JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 105, NO. D12, PAGES 15,525–15,542, JUNE 27, 2000

Uncertainty levels in predicted patterns of anthropogenic climate change

Tim P. Barnett,[1] Gabriele Hegerl,[2] Tom Knutson,[3] and Simon Tett[4]

Abstract. This paper investigates the uncertainties in different model estimates of an expected anthropogenic signal in the near-surface air temperature field. We first consider nine coupled global climate models (CGCMs) forced by CO_2 increasing at the rate of 1%/yr. Averaged over years 71–80 of their integrations, the approximate time of CO_2 doubling, the models produce a global mean temperature change that agrees to within about 25% of the nine model average. However, the spatial patterns of change can be rather different. This is likely to be due to different representations of various physical processes in the respective models, especially those associated with land and sea ice processes. We next analyzed 11 different runs from three different CGCMs, each forced by observed/projected greenhouse gases (GHG) and estimated direct sulfate aerosol effects. Concentrating on the patterns of trend of near-surface air temperature change over the period 1945–1995, we found that the raw individual model simulations often bore little resemblance to each other or to the observations. This was due partially to large magnitude, small-scale spatial noise that characterized all the model runs, a feature resulting mainly from internal model variability. Heavy spatial smoothing and ensemble averaging improved the intermodel agreement. The existence of substantial differences between different realizations of an ensemble produced by identical forcing almost requires that detection and attribution work be done with ensembles of scenario runs, as single runs can be misleading. Application of recent detection and attribution methods, coupled with ensemble averaging, produced a reasonably consistent match between model predictions of expected patterns of temperature trends due to a combination of GHG and direct sulfate aerosols and those observed. This statement is provisional since the runs studied here did not include other anthropogenic pollutants thought to be important (e.g., indirect sulfate aerosol effects, tropospheric ozone) nor do they include natural forcing mechanisms (volcanoes, solar variability). Our results demonstrate the need to use different estimates of the anthropogenic fingerprint in detection studies. Different models give different estimates of these fingerprints, and we do not currently know which is most correct. Further, the intramodel uncertainty in both the fingerprints and, particularly, the scenario runs can be relatively large. In short, simulation, detection, and attribution of an anthropogenic signal is a job requiring multiple inputs from a diverse set of climate models.

1. Introduction

Attempts to detect an anthropogenic signal in the observations require first that the signal be defined a priori. This is usually accomplished by sophisticated climate models forced by various time-dependent anthropogenic "scenarios," i.e., quantitative estimates of how various pollutant gases have changed in the 20th century [cf. *Santer et al*., 1996]. The anthropogenic "signal" is defined, after removal of long-term means and the seasonal cycle, from the model output of such scenario runs. Almost all of the prior detection studies have taken this approach to signal definition, e.g. *Barnett et al*., 1991; *Mitchell et al*., 1995; *Tett et al*., 1996; *Hegerl et al*., 1996, 1997, 1999a; *North and Stevens*, 1998; *North and Kim*, 1995.

Until very recently, detection studies did not address questions such as how similar are the predicted anthropogenic signals that we wish to detect when they are obtained from different models? Certainly, we might expect models forced by the same anthropogenic source to differ if for no other reason than the way the forcing is incorporated into the models. Also, the expressions for internal atmospheric physics in the models vary, for example, the radiation code, and that may introduce differences in response and sensitivity via cloud formulations [e.g., *Cess et al*., 1990; *Senior and Mitchell*, 1993]. Intermodel differences of this kind lead to uncertainty in conclusions drawn from a single model run, and here we refer to this as intermodel variability [cf. *Hegerl et al*., 1999a]. Finally, the large levels of internal dynamical variability within the complex climate models are well documented [e.g., *Palmer et al*., 1994; *Barnett*, 1995]. So we can expect that repeated model simulations made using the same model and identical representations for the forcings will give somewhat different results (intramodel variability) if they are started from slightly different

[1]Scripps Institution of Oceanography, La Jolla, California.
[2]Texas A&M University, College Station.
[3]Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey.
[4]Hadley Centre for Climate Prediction and Research, Meteorological Office, Bracknell, England.

Paper number 2000JD900162.
0148-0227/00/2000JD900162$09.00

Table 1. CGCM Control and 1% CO_2 Runs

Model	Flux Adjustment	Resolution* in Atmosphere and Ocean Components
CERFACS (European Centre for Research and Advanced Training in Scientific Computation)	none	A: 5.6 × 5.6 L30 O: 2.0 × 2.0 L31
CSIRO (Commonwealth Scientific and Industrial Research Organization)	heat, water momentum	A: 3.2 × 5.6 L9 O: 3.2 × 5.6 L21
GFDL (Geophysical Fluid Dynamics Laboratory)	heat, water	A: 4.4 × 7.5 L9 O: 4.5 × 3.8 L12
GISS (Goddard Institute for Space Sciences)	none	A: 4.0 × 5.0 L20 O: 4.0 × 5.0 L13
LMD (Laboratoire de Meteorologie Dynamique)	none	A: 1.6 × 3.8 L15 O: 2.0 × 2.0 L31
MPI (Max Planck Institute for Meteorology) (ECHAM3 + LSG)	heat, water momentum	A: 5.6 × 5.6 L15 O: 3.5 × 3.5 L11
MRI (Meteorological Research Institute)	heat, water	A: 5.0 × 4.0 L17 O: 2.0 × 2.5 L21
NCAR (National Center for Atmospheric Research) (CSM)	none	A: 2.8 × 2.8 L18 O: 2.0 × 2.0 L45
HadCM2 (Hadley Centre, UK Meteorological Office)	heat, water	A: 2.5 × 3.8 L15 O: 2.5 × 3.8 L20

*Approximate latitude and longitude intervals and number of levels in the vertical. Latitude intervals are variables in some cases, for example, atmosphere models using spectral transform techniques, or ocean models with enhanced resolution near the equator. Latitude-longitude resolution in the CMIP database is interpolated from the original (finer) grid.

initial conditions. This effect will be particularly marked during times when the anthropogenic signal predicted by the models is weak relative to the internal model noise [*Cubasch et al*., 1994; *Santer et al*., 1995a].

The key question we address here is how important to the detection problem are the uncertainties in predicted anthropogenic signals due to both intermodel and intramodel variability (i.e., due to differences in model formulation and internal climate variability within a single model). We note the current study is complementary to that of *Hegerl et al*. [1999a] which approaches the problem from another perspective than used here. To answer the question, we not only compare the model signals with observations but also estimate the size of the intermodel and intramodel differences relative to the predicted signal in a selected set of models. We will also estimate the pattern similarity between different estimates of the "same" anthropogenic response signal, as most modern detection methods use, at least partially, some form of pattern recognition. Large differences in the spatial patterns of model response to the same forcing would make detection and attribution more difficult, for we do not know which model, if any, is correct. Note our approach may not detect biases common to all models, unless they appear in the comparisons with observations.

The paper is arranged as follows: Section 2 describes the model data used in this study and some of the analysis techniques employed. We next explore the behavior of nine coupled climate models (CGCMs) forced by increasing CO_2 in an effort to see if large differences exist between different models' response to this relatively strong but idealized anthropogenic forcing. Section 4 concentrates mainly on intramodel differences arising in 11 simulations from three different models forced by greenhouse gases (GHG) and the direct sulfate aerosol effect (SUL). Sections 5 and 6 discuss implications for detection studies based on the results of sections 3 and 4. A final section summarizes the main points of this work.

2. Methods and Data

2.1. Data Sources

Two major sets of CGCM scenario run data were used in this study. The first comes from the coupled model intercomparison project (CMIP) [*Meehl et al*., 1997], an element of the Program for Climate Model Diagnosis and Intercomparison (PCMDI) [*Gates*, 1992; *Gates et al*., 1999]. The annual mean data we used here were part of the CMIP2 subproject and kindly provided by Curt Covey. In CMIP2, nine different CGCMs were first run for 80 years in control run mode. The initial conditions for these runs were selected from a longer control run of the individual models that had achieved steady state. The same models were then run for another 80 years with forcing corresponding to an increase of CO_2 at a rate of 1% per year compounded. The long-term mean, annual near-surface temperature from the 80 year control run was computed for each model and subtracted from that model CO_2 run to form annual anomalies that would be subjected to future analysis. The models, their resolution, etc., are given in Table 1, and additional information may be found on the World Wide Web (www.pcmdi.llnl.gov/modeldoc/cmip/).

The second set of scenario runs represents a more realistic set of experiments, albeit still limited by forcings known to be radiatively important that were omitted from the numerical experiments. Three different CGCMs were forced with a combination of greenhouse gases (GHG) and the direct effects of sulfate aerosols (SUL), the latter being expressed as changes in surface albedo in the models. The simulations began in the late 1800s and continued into the 21st century (see Table 2 for details). The observed/estimated gas concentrations were used up to modern times to force the models, and then an IPCC scenario for concentrations into the future was used. The important feature of these runs is that they were conducted in ensemble mode. There were two realizations of the Max Planck Institute (MPI) run [*Cubasch et al*., 1997], five realiza-

Table 2. GHG + SUL Scenario Runs

Model	Flux Correction	Control Run Length (year)	Run Period	Number Realizations	Resolution in Atmosphere and Ocean Components
GFDL	yes	675	1865–2014*	5	A: 3.75 × 2.25, L = 14 O: 1.9 × 2.25, L = 18
HadCM2	yes	1700	1861–2099	4	A: 3.75 × 2.5, L = 19 O: 3.75 × 2.5, L = 20
MPI	yes	1500	1880–2049	2	A: 5.625 × 5.625, L = 19 O: 3.5 × 3.5, L = 11

*Runs continuing.

tions of the Geophysical Fluid Dynamics Laboratory (GFDL) run [*Knutson et al*., 1999], and four realizations of the Hadley Center run [*Tett et al*., 1999]. In all cases the different realizations varied only in that their initial conditions which were taken from different states of a long control run separated by 50–130 years. The mean near-surface air temperature averaged over June, July, and August (JJA) derived from a long control run was used with the scenario runs to compute anomalies for the JJA period. This part of the year was selected because recent detection and attribution studies have concentrated on this period [e.g., *Santer et al*., 1995b; *Hegerl et al*., 1996, 1999a; *Barnett et al*., 1998].

Observed near-surface temperature data were also used in the study. This data set is a combination of land air temperature anomalies [*Jones*, 1994] and sea surface temperature anomalies [*Parker et al*., 1995] on a 5° × 5° grid-box basis. The merging of the two data sets is discussed by *Parker et al*. [1994]. Both components of the data set are expressed as anomalies from 1961 to 1990. The data set has been extensively used in the various IPCC reports [e.g., *Nicholls et al*., 1996]. The recent study of *Hegerl et al*. [1999b] showed that the observational sampling errors in these data were not large enough to seriously affect detection studies.

All of the data used below has been projected onto a common T42 grid, ~2.8° × 2.8° on a side. This was done via bidirectional linear interpolation. Notice that in most cases the original model runs were at a resolution lower than the T42, so the interpolation process did not introduce high wave number noise to the data. The fields were further limited to the domain where a reasonable amount of observations exist, 5219 grid points out of a possible 8192. In practice, this means regions in the Southern Hemisphere below about 40°S were largely ignored, as were the highest latitudes of the Northern Hemisphere.

2.2. Analysis Methods

Two main types of analysis were used. Pattern correlations were used to quantify the level of similarity between various patterns of climate change produced by the models. Also, common empirical orthogonal functions (cEOFs) were used to study the relative magnitudes of the differences between model simulations of anthropogenic responses, something simple pattern correlations cannot do.

