative skills of MME_20C to ECHO-G_20C: worse at simulating mean changes and trends for 1900–49 while better at simulating trends for 1950–99.

Fuzzy ranges of the decision boundary can be considered on the basis of the descriptive scales for the likelihood ratio, for example, logarithm ratio less than 1, of the two scenarios existing across the boundary similar to those for the Bayes factors given in Table 2. Sensitivity tests of the fluctuations of decision boundaries can also be done to prior probabilities (cf. Fig. 4a of M04) and other intermodel uncertainties (not only MULTI5 in Fig. 8). The latter test includes the effects of means, covariance matrices, and number of model simulations applied (Table 4). The fuzziness of the boundary is difficult to present in a single figure.

c. Result for higher-dimensional variables

Next we extend the 1D and 2D Bayesian decisions above to higher dimensions by increasing the degree of retained Legendre polynomials from 0 to 12 ($q = 1$ to 13). Inclusion of higher degrees means shorter time scales to be considered additionally in the Bayesian decision process (Fig. 1). Figure 9 shows the Bayesian decision results for higher-dimensional representations of SATs for the three analysis periods (1900–99, 1900–49, and 1950–99) from SINGLE and MULTI experiments. The logarithm of the Bayes factor is used as a decision function. For clarity, identical priors of the scenarios are assumed again. If we use the criteria for Bayes factors from Table 2 for the values in Fig. 9, we find that decisive evidences (logarithm of Bayes factors > 5) can be only seen for the All scenario for the periods of 1900–99 and 1950–99. This result is robust to both intermodel uncertainties and temporal scales concerned. For SAT changes during the earlier part 1900–49, Nat and All scenarios have strong evidences (loga-