Similarities between patterns of change produced by different models (i and j) are easily quantified via the pattern (P) correlation [e.g., *Richman*, 1986] defined as

$$P_{ij} = \frac{\langle (T_i - \langle T_i \rangle_x)(T_j - \langle T_j \rangle_x) \rangle}{[\langle (T_i - \langle T_i \rangle_x)^2 \rangle \langle (T_j - \langle T_j \rangle_x)^2 \rangle]^{1/2}}, \quad (1)$$

where $T_i = T_i(x, t)$, say, is the near-surface air temperature anomaly field from the ith model or realization relative to its climatology from a specific ensemble. The brackets represent an averaging operation over x. The mean that is removed in the calculation should be computed as a spatial average over the T field at a single time "t." Given the normalization in (1), P will have value 1.0 if the two patterns being compared are identical and 0.0 if they are orthogonal. Note that the T field is area weighted by the cosine of the latitude prior to estimation of P.

The significance of the pattern correlations was determined relative to the GFDL control run. Successive patterns of 50 year summer trends were estimated from this run. The pdf of these trend patterns was used to determine the significance of the same length trend patterns estimated from the scenario runs (section 4). This same procedure was used to estimate significance of trends in smoothed data, the control run smoothing being the same as that for the scenario runs (section 5.1).

The common EOFs were computed in the manner described by *Barnett* [1999]. The model temperature fields were concatenated as shown by *Barnett and Preisendorfer* [1987] to form a "single" data set; that is,

$$T'(x, t') = \begin{cases} T_1(x, \eta), & t' = 1, 2, \cdots, m; \quad \eta = 1, 2, \cdots, m \\ T_2(x, \eta), & t' = m+1, \cdots, 2m \quad \eta = 1, 2, \cdots, m \\ \vdots & \vdots \quad \vdots \end{cases} \quad (2)$$

where T_i is the near-surface air temperature anomaly field for the ith CGCM ($i = 1, 2, \cdots, N$), x is a spatial grid point counter for locations where data exist ($x_{\max} = 5219$) and is the same for all the models, each grid point time series is m terms long, and t' is a dummy time variable that describes the order of concatenation.

Each model's time series has been adjusted to have zero mean at each grid point. This means the array T′ can be immediately subjected to a standard EOF (empirical orthogonal function) analysis of its covariance matrix. Prior to estimation of the covariance matrix, each grid point time series is weighted by the cosine of its latitude. Note also that models with higher variability will tend to dominate the analysis if the covariance matrix is used. Since such model properties are something of keen interest, the covariance matrix is more informative than analysis of the correlation matrix which would suppress intermodel variance differences.

The results of the above analysis produce what we term common EOFs or cEOFs (not be to be confused with complex

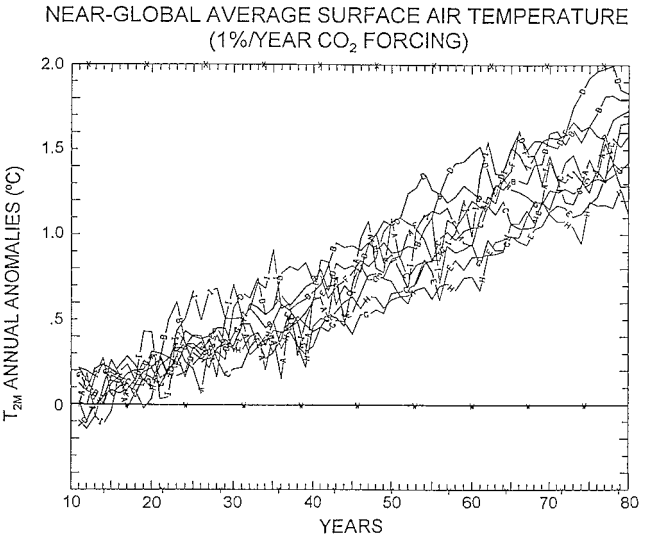

Figure 1. Near-global average surface temperature (degrees C) from nine CGCMs forced by idealized CO_2 signal increasing at 1% per year. The last 70 years of 80 year integrations are shown.

EOFs or CEOFs). The eigenvectors, after correction for the cosine weighting, represent the patterns of variability that the N CGCM runs share in common. It is convenient to define a partial eigenvalue, S_{ni}, to represent the relative contribution of the ith model to the energy of the nth eigenmode as follows:

$$S_{ni} = \frac{1}{m} \sum_{j=1}^{m} a_n^2(t_j'), \tag{3}$$

where the sum is over the ith individual m-component data blocks defined by (3), of which there are N. The distribution of the S_{ni} with mode n gives an approximation of the eigenvalue spectrum for model "i." This offers an immediate way to quantitatively intercompare both the patterns and the levels of internal variability of the runs as a function of pseudo-wavenumber (EOF mode number). See also the use of common EOFs described by *Stouffer et al*. [1999].

3. Results: Idealized CO_2 Forcing

This set of runs, although for an idealized anthropogenic scenario, offers an excellent opportunity to determine the magnitude of intermodel differences relative to the signal they are producing since they have similar forcing. Some of the uncertainty will come from internal model variability, but time averaging over the last 10 years of each run should reduce potential impacts of this intramodel variability. We expect the larger differences will come from differences in model physics, for example, the manner in which the changes in CO_2 affect the radiative forcing in each model and the associated feedbacks.

3.1. Traditional Analysis

It is useful when analyzing anthropogenically forced runs to show a map of the expected signal at some time in the future and/or the global average temperature changes expected as a function of time. These quantities are summarized in this section.

The time-dependent near-global mean was computed for each model and is shown in Figure 1. All models produce a change of about 1.5°C over the last 70 years (years 11–80) of the integrations, the first 10 years being omitted from analysis to avoid any potential start-up problems. The results look rather similar. Close inspection shows that by the end of the runs the global temperature estimate varies between 1.2° and 2.0°C, and the final decadal averages vary between 1.1° and 1.8°C. These ranges, centered on the mean of about 1.5°C, suggest the models agree to within about ±25% at the end of the 80 years of integration, a surprisingly small value given the differences in natural variability in the models and differences in the climate sensitivity due to different physical parameterizations.

The spatial pattern of change was computed by first averaging the last 10 years of each integration. (Averaging over the last 20 years produced basically the same result. A complementary analysis via cEOFS of the CMIP2 models giving much the same results and is given in the Appendix.) The resulting nine maps of the spatial signal over this period were then averaged together at each grid point, i.e., a nine model average of the spatial distribution of the "signal" (SMEAN) with the global average temperature left in. The intermodel standard deviation of this mean, due to both response differences and internal climate noise, was computed at each grid point from the nine model realizations and denoted by "N." The global mean of the nine model average computed over all grid points was removed to isolate better the spatial characteristics of the anthropogenic signal (SNOMEAN) without the global average temperature. The signal to noise ratios (SMEAN/N) and (SNOMEAN/N) are shown in Figure 2, along with the nine model average before removal of the global mean.

The results show that retention of the global mean provides a signal that is robust relative to the intermodel variability, i.e. the ratio SMEAN/N is greater than 1.0 over most of the globe (Figure 2, middle). Conversely, removal of the global mean (Figure 2, bottom, SNOMEAN/N) gives a signal that is poorly defined over most of the tropical and midlatitude ocean areas of the Earth (stippled areas), since SNOMEAN/N is less than 1.0. Together, these results mean the agreement of the model is driven by their estimates of global mean temperature and enhanced warming over land. Removal of this mean produces signal patterns that vary greatly between models. In detection work, the changes in global mean would be useful for detecting "climate change" but would be of modest use in "attributing" changes to specific physical processes (see *Tett et al*. [1999] for an example).

3.2. Pattern Similarity

The strongest signal was at the end of the integrations and was isolated by taking an average of the last 10 years of each model run. The pattern correlations between the nine estimates of this spatial signal are given in Table 3. Values range from 0.08 to 0.72 with values above 0.46 being significantly different than the 80 year long CMIP2 control run values at the 0.05 level. Note that most of the anthropogenic model pattern correlations are significantly higher than were obtained from their respective control run alone.

The reasons for the largest disparities in Table 3 become clear when we investigate the patterns that contribute to the extremes of this range of correlations (Figure 3). The global mean temperature has been removed from each panel to emphasize better the patterns. The MPI-NCAR patterns (corre-

lation 0.08, Figures 3a–3b) show the former model to produce a much larger signal over most of the continental regions than the latter model, especially over South America, North America, and Africa. We speculate that land processes, for example, soil moisture, are handled quite differently between these models, an idea supported by the recent work of *Räisänen* [1999]. Note the large response in the NCAR simulation in the Bering Sea and Greenland-Iceland area, regions where MPI has little response. The most likely cause for these differences is in the way the models account for sea ice. In contrast the GFDL-MPI comparison (correlation 0.72, Figures 3b–3c) shows most of the main features are in qualitative agreement. The main differences are in small displacements between the main response regions, for example, compare results over North America and, in the highest latitudes, for example, the Bering Sea region. This is the best case of pattern agreement in Table 3.

3.3. Summary

Patterns of near-surface air temperature change predicted to occur as CO_2 increases at a rate of 1%/yr by nine (9) different global climate models agree well with respect to the global mean. The spatial patterns of change which accompany the changing mean agree moderately well and demonstrate clearly the strong contrast in land-ocean temperature change and high-latitude warming. Consideration of the largest spatial response shows that the various change patterns predicted by the model share a substantial 46% of their variance in common (compare the Appendix). Visual inspection of the individual model results leads us to speculate that the differences in responses may be due to the basic model physics, or parameterizations, especially those affecting land and sea ice processes, an idea supported by the recent work of *Räisänen* [1999].

4. Results: GHG and Direct Sulfate Aerosols

The group of simulations discussed below is more realistic than those presented above but still lacking forcing by key chemical constituents, for example, indirect sulfate aerosols, ozone, etc. Even the forcings they do use differ in fundamental ways; for example, the direct sulfate forcing has a time-independent pattern, with a time-varying amplitude, in the Hadley Centre simulations [*Mitchell et al*., 1995] but a time varying pattern and amplitude in the MPI runs. Nevertheless, at least two models forced by the combination of GHG and direct sulfate aerosols (SUL) produce surface temperature changes over 1945–1995 which are consistent with those found in the observations [e.g., *Hegerl et al*., 1997, 1999a; *Tett et al*., 1996, 1999; *Knutson et al*., 1999; *Barnett et al*., 1999, section 6], although these results carry numerous caveats (compare same authors). It is for this reason that this section will concentrate on the characteristics of those model signals during this recent (1945–1995) period, the latter year being selected to avoid trend estimate distortion associated with the huge El Niño/Southern Oscillation (ENSO) event of 1997–1998. We further restrict the analysis to the northern summer (June/July/August, JJA), a time of the year when signal to noise is a maximum and detection chances particularly for sulfate signals enhanced (*Hegerl et al*. [1997], *Barnett et al*. [1998], *Santer et al*. [1995b] results suggest SON might also be a good season for detection).

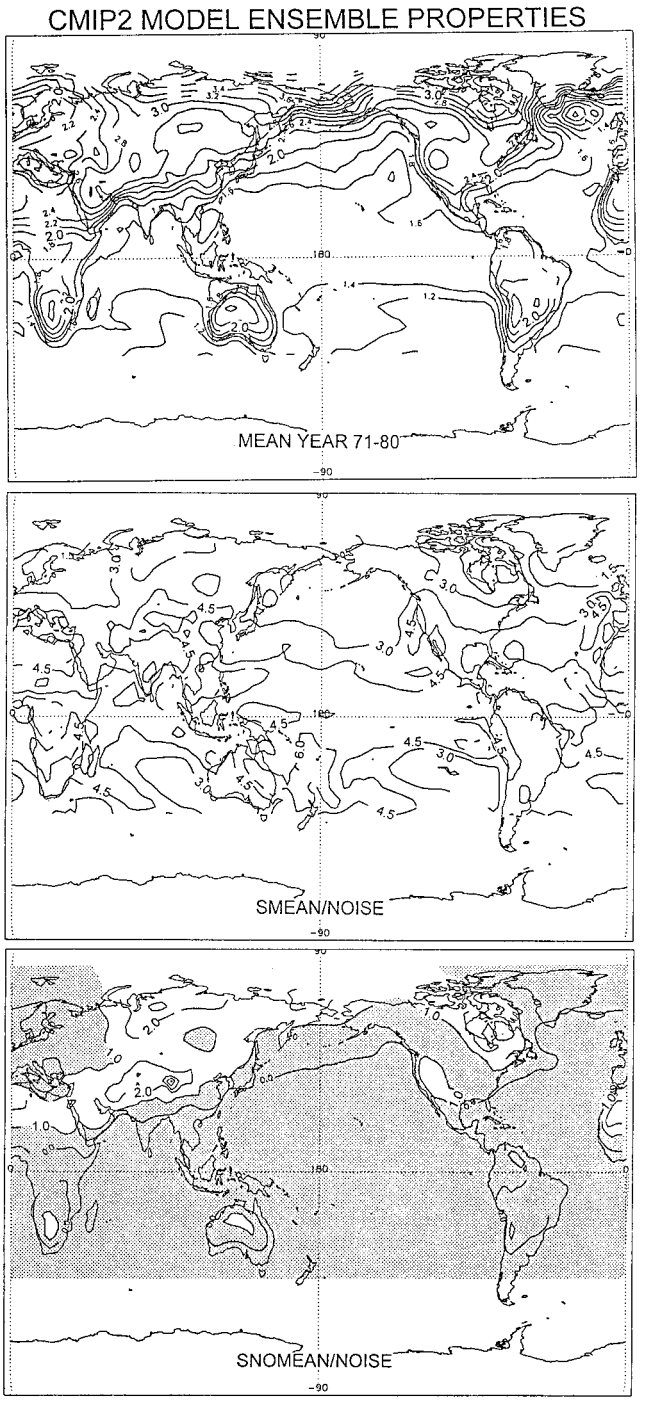

Figure 2. (top) Annual temperature anomaly averaged over the last decade over all nine model simulations forced by 1% per year increase in CO_2 concentrations. (middle) Ratio of the signal shown in the top panel divided by the standard deviation about this mean computed from the nine member ensemble. (bottom) Same as middle panel but with global mean removed.

4.1. Traditional Analysis

The smoothed near-global mean temperature for June/July/August (JJA) from 1910 to 2010 is shown in Figure 4 for all 11 realizations from the three models. We note other phases of the seasonal cycle might be useful for detection, for example, winter [*Santer et al*., 1996]. The observed near-global mean is shown for comparison. All series have been filtered with a 10

Table 3. Pattern Correlation for the Idealized CO_2 Runs

	HadCM2	NCAR	MRI	LMD	GISS	GFDL	MPI	CSIR	CERF
CERF	(0.59)	0.18	0.42	(0.60)	(0.56)	(0.64)	(0.70)	(0.57)	
CSIR	(0.58)	(0.48)	(0.47)	0.45	0.44	(0.60)	(0.51)		
MPI	(0.62)	0.08	0.39	(0.70)	0.45	(0.72)			
GFDL	(0.58)	0.22	0.44	(0.53)	(0.51)				
GISS	(0.58)	0.27	(0.51)	0.49					
LMD	(0.62)	0.17	0.35						
MRI	(0.57)	0.17							
NCAR	0.36								
HadCM2									

Values (in parentheses) of pattern correlations were likely to occur in the selected model control run less than 10% of the time.

year smoother. The trend over the period 1945–1995 from the simulations varies between 0.016° and 0.048°C per decade for the models. The trends of the averaged GFDL and Hadley Centre ensembles are 0.038° and 0.031°C per decade, respectively. Analysis of the observations gave a value of 0.028°C per decade over the 1945–1995 period. So the average properties of the raw trend from the models and observations are in relatively good agreement, a result whose statistical significance is verified in section 6. The model results do not take into account any volcanic or solar influences [e.g., *Cubasch et al*., 1997; *Tett et al*., 1999; *Stott et al*., 1999]. Note also the single model integration that shows warming until about 1940 and cooling afterward, much like the observations. This interesting result is from a GFDL realization (GF3 in Tables 4 and 5 and G3 in Plate 1) and discussed by *Knutson et al*. [1999].

The spatial patterns of trend over 1945–1995 are shown in Figure 5 for both the observations and the average of all of the 11 realizations from the three models. (Use of trend to represent change between 1945 and 1995 is a good first-order approximation but suboptimal from a detection point of view [cf. *Santer et al*., 1996; *Wigley et al*., 1998]. It is, however, the metric used in many current detection studies and so its use here facilitates comparison with the recent work.) Inspection of the figure shows the models have both overestimated and underestimated the rate at which the temperature has changed in some regions of the world. However, the fundamental patterns appear to have some key differences, especially over the Northern Hemisphere landmasses and oceans, where negative trends are observed, while average model trends are positive (see also *Knutson et al*. [1999] who obtained a similar result). These differences are thought to be due to a number of factors, including those associated with indirect sulfate aerosol forcing and differing representations of ENSO, Pacific Decadal Oscillation [e.g., *Trenberth and Hurrell*, 1994; *Barnett et al*., 1999b], and other "natural" climate modes. Note the averaged model pattern will be smoother and will have less distinctive features than the observations, simply as a result of the variance reduction produced by the averaging. There is only one realization of the observations, and so it will be noisy.

4.2. Pattern Similarity

The differences between the simulation of different models with what was supposed to be essentially the same forcing were investigated as above. Each realization from each model was used to estimate the linear trend in surface air temperature between 1945 and 1995, as a function of spatial location. This resulted in 11 different maps, one for each model/realization, of gridded trend values over the data-adequate regions. The globally averaged means of these trends, discussed in section 4.1 above, were removed from their respective maps and the pattern correlations computed. The observed trends also were included in this analysis set.

The results are shown in Table 4. The raw 50 year JJA trend maps correlate with each other in the range −0.11 to 0.45, with values exceeding 0.30 being likely from an identical analysis of the GFDL control run only 10% of the time. In this case, the GFDL control run is taken to represent natural, internal variability in a climate system with no anthropogenic forcing. Twelve of the 55 individual realization interpattern correlations equal or exceed 0.30. The models pattern-correlate with the observed pattern in the range −0.10 to 0.29. Values of 0.23 or greater are exceeded 10% of the time if one replaces the anthropogenically forced patterns with comparable length segments of the GFDL control run; that is, only two of the 12 individual realization correlations are significant. Note that this test basically asks if the agreement between the observations and the anthropogenic runs differs at all from that found between the observations and the control runs. We obtained essentially the same results by replacing the observations with comparable chunks from the GFDL control run and then comparing these with the anthropogenic runs.

The pattern correlation between the ensemble averages of the GFDL and HadCM2 ensembles is 0.37, a value that occurred just 10% of the time in an identical analysis of the pseudoensembles drawn from the GFDL control run, so the ensemble averages from the two anthropogenically forced models are said to be similar to each other. The correlations between each of these ensemble mean patterns and the observed pattern of trend are 0.09 and 0.17, respectively. Neither of these last values is significantly different than expected from the GFDL control run compared to observations. Finally, the grand average of all 11 runs ("AVG" in Table 4) significantly correlates with all of the individual runs (except HC4) and the observations. So ensemble averaging enhances agreement by reducing model-induced noise, such that there is clearly substantial agreement between the various model realization estimates of the temperature trend over the last 50 years. However, these simple patterns of trend change are not correlated significantly with the observations. As we shall see in section 6, use of optimal detection techniques gives just the opposite result.

The reasons for the largest disparities can be seen on the individual, unsmoothed trend maps (after removal of the globally averaged trends) associated with the largest positive and negative correlations in Table 4 (Figure 6). The top two panels correlate at −0.15, the largest model-model disagreement

found in Table 4. The Hadley Centre model (HadCM2) realization 4 (denoted by HC-4) shows a relative warming trend over most of the Northern Hemisphere, while the MPI model realization "A" shows large areas of relative cooling over the same region, especially over Eurasia, the western Pacific, and North Atlantic. These are the regions where the sulfate forcing is strongest. *Hegerl et al*. [1997, 1999a] show the sulfate signal

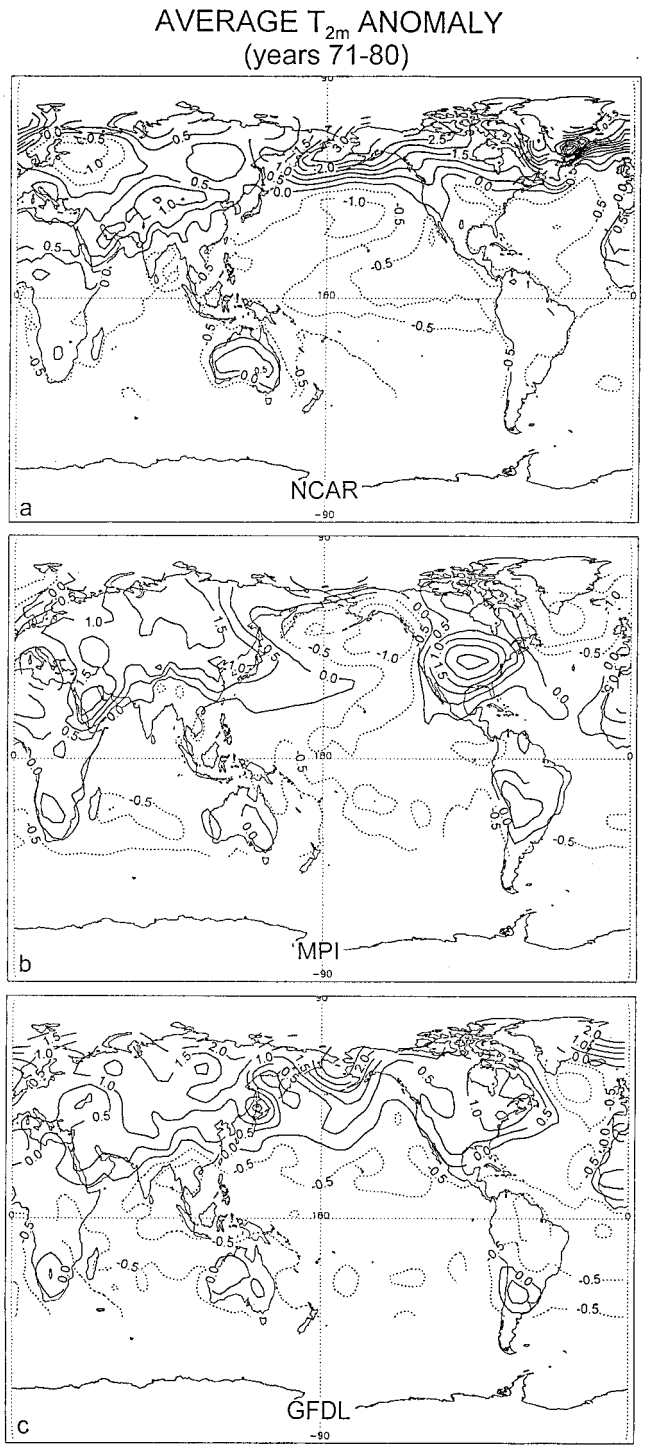

Figure 3. Average of the last decade annual near-surface temperature (degrees C) from three members of the ensemble. (a) from the NCAR model, (b) from the Max Planck Institute model, and (c) from the GFDL (R15) model.