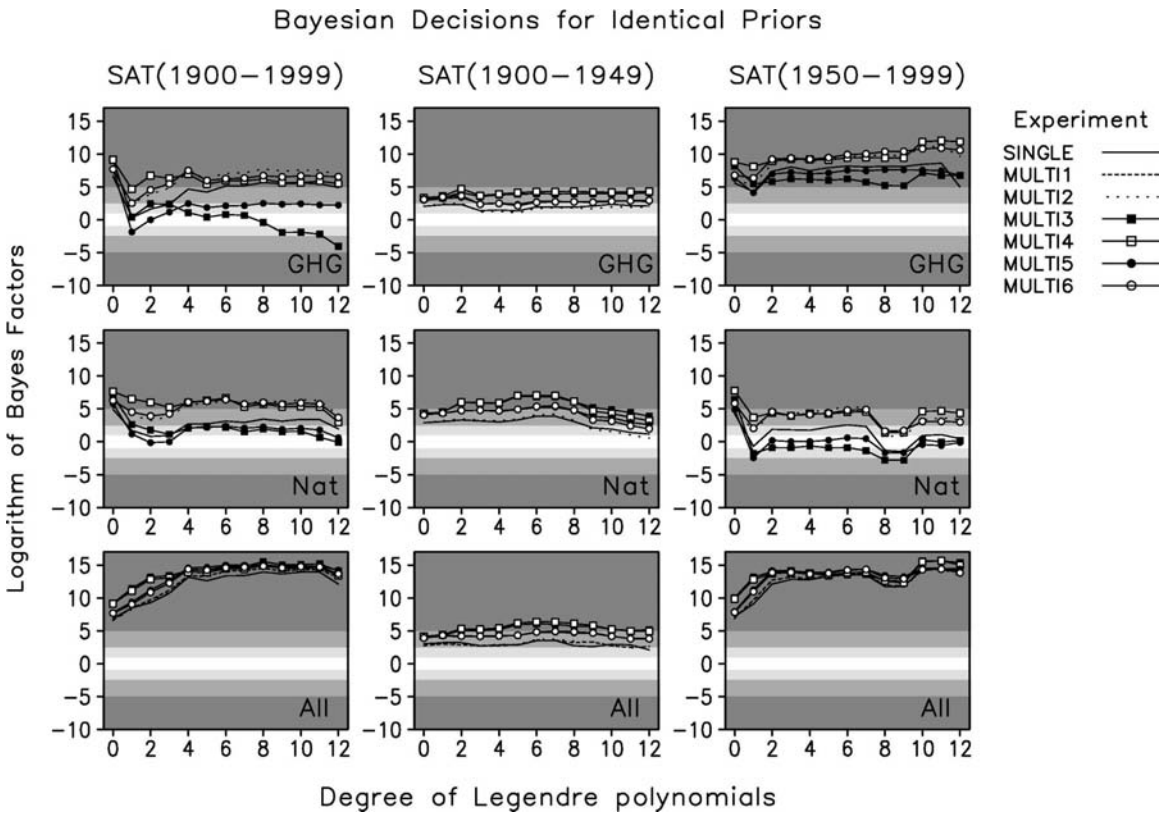

FIG. 9. Bayesian decision results for observed global mean SATs for the period of (left) 1900–99, (center) 1900–49, and (right) 1950–99 as represented by distributions of logarithm of Bayes factors for the GHG, Nat, and All scenarios with respect to the CTL scenario as retained degree of Legendre polynomials varies from 0 to 12 (as shorter time scales are added, see Fig. 1, e.g.) for SINGLE and MULTI experiments described in Table 4. Assuming identical priors for the scenarios, logarithms larger than 5 (less than –5) represent decisive evidence for the scenario concerned (for the reference scenario CTL), which are marked with dark shadings. Light and middle-dark shadings represent substantial and strong areas similarly (see Table 2 for descriptive scales).

rithm > 2.5) for Legendre degrees 0–8. It is reasonable to see that Nat and All results are very close to each other for 1900–49 when anthropogenic forcing is weaker than natural forcing.

Figure 9 also reveals large fluctuations of the Bayes factors between experiments especially in the GHG and Nat results for the two periods of increasing SATs (1900–99 and 1950–99). Remembering that the mean of the CTL scenario is always close to zero in all experiments and the means of GHG and Nat are not changed between models (Table 4), this indicates an important role of natural internal variability in the Bayesian decisions. Smaller values in MULTI3 and MULTI5 are caused by smaller variabilities estimated from MME_PI than from MME_20C for the MULTI2, MULTI4, and MULTI6 (not shown). This is because the intermodel uncertainty in MME_20C includes intermodel differences in responses to the external forcing as well as natural variability, while the variability in MME_PI is composed mostly of the latter.

d. Sensitivity to varying priors

As a last step the Bayesian decisions can be generalized by considering varying prior probabilities. Here the priors represent subjective beliefs in the given scenarios. It can be regarded as a weighting factor to the likelihood PDF of each scenario. As a simple way we take "noninformative" uniform priors (M04; Schnur and Hasselmann 2005; Lee et al. 2005). Another assumption is that total prior sum of the four scenarios (CTL, GHG, Nat, and All) equals one, and priors of the climate change scenarios excluding CTL are identical, which leads to $P(m_{2,3,4}) = [1 - P(m_1)]/3$. Next prior probability of CTL $P(m_1)$ is changed from 0.01 to 0.99. This corresponds to changing prior odds $P(m_1)/[1 - P(m_1)]$ from 0.01 to 99, which can be interpreted as giving a weighting to $P(m_{2,3,4})$ by 33 times of $P(m_1)$ to a weighting to $P(m_1)$ by about 337 times of $P(m_{2,3,4})$. Using the varying priors and the likelihood values (or Bayes factors in Fig. 9), posterior probabilities for each