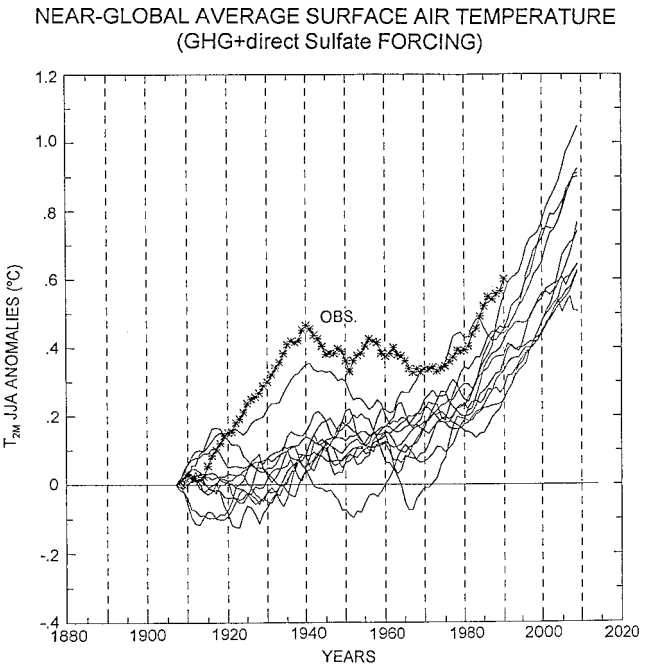

Figure 4. Near-global average surface temperature for JJA (degrees C) from the 11 member ensemble of coupled models forced by a combination of greenhouse gases (GHG) and direct sulfate effects (SUL) shown by solid lines. Observational equivalent shown by connected asterisks. All data were put through a 10 year filter.

in the MPI models is substantially stronger than in the Hadley Centre model. It seems probable, then, that the difference in forced response between the two models is due to this fact. We cannot, however, discount the fact that the HC run 4 is a statistical outlier produced mainly by internal model variability, a suggestion that will be confirmed in section 6 (see Plate 1). In any event the existence of results such as that from HC-4 clearly show that future detection work on CGCM simulations needs to be done on ensembles of scenario runs, not individual runs. Such an approach to detection is present in the studies of *Tett et al*. [1996, 1999], *Stott et al*. [1999], and *Hegerl et al*. [1999a].

The bottom two panels of Figure 6 have one of the highest pattern correlations of any of the simulations (0.40). Here the agreement is due to the coexistence of large regions of tropical and Southern Hemisphere with relative warming and relative northern midlatitude cooling in both the MPI realization "A" and the Hadley Centre realization 3 simulations. Remember the global average trend was removed, so the apparent cooling is relative to this global averaged trend which was positive.

Close inspection of Table 4 reveals that different simulations made by the same basic model often have low correlation. In places, the simulation-to-simulation difference is as big or bigger than the mean signal itself. The conclusion we draw is that the intramodel variability, due to internal model dynamics, is generally at least as large in the three models (at this time) as the anthropogenic signal they are producing. We shall see below that this problem can be overcome through use of ensemble properties of the various model integrations but not for observations where we have only one realization.

4.3. Common EOFs

The cEOFs were computed over a uniform record length for each model from 1900 to 1995, the same length as the obser-

Table 4. Pattern Correlation of Surface JJA Temperature Trends (1945–1995)

	OBS	AVG	GF-5	GF-4	GF-3	GF-2	GF-1	EC-B	EC-A	HC-4	HC-3	HC-2	HC-1
HC-1	(0.29)	(0.60)	0.25	0.16	0.06	(0.33)	(0.34)	0.25	0.22	0.18	0.08	(0.44)	
HC-2	0.19	(0.59)	0.24	0.25	0.13	0.15	(0.37)	0.28	0.21	0.13	0.16		
HC-3	0.19	(0.52)	−0.02	(0.41)	−0.01	0.29	0.12	(0.34)	(0.40)	0.09			
HC-4	−0.10	0.23	0.03	−0.02	−0.03	0.05	0.04	−0.05	−0.11				
EC-A	(0.26)	(0.57)	0.09	(0.32)	0.19	0.21	0.23	(0.45)					
EC-B	0.19	(0.57)	0.01	0.19	0.05	(0.36)	0.26						
GF-1	0.13	(0.62)	(0.39)	0.26	0.22	0.29							
GF-2	0.17	(0.56)	0.05	0.17	0.10								
GF-3	0.04	(0.37)	(0.35)	0.01									
GF-4	0.12	(0.50)	0.06										
GF-5	0.20	(0.44)											
AVG	(0.30)												
OBS													

Values (in parentheses) of pattern correlations were likely to occur from analysis of the GFDL control runs less than 10% of the time. HC-1 = HadCM2, realization 1; HC-2 = HadCM2, realization 2, etc.; AVG is the average of all 11 runs.

vational record used in this paper. The data used to form the covariance matrix were anomalies of near-surface temperature from the models only, computed as described above and smoothed with a 10 year filter to eliminate interannual variability such as ENSO and weighted by cosine of the latitude. Note that a simple analysis will be most influenced by the GFDL and Hadley Centre realizations simply because there are more of them. A weighting scheme inversely proportional to the number of individual model realizations was used prior to estimation of the covariance matrix to correct for this potential bias. We note that there are a number of ways this analysis could have been done (e.g., correlation matrix, different spatial weighting, etc.) which could potentially change the details of the results shown here. Hence we concentrate only on the major results which we feel are insensitive to the details of the cEOF analysis.

The leading cEOF (Figure 7) accounted for 39.4% of the common variance. The signal has a pattern of almost uniform sign over the study region with positive maxima over the land masses, especially the tropics. A notable minimum exists over the central North Pacific Ocean. However, the results suggest the response is far from spatially uniform, so there is a clear fingerprint for detection studies to attempt to find in the observations. That fingerprint essentially is one of land/sea temperature contrasts, with larger response in the middle of the major continents. It is this pattern that detection schemes must seek to discover. Note the high-latitude signal is weaker than found with CO_2 forcing only (Figure A1). The second cEOF (not shown) accounts for only about 6% of the variance and is statistically degenerate. Such information may represent a "correction" to the leading mode associated with time-dependent spatial changes in the cEOF1 pattern and so cannot be neglected out of hand.

The eigenvalue spectrum (Figure 8) shows the energy (amplitude squared) of the principal anthropogenic signal illustrated in Figure 7 varies considerably between models, being strongest in the GFDL simulations and weakest in the MPI runs. Interestingly, the signal energy variations within the GFDL or HadCM2 model ensembles are about the same size as the intermodel ensemble differences. The MPI runs contain considerably less energy (a smaller signal) than the other two models. This is likely due to the ocean component of that model, the large-scale geostrophic model, which is known to demonstrate weak interannual and interdecadal variability [e.g., *von Storch*, 1994; *Pierce et al*., 1995]. The relative large difference in signal strength between the models suggests that detection schemes ought to include signal magnitude (e.g., optimal detection methods), something simple pattern correlation schemes do not do.

Also shown in Figure 8 is the projection of the observational record onto cEOF1. Obviously, the recent changes in an annual temperature project only moderately well onto the signal pattern predicted by the combined models, with weaker loading than the models. This result is somewhat better than ex-

Table 5. Pattern Correlation of Smoothed Surface JJA Temperature Trends (1945–1995)

	OBS	AVG	GF-5	GF-4	GF-3	GF-2	GF-1	EC-B	EC-A	HC-4	HC-3	HC-2	HC-1
HC-1	0.37	(0.74)	0.45	0.43	0.29	(0.52)	(0.69)	0.44	0.39	0.16	0.25	(0.65)	
HC-2	0.25	(0.76)	(0.50)	(0.56)	0.39	0.34	(0.62)	0.42	0.36	0.34	0.33		
HC-3	0.26	(0.63)	0.12	(0.60)	0.16	0.45	0.30	(0.53)	(0.70)	−0.01			
HC-4	−0.20	0.18	0.21	0.11	0.00	−0.14	0.19	−0.11	−0.27				
EC-A	0.34	(0.71)	0.22	(0.54)	0.36	(0.46)	0.41	(0.72)					
EC-B	0.29	(0.68)	0.11	0.35	0.19	(0.61)	0.45						
GF-1	0.23	(0.81)	(0.61)	(0.53)	0.44	(0.53)							
GF-2	0.26	(0.66)	0.15	0.36	0.32								
GF-3	0.12	(0.58)	(0.61)	0.31									
GF-4	0.15	(0.72)	0.32										
GF-5	0.23	(0.61)											
AVG	0.34												
OBS													

Values (in parentheses) of pattern correlations were likely to occur from analysis of the GFDL control runs less than 10% of the time. HC-1 = HadCM2, realization 1; HC-2 = HadCM2, realization 2, etc.; AVG is the average of all 11 runs.

pected from the low correlations between observations and various realizations (compare Table 4), principally due to the filtering of high wave number information by the cEOF analysis. Remember, however, the observations contain signals due to volcanoes, solar variability, etc., signals not included in the model simulations and not completely eliminated by the temporal filtering.

The PCs of the leading cEOF (Figure 9) explain the results obtained in the eigenvalue spectrum. The trend in recent years is slightly larger in the GFDL runs, hence the larger partial eigenvalues. The opposite situation is seen for the MPI runs, which have relatively little variability outside an increasing temperature trend in the last 30 or so years of integration. The projection of the observations onto the anthropogenic signal is represented by a pseudo-PC that has low variability and shows a small increasing trend, a trend similar to that produced by many of the simulations.

4.4. Summary

The global average trends predicted by all the models are in rather good agreement among themselves and with the observations over the last 50 or so years. The raw unsmoothed individual patterns of the near-surface air temperature trend predicted to occur as a result of GHG and direct sulfate aerosol forcing by three different models exhibit a moderate level of similarity among themselves in their spatial characteristics, but they do not have a statistically significant relation with observed patterns of change. Consideration of the lowest wavenumber response, the scales used in detection studies, shows that the various models share 39% of their variance in common. The intermodel differences appear about the same size as

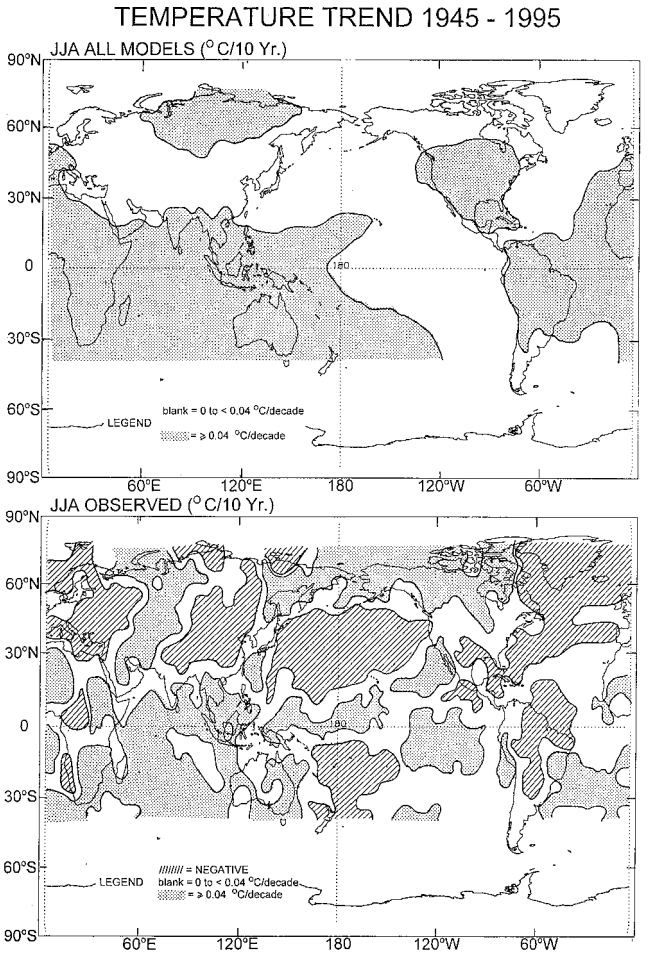

Figure 5. (top) Trend in JJA temperature averaged over all 11 GHG + SUL runs over the period 1945–1995 (degrees C per decade). (bottom) Same as top panel but for the observations. Hatched areas correspond to negative trends.

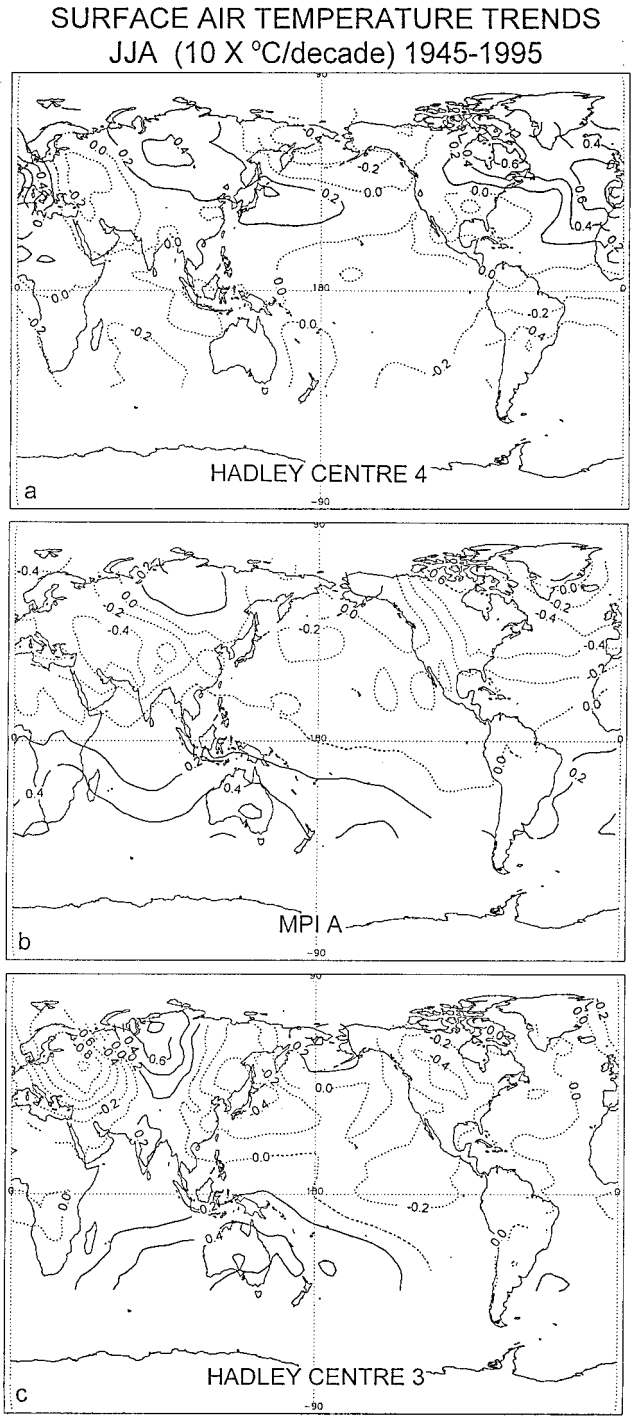

Figure 6. Patterns of temperature trend (10 × C/decade) between 1945 and 1995 from three members of the GHG + SUL ensemble. The global mean of each pattern was removed before plotting. See Table 2 for model identification codes.

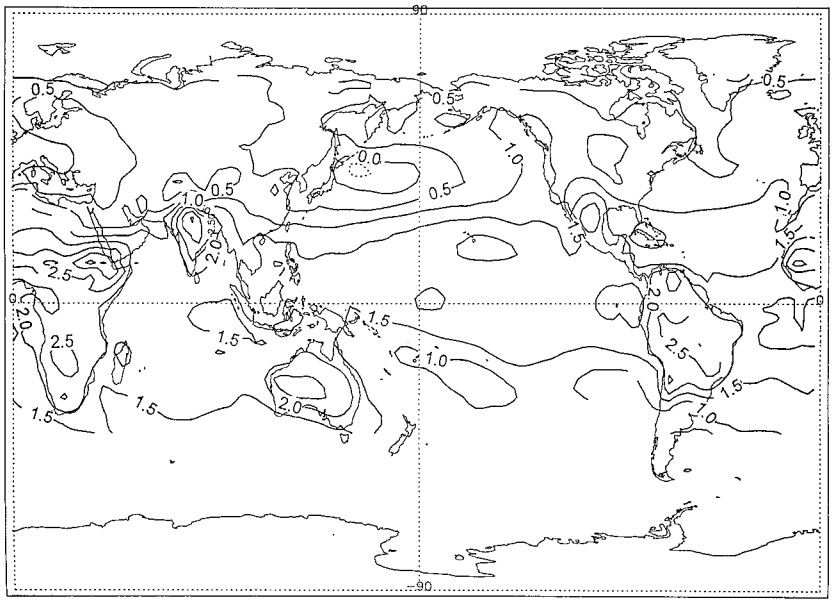

Figure 7. First common empirical orthogonal function of the 11 model ensemble forced by GHG + SUL. This mode captured 39.4% of the joint model variance in near-surface JJA air temperature. Data were put through a 10 year smoother prior to analysis.

the intramodel differences due to internal variability for two of the models.

5. Implications for Detection

5.1. Noise Suppression

The above results make it abundantly clear that several types of noise conspire to obscure the anthropogenic signal produced by the models. We demonstrate briefly here that in principle this noise contamination largely can be overcome. The outstanding message is that spatial filtering and ensemble averaging are nearly mandatory prior to any attempts to detect and attribute climate change and anthropogenic signal predicted by CGCMs. The most recent anthropogenic signal detection studies have made good use of both of these preanalysis operations [cf. *Tett et al*., 1999; *Hegerl et al*., 1999a].

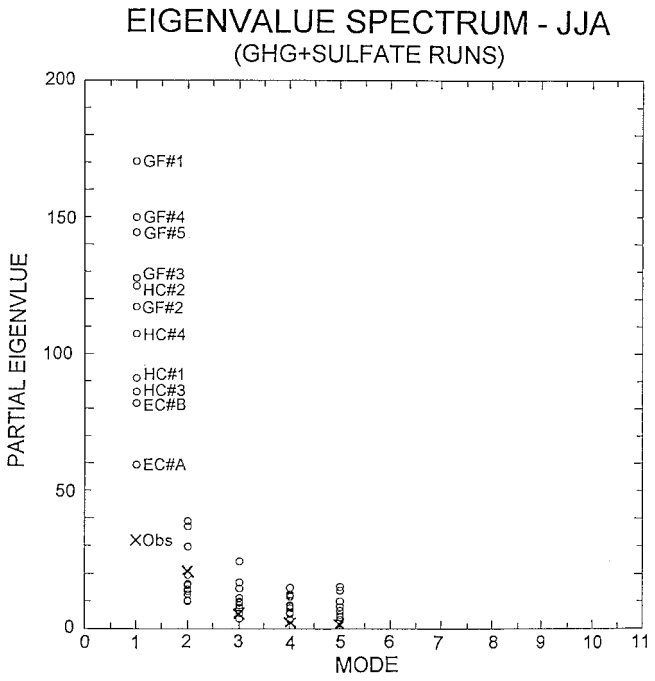

Figure 8. Eigenvalue spectrum of the 11 member ensemble common EOF analysis (units of $°C^2$).

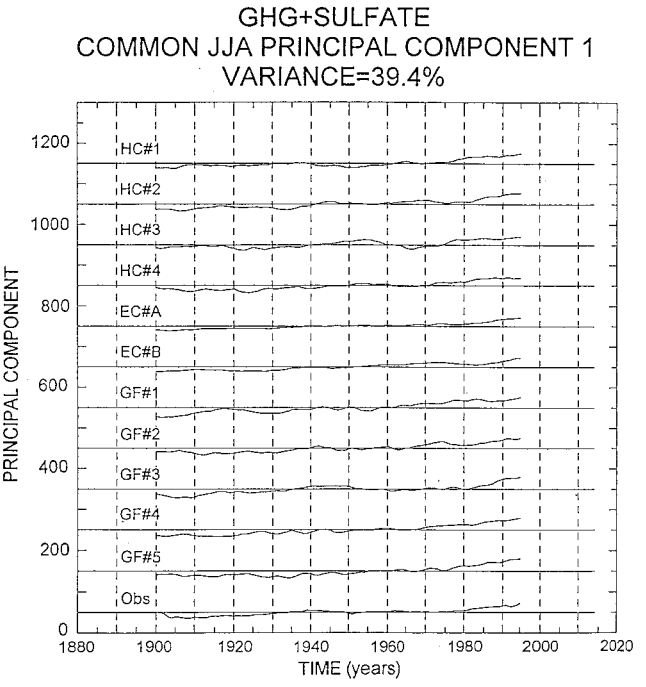

Figure 9. Leading principal components that go with common EOF 1 of the temporally smoothed data shown in Figure 7. The model identifications are given in Table 4. The pseudo-PC for the observations projected onto the model cEOF1 is shown in the bottom panel. The ordinate is offset in increments of 100 units to separate the models.

It was noted that regional climate noise was an obvious contributor to the low correspondence between the individual simulations. To minimize this factor and assess the effect of spatial filtering, we spatially smoothed the trend patterns used to compute Table 4, eliminating scales less than those corresponding to about zonal wavenumber 6. The remaining information is similar in scale to that suggested by *Stott and Tett* [1998] as the minimum useful for detection purposes. Table 5 shows the pattern correlations computed from this smoothed data.

Inspection of Table 5 shows the correlations are generally higher, as expected, with values equaling or exceeding 0.46 being likely from an identical analysis of the GFDL control run only 10% of the time. Seventeen of the pattern correlations between individual realizations are now significant versus 12 for the unsmoothed data. For instance, the GF 4 versus EC "A" pattern correlation has increased from 0.32 to 0.55. Note, however, that the spatial smoothing does not eliminate apparent outlier simulations; for example, the correlation between HC-4 and EC "A" has gone from −0.11 to −0.27. These two runs generally have the opposite spatial trend patterns and no amount of spatial smoothing will change that fact.

The intramodel variability that produced HC-4 can be partially overcome by ensemble averaging. We noted above the pattern correlation between the HadCM2 and the GFDL ensemble averages was 0.37. If we first spatially smooth and then ensemble average the simulations from the GFDL and HadCM2 model, the pattern correlation between them rises to 0.64, a value likely to occur less than 5% of the time in an identical analysis of the GFDL control run. This shows that the low wavenumber properties of the ensembles from those two models are rather similar. Although we have only one realization of the observations, we can still spatially smooth it and correlate it with the smoothed ensemble averages from GFDL and HadCM2. The correlations were between 0.21 and 0.23, modestly better agreement between the prediction of the models of the last 50 year temperature trend and that observed than the values noted in section 4.2, but still likely to occur in an identical analysis of the GFDL control run over 20% of the time. Finally, if all 11 realizations are smoothed and then averaged together, they were found to correlate with the smoothed observations at 0.34 (cf. *Wigley et al*. [1998] who estimated such a current value from a perfect model study). This suggests there is good gain to be realized through simple ensemble averaging and that an appropriate ensemble size could well be of the order of 10.

In summary, ensemble averaging and spatial smoothing are prerequisite analyses prior to any attempts to detect and attribute climate change using CGCM results. The ensemble averaging is especially important: If one had only the HC-4 run as the estimated anthropogenic signal, then one would conclude that there was no anthropogenic signal in the atmosphere (smoothed pattern correlation with observations equal to 0.20). By the same token, one could just as easily obtain a single simulation with a large positive correlation with the observations, thereby concluding an anthropogenic signal has been both detected and attributed to human activity (compare Figure 4 and *Knutson et al*. [1999, Figure 1] for impressive examples). Either conclusion, based on a single simulation, is apt to be misleading. This again shows the pressing need to analyze ensembles of scenario runs for detection work and to account for the statistical properties of these finite ensembles, something that the latest studies are beginning to do [cf. *Hegerl et al*., 1999a; *Tett et al*., 1999]. It also highlights the problems inherent in using the single realization of real-world observations which is available to us.