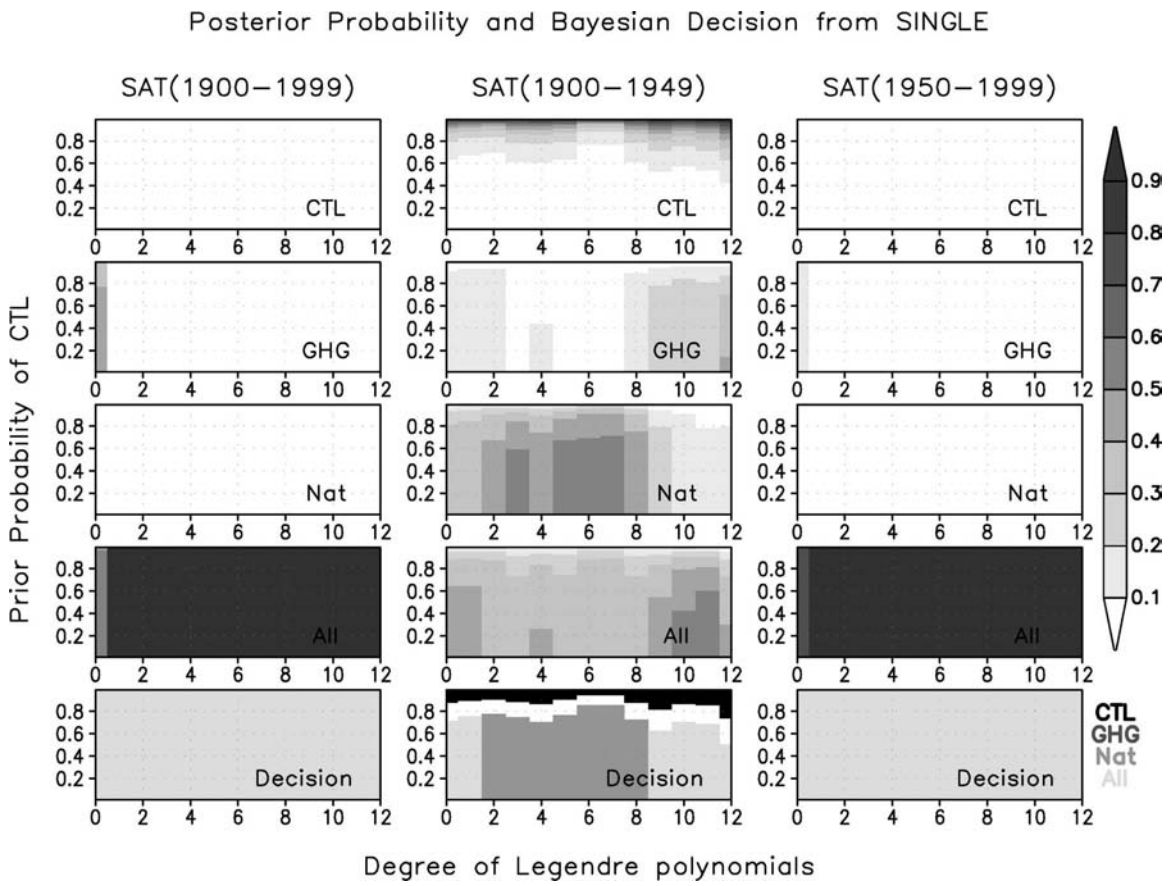

FIG. 10. Distributions of posterior probabilities for the four scenarios of (top four rows) CTL, GHG, Nat, and All and (bottom row) corresponding Bayesian decisions of the most probable scenario for observed global mean SATAs for three periods [(left) 1900–99, (middle) 1900–49, and (right) 1950–99] in case of varying prior probabilities from the SINGLE experiment. Ordinate depicts the prior probability of CTL P(m_1) and abscissa represents retained Legendre degrees from 0 to 12. It is assumed that priors of the climate change scenarios except for CTL are identical with a constraint that a total summed prior is unity. White-colored area in the decision plot indicates decision area where logarithm posterior odds of the decided scenario with respect to CTL are less than 1, i.e., climate change signal is "not worth more than a bare mention" (see Table 2).

scenario can be obtained from Eq. (1). The posterior probability of each scenario represents its contribution to total summed posteriors, which always equals one as the prior does (cf. see Fig. 2 of M04). Bayesian decisions are made by selecting the scenario of maximum posterior at each point of varying priors and Legendre degrees retained.

Figure 10 shows these posterior probabilities of the four scenarios and the Bayesian decisions for global mean SATA from the SINGLE experiment as a function of the varying priors. The periods of 1900–99, 1900–49, and 1950–99 are considered as before. For the whole twentieth century and its second half, All has the largest posterior probabilities for all ranges of Legendre expansions and prior values even if P(m_1) is 0.99. The result for the first half shows a different behavior. Posterior of All is largest for LP0, LP1, and LP9–LP12 while Nat signals are dominant from LP2 to LP8. Also the decision is in favor of CTL if P(m_1) is larger than 0.8 or prior odds are larger than 4 (see above). This result from the single-model ECHO-G demonstrates that Bayesian decision results for global mean SATs are largely insensitive to prior changes if the whole twentieth century or the second half is considered.