5.2. Model Spread

Clearly, the models do not all produce the same estimate of the anthropogenic signal nor should we expect them to. So it is of interest to see how the range of model estimates of near-surface temperature change compares with observations. In this case we use a display akin to that shown by *Knutson et al*. [1999] to determine where in physical space the models/observations might be in agreement. The GHG + SUL forced model grid point temperature trends between 1945 and 1995 (spatially smoothed), predicted by the 11 model runs, represent a range of possible trend values.

We checked this range with the spatially smoothed, observed trend at each grid point. The regions where the observed trend fell within the extremes of the model trend estimates were left blank in Figure 10. The regions where the observed trend was above/below the model extrema were stippled/hatched, respectively. Of the grid points inspected, 18% had observed trends greater than the largest positive trend predicted by the 11 model group, while 29% had observed trends less than the smallest of the model group.

Over roughly one half of the grid points, the trends from at least some of the model runs bounded the observed trend. This is a reasonably good correspondence given the facts that important forcings are omitted, linear trends used, etc. However, what of the regions where the observed trend was different from any of the model runs? Inspection of Figure 10 shows these to be the regions mostly inhabited by the Pacific Decadal Oscillation [cf. *Latif and Barnett*, 1994; *Barnett et al*., 1999b] and the North Atlantic Oscillation. *Knutson et al*. [1999] found model/observed differences in the same regions for the GFDL model. This suggests that the models are not adequately representing the changes expected in these major climate modes. Whether this is due to the fact the models simply do not represent the climatological character of these modes well or to the fact that they do not capture the anthropogenic impact on the variability of these modes is an open question, one that will be considered elsewhere.

6. Optimal Detection

The results presented above show that the various models' estimates of an anthropogenic signal appear to differ substantially when viewed with conventional analysis methods. However, modern detection schemes first optimize the anthropogenic signal to maximize the signal-to-noise ratio prior to attempting to find the signal in the observations [e.g., *Hasselmann*, 1997; *Hegerl et al*., 1997; *North and Stevens*, 1998]. This transformation can substantially change the appearance of the signal to be detected. Modern detection methods also consider the largest spatial scales by retaining only low-order EOFs, a filtering operation designed to avoid the small-scale noise. Both of these actions appear to avoid many of the concerns noted above for detection and attribution purposes.

To determine directly the impact of intermodel and intramodel differences on detection, we applied the fingerprint detection methods described by *Hegerl et al*. [1997] to each of the 11 GHG and direct sulfate aerosol scenario runs analyzed in section 4, as well as their ensemble averages. The analysis was applied to various runs both including their global means and after the global means had been removed. The results are

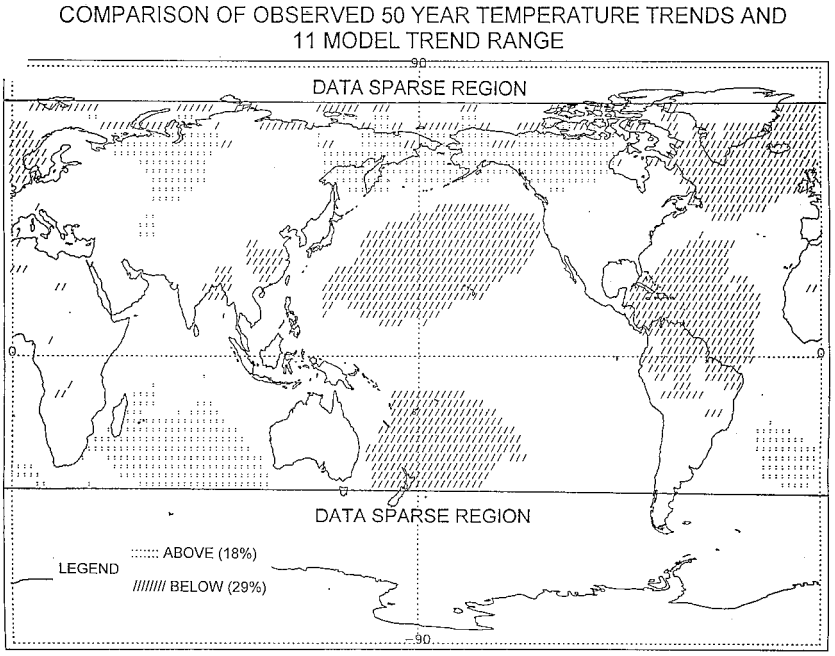

Figure 10. Range test for model trends versus observation. Areas where the observed 1945–1995 trend is within the 11 model trend range are open. Areas where the observed trend is above or below the model range are stippled or hatched, respectively. The percent of grid points where observed trend is above/below the model trend range is given in the legend.

presented in a detection and attribution diagram (see *Hegerl et al*. [1997, 1999] for details) based on the MPI fingerprint from model ECHAM3/LSG and the Hadley Centre fingerprint from model HadCM2 with (Plate 1) and without the global mean (Plate 2). "Fingerprint" here refers to the patterns of anthropogenic change predicted by the various models due to GHG and SUL forcing.

When projected onto the fingerprints, each model run and the observations become single points in the "detection space" defined by the fingerprints. The uncertainty in the position of the observations is represented by an elliptical region (thick line) defining the region of the space where the actual observations are expected to lie with 90% confidence. This uncertainty region is estimated from the levels of natural variability inherent in the HadCM2 model control run which we assumed to be the same as the observations [e.g., *Hegerl et al*., 1997, 1999a]. Single ensemble members are surrounded by a similar range of uncertainty (not shown) as the observations, so even simulations outside the observed uncertainty ellipse can be consistent with the observations. For ensemble averages, the uncertainty range of the average is substantially smaller, since the variance of the Gaussian distribution shown in the ellipse diminishes by a factor of 1 over the ensemble size. Therefore comparing ensemble averages with the observations is a much more rigorous test for the model. Note that each of the three ensembles has different uncertainties due to the differing number of realizations and because the internal variability of each model is different.

Projecting the various GHG + SUL runs onto the MPI fingerprint (Plate 1, top) illustrates at once the behavior of the different model ensembles. The five GFDL runs (pluses) and their ensemble mean (thick pluses) group closely together. The simulation that reproduces the early century warming well (labeled GF-3, Plate 1, top [*Knutson et al*., 1999]) is rather close to the observations. A Hotelling-T^2 test on the difference between individual ensemble members and the observations shows that two of the GFDL ensemble members are not consistent with the observations (shown with gray pluses and denoted GF-2 and GF-5 in Plate 1), while the other three simulations are consistent with observations (shown with pink pluses). "Consistency" in this case means the observed trend patterns and those produced by the model forced by a combination of GHG and direct sulfate aerosols are statistically indistinguishable.

The realizations from the Hadley Centre are more scattered in the detection space. Three of the runs (blue, circle) and the ensemble mean (HadC, thick circle) are consistent with the observations, while HC-4 is not (gray, circle). By itself, HC-4, lying below the zero line of the orthogonal sulfate patterns signal strength, implies a Northern Hemisphere warming due to direct sulfate effects, just the opposite of what is expected. In other words, an event of internal model variability overwhelms the effect of sulfate forcing in this case. Both of the MPI runs (red, cross) and the ensemble mean (H3L, thick cross) are consistent with the data.

The average of all of the 11 runs ("all," thick black circles) falls well inside the observed uncertainty. This means the observed and grand ensemble averaged near-surface air temperature trends over the last 50 years from the models are consistent with each other in the detection space used here. Exclusion of the global mean trend (Plate 2, top) gives much the same story as above except that all GFDL simulations are now consistent with observations. Note that either with (Plate 1, top) or without (Plate 2, top) the spatial mean, the ellipses do not include the origin. This means both greenhouse and sulfate signals have been detected, subject to the caveats given below.

The same scenario runs were next projected onto the Hadley Centre fingerprint, a different measure of the expected re-

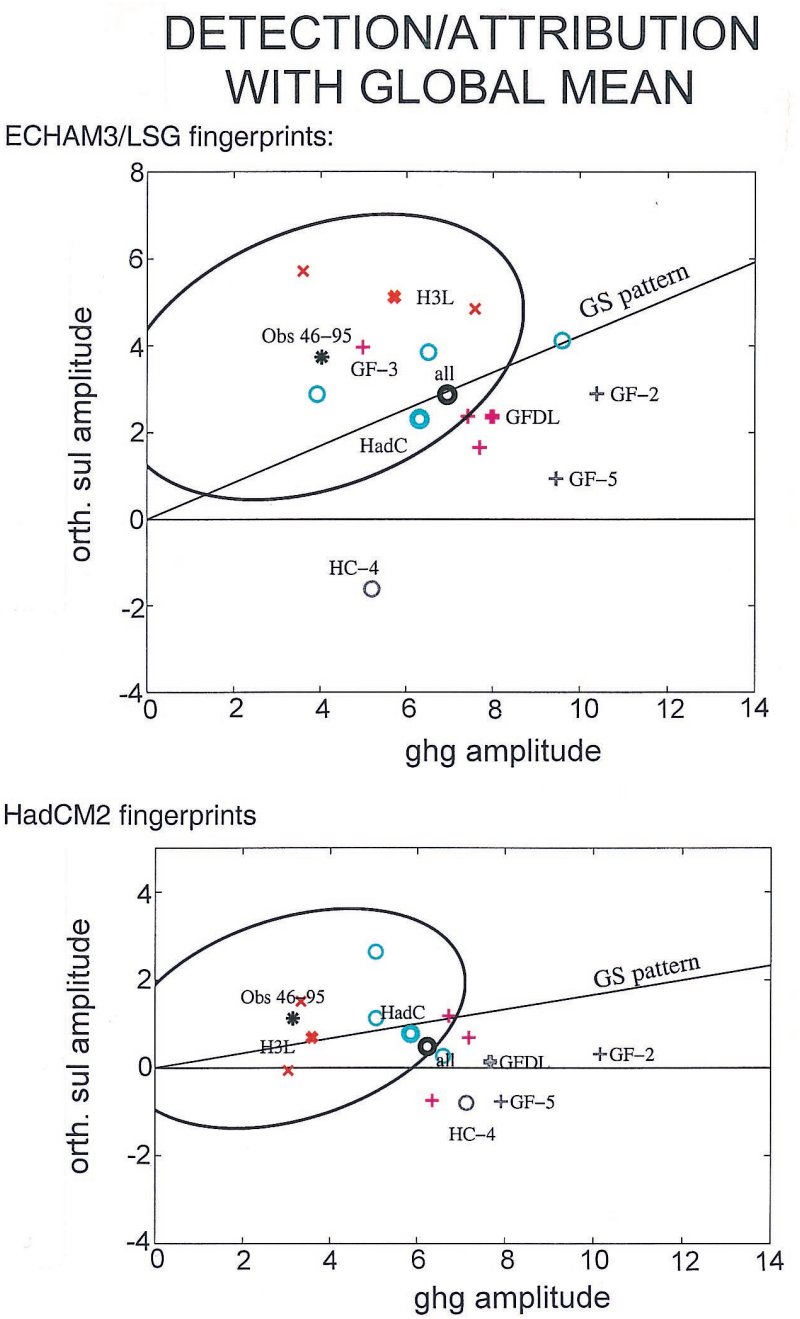

Plate 1. Detection diagram based on MPI (top) and HadCM2 (bottom) anthropogenic fingerprints (see *Hegerl et al*. [1997] for details of figure construction). Horizontal axis gives the amplitude of the GHG signal. Vertical axis gives the amplitude of the SUL signal that is orthogonal to the GHG signal. The five GFDL runs (pink) are noted with pluses and their ensemble mean with a boldface plus and denoted GFDL. The four HadCM2 runs (blue) are noted with circles, their ensemble mean with a boldface circle and denoted HadC. The two MPI runs (red) are noted with crosses and their ensemble mean with a boldface cross and denoted H3L. The average of all 11 runs is denoted by a thick black circle and "all." The current observed climate trend 1945–1995 is given by a boldface asterisk and denoted "obs." The ellipse represents the 90% confidence limits on the current observed state. The near-global mean trend is included in this analysis. Points "consistent" with the observations according to a Hotelling-T^2 test are colored. Runs not consistent with the observations at the 10% level (ensemble average, GF-2, GF-5 from GFDL and HC-4 from HadCM2) have gray symbols. The line marked "GS" is the one-dimensional projection of the GHG + SUL signal (see *Hegerl et al*. [1997] for details).

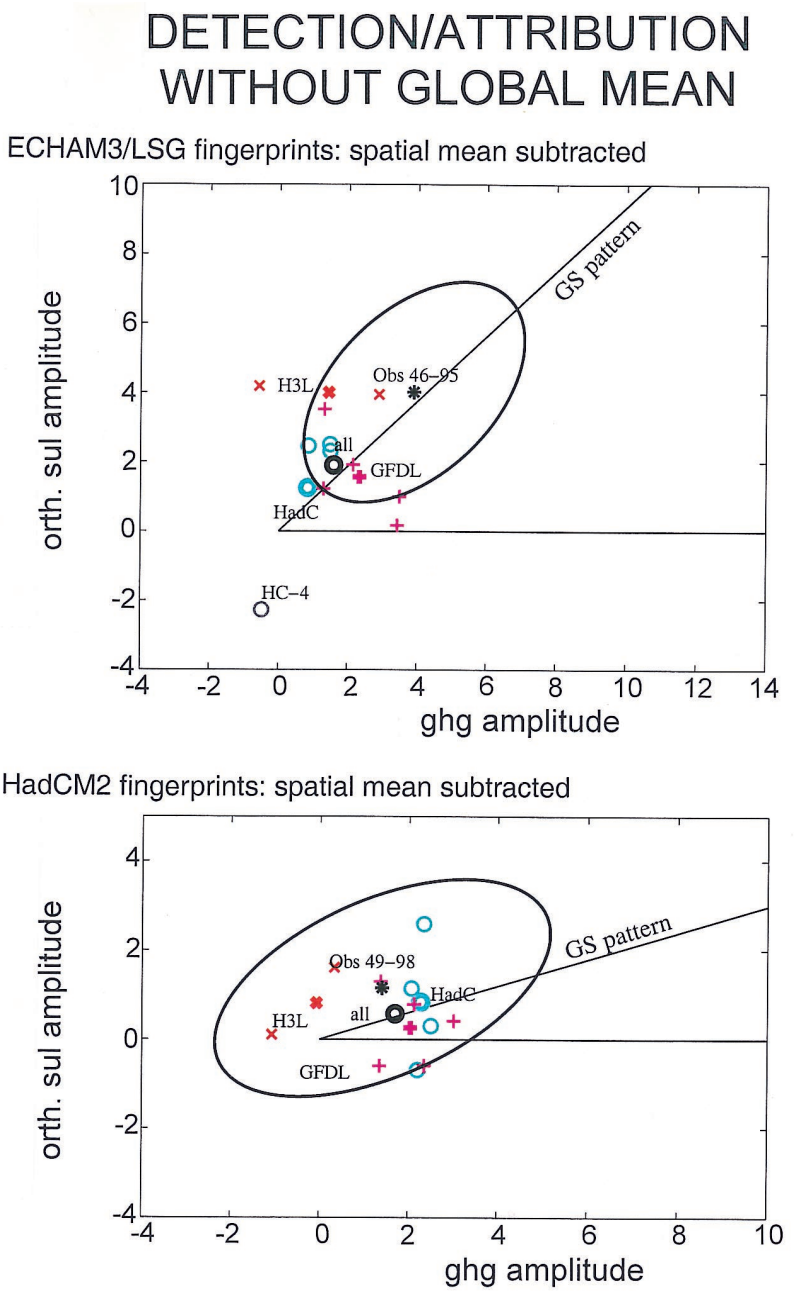

Plate 2. Same as Plate 1 except the near-global mean trend has been removed prior to analysis.

sponse to GHG and SUL forcing Plate 1, bottom (see *Hegerl et al*. [1999a] for a more complete discussion). In this case, the interpretation is rather different. The ensemble mean of the GFDL runs in this case is not consistent with the observations, but three out of five of its realizations are consistent. The Hadley Centre realizations and their mean are consistent with the observations except for the previously discussed HC-4 simulation. The MPI runs and their mean (H3L) are all consistent with the observations. Note that the observational uncertainty ellipse includes the origin. This means the observed changes in climate, in this coordinate system, could be due also to natural variability. However, an anthropogenic signal can be detected beyond that expected from natural variability if the test is conducted in a one-dimensional space defined by the total anthropogenic signal (cf. *Hegerl et al*. [1999a], and *Barnett et al*. [1999a] for details). This latter type of test makes it impossible to estimate the relative roles of GHG and SUL in producing the observed signal.

Removing the global mean from the realizations and then projecting the results onto the Hadley fingerprint gives rather different results (Plate 2, bottom). Remarkably, all of the 11 realizations and their ensemble means fall on or within the confidence ellipse of the observations, seemingly a resounding case for detection and attribution (D/A) of an anthropogenic signal. Unfortunately, the confidence ellipse for the observations still includes the origin, so the observed climate change could still be due to natural variability. Put another way, we cannot distinguish between the anthropogenic and the natural forcing mechanisms, and so we cannot attribute the climate change to a specific cause if we disregard the global mean in the Hadley fingerprint. However, *Stott et al*. [1999] and *Tett et al*. [1999] detect a HadCM2 greenhouse gas and a sulfate signal, with and without spatial mean, if both spatial and temporal patterns were used in a fingerprint approach. This suggests that using information about the time evolution of anthropogenic signals beyond simple linear trends, as we do here, enhances the prospects for detecting such signals if they are present in the observations.

In summary, the filtering and ensemble averaging techniques used in D/A studies are largely adequate to suppress the types of noise and uncertainty discussed above. The results show the need to use ensembles of scenario runs in D/A work to avoid being fooled by an "outlier" simulation, for example, the single realization (gray circle, HC-4) lying below the abscissa in Plate 1. It is also clear that leaving the global mean in the analysis or removing it can make a difference in the answer one obtains (compare Plates 1 and 2 [*Hegerl et al*., 1999b]). However, the average of all simulations was consistent with the observations in either case. Finally, while many features and results derived from two different estimates of the expected fingerprints agree, some do not. This, in turn, suggests we cannot rely on just one model estimate of expected anthropogenic change for detection work, a result already seen in sections 3 and 4.

7. Conclusions

This study has examined the relative differences between different model estimates of anthropogenically forced signals. Our main conclusions are as follows:

1. Nine models forced by CO_2 increasing at the rate of 1%/yr produce near-global mean near-surface temperature signals at the end of an 80 year integration which agree to within about 25% of the nine model average.

We next analyzed 11 different runs from three different CGCMs, each forced by observed/projected GHG and direct sulfate aerosol effects. Concentrating on the trend of near-surface temperature change over approximately the last 50 years (1945–1995), we found the following:

2. The individual, unsmoothed model simulations bore only weak similarity to each other and to the observations.

3. One cause of the above result was the small-scale spatial noise that characterized all the model runs. Spatial smoothing improved the magnitude of statistically significant agreement between individual model runs and also with data, a result anticipated by *Stott and Tett* [1998].

4. A major reason for the apparent dissimilarity between the individual runs was due to internal model variability. Ensemble averaging the runs produced patterns that were more similar between models and observations than noted (item 3) above. Intermodel ensembles produced even better results. The existence of large differences between members of a single CGCM ensemble almost requires that detection and attribution work be done with ensembles of scenario runs, for single runs cannot represent the range of possible results due to internal variability [see also *Tett et al*., 1996, 1999; *Stott et al*., 1999; *Hegerl et al*., 1999a]. This also raises the issue of whether it is better to go for lower-resolution runs with larger ensembles or higher resolution and smaller ensembles.

5. Recent detection and attribution methods, coupled with ensemble averaging methods, produced a reasonably consistent match between model predictions of expected temperature trends due to a combination of GHG and direct sulfate aerosols and those observed [see also *Tett et al*., 1999; *Hegerl et al*., 1999a]. The reader should weigh this statement carefully, for the runs we studied do not include many of the anthropogenic pollutants thought to be important (e.g., indirect sulfate aerosol effects, tropospheric ozone) nor do they have additional natural forcing by volcanoes, solar changes, etc. Recent results show that these latter natural forcings cannot by themselves fully explain recent climate change signals [cf. *Cubasch et al*., 1997; *Tett et al*., 1999; *Stott et al*., 1999; *Hegerl et al*., 1997, 1999a; *North and Stevens*, 1999.]

6. Finally, our results demonstrate the need to use different estimates of the anthropogenic fingerprint in detection studies. Different models give different estimates of these fingerprints, and we do not currently know which is most correct. Further, the intramodel uncertainty in both the fingerprints and the scenario runs can be relatively large. In short, simulation, detection, and attribution of an anthropogenic signal is a job requiring inputs from different, high-quality models.

Appendix

The common EOFs (cEOFs) of the CMIP2 models were computed from the last 70 years of the nine model data. For each model grid point, the long-term annual mean from its respective control run was removed first. The nine fields were then concatenated and the EOFs computed from the associated covariance matrix.

The main feature accounting for the substantial energy in cEOF1 is (Figure A1, 45.8% of total variance) clearly the land-sea contrast in temperature change expected with anthropogenic warming [cf. *North and Stevens*, 1998]. Also notable is the relatively strong warming at high latitudes. All models seem to produce these features, although as shown in Figure 3 there is considerable variation in its relative strength between models.

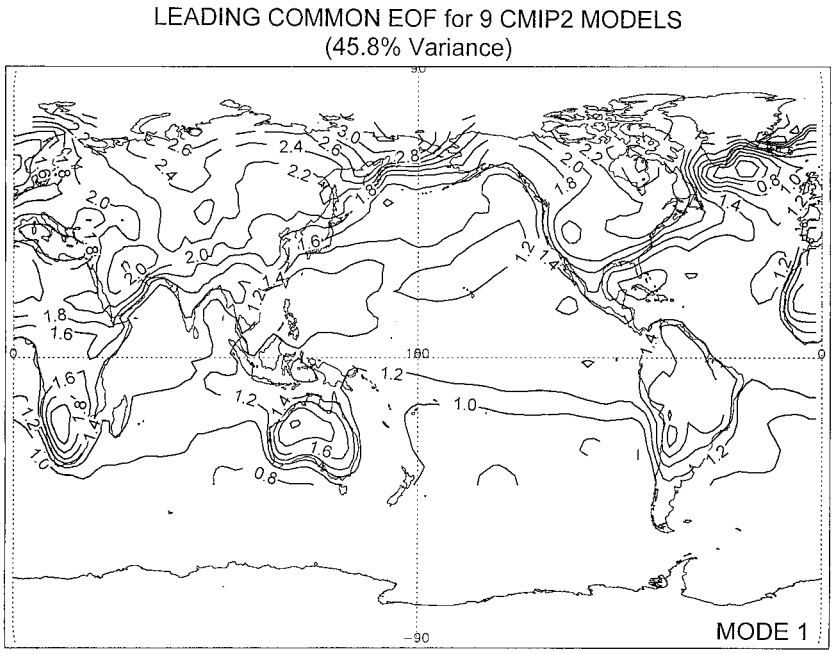

Figure A1. First common empirical orthogonal function of the nine CMIP2 model ensemble. This mode captured 45.8% of the joint model variance.

The small-scale details of the spatial patterns of surface temperature change produced by the various models in response to a 1% CO_2 increase are only moderately similar (Table 3). This result, seen in section 3.2, can be explained partially by the presence of high wave number noise in the patterns (in spite of the decadal time averaging) since the second- and higher-order modes are, for all intents and purposes, statistically degenerate. The leading PC in Figure A2 contains little of this noise since it is dominated by trends. The cEOF analysis focuses on the largest scales of variability, the same ones that detection schemes rely on [e.g., *Allen and Tett*, 1999], and so effectively, low-pass filters this noise for mode 1.