To see the impact of intermodel uncertainties on Bayesian decision results in the case of varying priors, posteriors and Bayesian decisions from the MULTI5 experiment are shown in Fig. 11. Other MULTI experiments exhibit very similar results (not shown). One can clearly find that the scenarios selected by the decision in the MULTI5 experiment are very close to those in the SINGLE experiment. Only one minor difference from the SINGLE result can be seen in that Nat is dominant in LP0 and LP1 for 1900–49. This MME effect on long-term components (*scale* and *trend*) is asso-

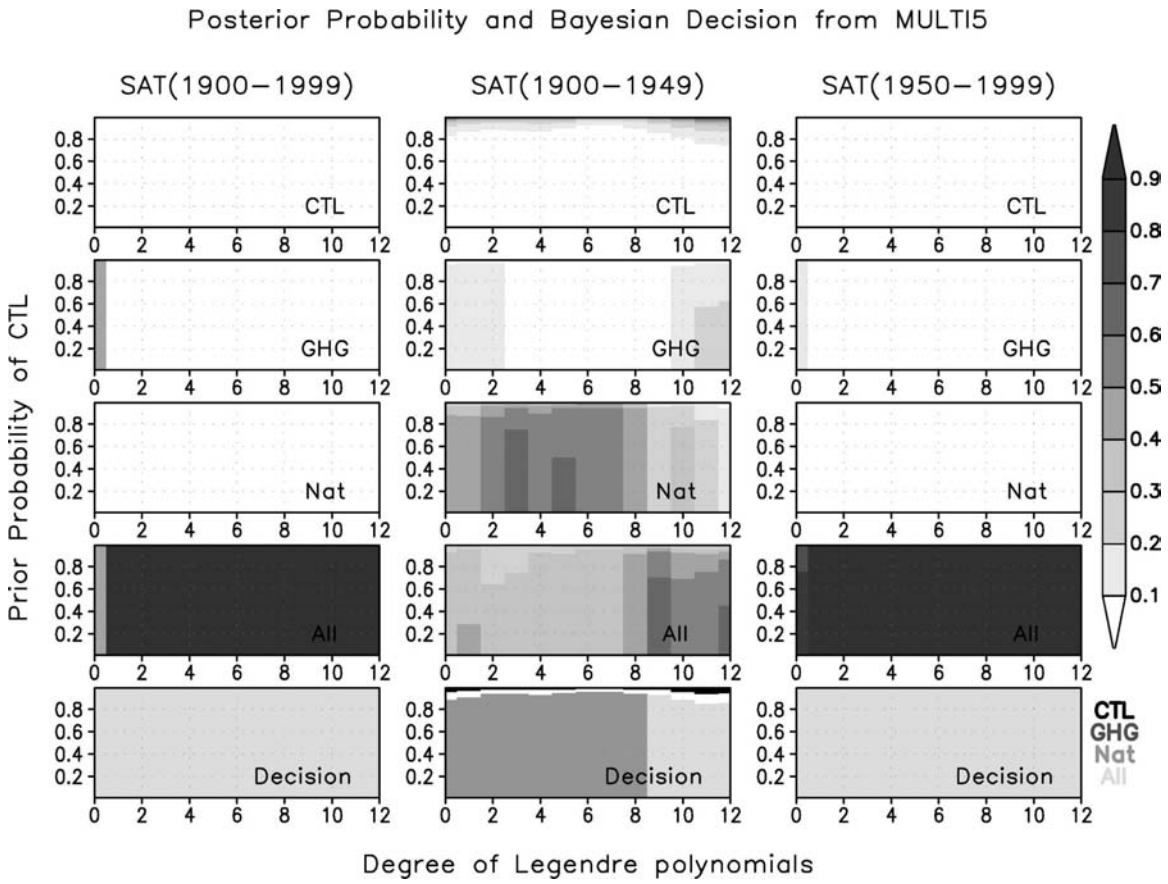

FIG. 11. Same as in Fig. 10, but for the MULTI5 experiment.

ciated with movement of means as discussed above in the identical prior case. A good consistency between the SINGLE and MULTI experiments demonstrates that our Bayesian decision results for the observed twentieth-century global mean SATs are robust to the uncertainties arising from intermodel differences as well as prior probabilities.

6. Conclusions and discussion

A Bayesian decision method developed by M04 is applied to observed and simulated twentieth-century global mean SATs considering four scenarios ("CTL," "GHG," "Nat," and "All") whose parameters are estimated from single-model ECHO-G or multimodel IPCC AR4 simulations. To consider temporal scales in the analysis, the time series of global mean SATs are expanded into 13 components of Legendre series expansions. The coefficients represent an overall warming scale (zeroth degree), the linear trend (first degree), and shorter-term decadal (second degree and higher) variations. The Legendre coefficients serve as detection variables for the Bayesian decision process. Two subperiods of 1900–49 and 1950–99 as well as the whole twentieth century (1900–99) are considered to examine relative importance of different forcing factors in explaining observed climate change in different subperiods.

Simple comparison of Legendre coefficients shows good consistencies between observations and All forcing simulations, which does provide some hints on the results of climate change detection and attribution. Our Bayesian decision quantifies this comparison with consideration of the changing prior information and uncertainties from internal variability and intermodel difference as well. As in other Bayesian studies (Min et al. 2005a; Schnur and Hasselmann 2005; Lee et al. 2005), the Bayes factor or likelihood ratio is used as an indicator of observed evidence for the scenario concerned, given a predefined reference scenario (CTL in this study). Numerical values of the Bayes factor are transferred into the descriptive scales "substantial," "strong," and "decisive" through thresholds provided by Kass and Raftery (1995).

Analysis results using single-model simulations (SINGLE experiment) show that global mean SATs for the whole twentieth century and its second half exhibit decisive evidences only for the All scenario for all ranges of Legendre degrees retained. On the other hand temperature changes in the first half of the century are classified into both Nat and All scenarios with strong evidences. These results are very consistent with previous works based on conventional or Bayesian statistics (Mitchell et al. 2001; IDAG05, and references therein).

When multimodel simulations are applied (MULTI experiments), the distributions of the Bayes factors are not changed notably, implying the insensitivity of Bayesian decisions to intermodel uncertainties arising from different model sources. It is also demonstrated that Bayesian decision results are largely robust to varying priors. This is well shared with previous results for a regional-scale detection by Min et al. (2005a) and global results by M04, Schnur and Hasselmann (2005), and Lee et al. (2005).

In this study, we handle multiple signals (GHG, Nat, and All) in the presence of the unforced background noise (CTL). The unforced noise can contain the intermodel differences if we use MMEs rather than SMEs. Suppose that one wants to detect anthropogenic signal only (e.g., "Anthro" scenario defined as GHG plus sulfate aerosol forced change) from a single background noise. This becomes a single signal versus a single noise problem, and the Bayesian decision procedure and result would be very similar to those in the 2D case of M04 and Min et al. (2005a) although details might be highly different because of the different data domain and model simulations used. As the noise model, we can think about CTL alone or CTL-plus-Nat scenarios and assess how this would change the Bayesian decision result. However, in practice, since there are not enough model simulations available with natural forcing only, parameters characterizing the noise model should be estimated mainly from CTL, which would reduce a possible change in the decision result.

One important result is that Bayesian decision results are highly dependent upon a suite of forcing scenarios defined, not the choice of models. This implies that one should construct proper scenarios with care by which observations might be explained better. It should be also noted that forcing uncertainty is partly included in the intermodel uncertainties analyzed here since models were not integrated with exactly identical forcing for a given scenario (e.g., solar and volcanic forcing). Although, for simplicity, we assume that forcing uncertainty is very small because generally the uncertainty in the response from different models to the same forcing is much larger than ranges of forcing uncertainty; separation of the forcing uncertainty out of the intermodel differences and its effect on the Bayesian climate change assessment would be an interesting subject for future work.

Acknowledgments. We thank four anonymous reviewers for their helpful comments. We are also grateful to Gabi Hegerl for fruitful discussion; Stephanie Legutke, Jean-Francois Royer, Ulrich Schlese, and Eduardo Zorita for their help in our ECHO-G simulations; and Won-Tae Kwon and Hyo-Shin Lee for providing their ECHO-G dataset. This work was supported by the German Research Foundation (DFG) with Grant He1916/8. ECHO-G model simulations were performed at NEC supercomputers in DKRZ, Germany, and in KMA and KISTI, South Korea. We acknowledge the international modeling groups for providing their data for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the JSC/CLIVAR Working Group on Coupled Modelling (WGCM) and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panel for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The IPCC Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, U.S. Department of Energy.

REFERENCES

Barnett, T. P., and Coauthors, 1999: Detection and attribution of recent climate change: A status report. *Bull. Amer. Meteor. Soc.,* **80,** 2631–2659.

Berger, J. O., 1985: *Statistical Decision Theory and Bayesian Analysis*. 2d ed. Springer, 617 pp.

Berliner, L. M., R. A. Levine, and D. J. Shea, 2000: Bayesian climate change assessment. *J. Climate,* **13,** 3805–3820.

Blackman, R. B., and J. W. Tukey, 1958: *The Measurement of Power Spectra from the Point of View of Communication Engineering*. Dover Publications, 190 pp.

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.

Crowley, T. J., 2000: Causes of climate change over the last 1000 years. *Science,* **289,** 270–277.

Duda, R. O., and P. E. Hart, 1973: *Pattern Classification and Scene Analysis*. John Wiley, 482 pp.

Feichter, J., U. Lohmann, and I. Schult, 1997: The atmospheric sulfur cycle in ECHAM-4 and its impact on the shortwave radiation. *Climate Dyn.,* **13,** 235–246.

Gillett, N. P., M. F. Wehner, S. F. B. Tett, and A. J. Weaver, 2004: Testing the linearity of the response to combined greenhouse

gas and sulfate aerosol forcing. *Geophys. Res. Lett.,* **31,** L14201, doi:10.1029/2004GL020111.

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

——, P. D. Jones, and T. P. Barnett, 2001: Effect of observational sampling error on the detection of anthropogenic climate change. *J. Climate,* **14,** 198–207.

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.

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.

Jeffreys, H., 1935: Some tests of significance, treated by the theory of probability. *Proc. Cambridge Philos. Soc.,* **31,** 203–222.

——, 1961: *Theory of Probability*. 3d ed. Oxford University Press, 470 pp.

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.

Kaplan, W., 1992: *Advanced Calculus*. 4th ed. Addison-Wesley, 746 pp.

Kass, R. E., and A. E. Raftery, 1995: Bayes factors. *J. Amer. Stat. Assoc.,* **90,** 773–795.

Lee, T., F. Zwiers, G. Hegerl, X. Zhang, and M. Tsao, 2005: A Bayesian approach to climate change detection and attribution assessment. *J. Climate,* **18,** 2429–2440.

Legutke, S., and R. Voss, 1999: The Hamburg atmosphere-ocean coupled circulation model ECHO-G. Tech. Rep. 18, German Climate Computre Centre (DKRZ), Hamburg, Germany, 62 pp.

Meehl, G. A., W. M. Washington, C. M. Ammann, J. M. Arblaster, T. M. L. Wigley, and C. Tebaldi, 2004: Combinations of natural and anthropogenic forcings in twentieth-century climate. *J. Climate,* **17,** 3721–3727.

——, ——, H. Paeth, and W.-T. Kwon, 2004: A Bayesian decision method for climate change signal analysis. *Meteor. Z,* **13,** 421–436.

——, ——, and W.-T. Kwon, 2005a: Regional-scale climate change detection using a Bayesian decision method. *Geophys. Res. Lett.,* **32,** L03706, doi:10.1029/2004GL021028.

——, S. Legutke, A. Hense, and W.-T. Kwon, 2005b: Internal variability in a 1000-year control simulation with the coupled climate model ECHO-G—I. Near-surface temperature, precipitation and sea level pressure. *Tellus,* **57A,** 605–621.

——, ——, ——, and ——, 2005c: Internal variability in a 1000-year control simulation with the coupled climate model ECHO-G—II. El Niño Southern Oscillation and North Atlantic Oscillation. *Tellus,* **57A,** 622–640.

——, ——, ——, U. Cubasch, W.-T. Kwon, J.-H. Oh, and U. Schlese, 2006: East Asian climate change in the 21st century as simulated by the coupled climate model ECHO-G under IPCC SRES scenarios. *J. Meteor. Soc. Japan,* **84,** 1–26.

Mitchell, J. F. B., D. J. Karoly, G. C. Hegerl, F. W. Zwiers, M. R. Allen, and J. Marengo, 2001: Detection of climate change and attribution of causes. *Climate Change 2001: The Scientific Basis,* J. T. Houghton et al., Eds., Cambridge University Press, 695–738.

Roeckner, E., and Coauthors, 1996: The atmospheric general circulation model ECHAM-4: Model description and simulation of present-day climate. Max Planck Institute Rep. 218, Hamburg, Germany, 90 pp.

——, L. Bengtsson, J. Feichter, J. Lelieveld, and H. Rodhe, 1999: Transient climate change with a coupled atmosphere–ocean GCM including the tropospheric sulfur cycle. *J. Climate,* **12,** 3004–3032.

Schnur, R., and K. Hasselmann, 2005: Optimal filtering for Bayesian detection and attribution of climate change. *Climate Dyn.,* **24,** 45–55.

Stott, P. A., and S. F. B. Tett, 1998: Scale-dependent detection of climate change. *J. Climate,* **11,** 3282–3294.

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