The eigenvalue spectrum (Figure A3), together with Figure A2, illustrates that even moderate dissimilarities on the largest scales might have an impact on detection results. The partial eigenvalues of the leading mode, often taken in detection stud-

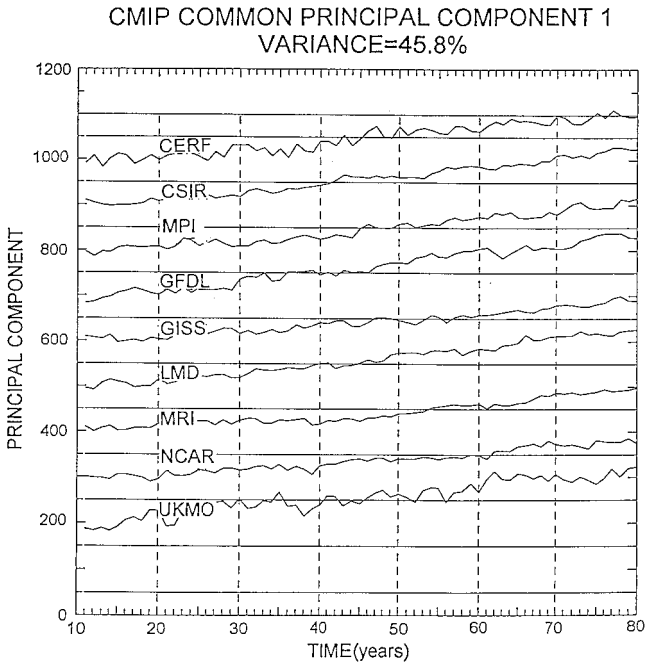

Figure A2. Leading principal component that goes with common EOF 1 shown in Figure A1. The model identifications are given in Table 1. The ordinate is offset in increments of 100 units to separate the various models.

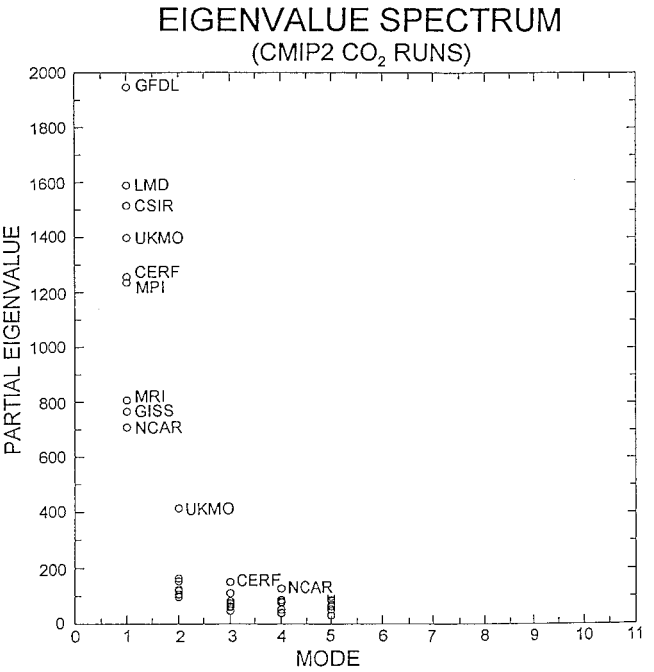

Figure A3. Eigenvalue spectrum from the common EOF analysis of the 1% per year CO_2 ensemble (units of $°C^2$).

ies as the "signal to be detected," vary between models approximately in the range 1950 to 700 variance units (degrees C^2), with a mean of about 1250°C^2. The ratio of signal amplitudes (square roots of the variance) to the mean signal ranges from 1.23 to 0.56. The GFDL simulation gives the largest response to the 1%/yr CO_2 forcing and will tend to dominate the results of the cEOF analysis, while the NCAR simulation gives the weakest response and will be least well represented in the analysis. Without further information we conclude that the simplest anthropogenic scenario (increasing CO_2 forcing case), as represented by the leading eigenmode, is in agreement between different model simulations to within a scaling factor ranging from 0.56 to 1.23. The detection scheme discussed by *Stott and Tett* [1998] estimates a scaling factor that can be directly compared with this range.

Acknowledgments. This work was made possible by the NOAA Office of Global Programs and the Department of Energy Office of Biological and Environmental Research in conjunction with the Climate Change Data and Detection element (a part of the U.S. Global Change Research Program). The Scripps Institution of Oceanography provided partial support for TPB. GCH was supported by the Alexander von Humboldt-Stiftung, JISAO of the University of Washington, and NSF (ATM-9707069). The model data were provided as follows: The HadCM2 runs have been supplied by the Climate Impacts LINK project (Department of Environment contract EPC 1/1/16) on behalf of the Hadley Center and U.K. Meteorological Office. SFBT and computer time to carry out the HadCM2 simulations was supported by U.K. DETR under contract PECD 7/12/372. The MPI runs were provided by the Germany Climate Computing Center (DKRZ), courtesy of L. Bengtsson and E. Roeckner. The Geophysical Fluid Dynamics Laboratory runs were provided by Tom Delworth. The Global Sea-Ice and Sea Surface Temperature data set comes from work of N. A. Rayner, E. B. Horton, and D. E. Parker via Phil Jones and the UEA Climatic Research Unit. Thanks are due to Karl Taylor for critical suggestions that improved the manuscript and for his early collaborative work on the CMIP2 runs. The authors also appreciate the efforts of three anonymous reviewers whose extensive comments substantially increased the quality of the paper.

References

Allen, M. R., and S. F. B. Tett, Checking for model consistency in optimal fingerprinting, *Clim. Dyn.*, *15*, 419–434, 1999.

Barnett, T. P., Monte Carlo climate forecasts, *J. Clim.*, *8*(5), 1005–1022, 1995.

Barnett, T. P., Comparison of near surface air temperature variability in eleven coupled global climate models, *J. Clim.*, *12*, 511–518, 1999.

Barnett, T. P., and R. W. Preisendorfer, Origins and levels of monthly and seasonal forecast skill for United States air temperatures determined by canonical correlation analysis, *Mon. Weather Rev.*, *115*, 1825–1850, 1987.

Barnett, T. P., M. E. Schlesinger, and X. Jiang, On greenhouse gas detection strategies, in *Greenhouse-Gas-Induced-Climatic Change: A Critical Appraisal of Simulations and Observations*, edited by M. E. Schlesinger, pp. 537–558, Elsevier Sci., New York, 1991.

Barnett, T. P., G. Hegerl, B. Santer, and K. Taylor, The potential effect of GCM uncertainties on greenhouse signal detection, *J. Clim.*, *11*(4), 659–675, 1998.

Barnett, T. P., et al., Detection and attribution of recent climate change: A status report, *Bull. Am. Meteorol. Soc.*, *80*(12), 2631–2659, 1999a.

Barnett, T. P., D. W. Pierce, R. Saravanan, N. Schneider, D. Dommenget, and M. Latif, Origins of the midlatitude Pacific decadal variability, *Geophys. Res. Lett.*, *26*(10), 1453–1456, 1999b.

Cess, R. D., et al., Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models, *J. Geophys. Res.*, *95*, 16,601–615, 1990.

Cubasch, U., B. D. Santer, A. Hellbach, G. C. Hegerl, H. Hock, H. Maier-Reimer, U. Mikolajewicz, A. Stossel, and R. Voss, Monte Carlo climate forecasts with a global coupled ocean-atmosphere model, *Clim. Dyn.*, *10*, 1–19, 1994.

Cubasch, U., G. C. Hegerl, R. Voss, J. Waszkewitz, and T. J. Crowley, Simulation with an O-AGCM of the influence of variations of the solar constant on the global climate, *Clim. Dyn.*, *13*, 757–767, 1997.

Gates, W. L., AMIP: The atmospheric model intercomparison project, *Bull. Am. Meteorol. Soc.*, *73*(12), 1962–1970, 1992.

Gates, W., et al., An overview of the results of the Atmospheric Model Intercomparison Project (AMIP), *Bull. Am. Meteorol. Soc.*, *80*(1), 29–55, 1999.

Hasselmann, K., Multi-pattern fingerprint method for detection and attribution of climate change, *Clim. Dyn.*, *13*(9), 601–611, 1997.

Hegerl, G. C., H. von Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, Detecting anthropogenic climate change with an optimal fingerprint method, *J. Clim.*, *9*, 2281–2306, 1996.

Hegerl, G. C., K. Hasselmann, U. Cubasch, J. F. B. Mitchell, E. Roeckner, R. Voss, and J. Waszkewitz, On multi-fingerprint detection and attribution of greenhouse gas and aerosol forced climate change, *Clim. Dyn.*, *13*, 613–634, 1997.

Hegerl, G. C., P. Stott, M. Allen, J. F. B. Mitchell, S. F. B. Tett, and U. Cubasch, Optimal detection and attribution of climate change: Sensitivity of results to climate model differences, *Clim. Dyn.*, in press, 1999a.

Hegerl, G. C., P. D. Jones, and T. P. Barnett, Effect of sampling uncertainty on anthropogenic signal detection, *J. Clim.*, in press, 1999b.

Johns, T. C., R. E. Carnell, J. F. Crossley, J. M. Gregory, J. F. B. Mitchell, C. A. Senior, S. F. B. Tett, and R. A. Wood, The second Hadley Centre coupled ocean-atmosphere GCM: Model description, spinup, and validation, *Clim. Dyn.*, *13*, 103–134, 1997.

Jones, P. D., Hemispheric surface air temperature variations: A reanalysis and an update to 1993, *J. Clim.*, *7*, 1794–1802, 1994.

Knutson, T. R., T. L. Delworth, K. Dixon, and R. J. Stouffer, Model assessment of regional surface temperature trends (1949–1997), *J. Geophys. Res.*, *104*, 30,981–30,996, 1999.

Latif, M., and T. P. Barnett, Causes of decadal climate variability over the North Pacific/North American sector, *Science*, *266*, 634–637, 1994.

Meehl, G. A., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer, Intercomparison makes for a better climate model, *Eos*, *78*, 445–446, 451, 1997.

Mitchell, J. F. B., T. C. Johns, J. M. Gregory, and S. F. B. Tett, Climate response to increasing levels of greenhouse gases and sulfate aerosols, *Nature*, *376*, 501–504, 1995.

Nicholls, N., G. V. Gruza, J. Jouzel, T. R. Karl, L. A. Ogallo, and D. E. Parker, Observed climate variability and change, in *Climate Change 1995: The IPCC Second Assessment*, edited by J. T. Houghton et al., pp. 133–192, Cambridge Univ. Press, New York, 1996.

North, G. R., and K. K. Kim, Detection of forced climate signals, part II, Simulation results, *J. Clim.*, *8*, 409–417, 1995.

North, G. R., and M. J. Stevens, Detecting climate signals in the surface temperature record, *J. Clim.*, *11*, 563–577, 1998.

North, G. R., K. K. Kim, S. S. P. Shen, and J. W. Hardin, Detection of forced climate signals, part I, Filter theory, *J. Clim.*, *8*, 401–408, 1995.

Palmer, T. N., R. Buizza, F. Molteni, Y. Q. Chen, and S. Corti, Singular vectors and the predictability of weather and climate, *Phil. Trans. R. Meteorol. Soc. A*, *348*, 459–475, 1994.

Parker, D. E., P. D. Jones, A. Bevan, and C. K. Folland, Interdecadal changes of surface temperature since the late 19th century, *J. Geophys. Res.*, *99*, 14,373–14,399, 1994.

Parker, D. E., C. K. Folland, and M. Jackson, Marine surface temperature: Observed variations and data requirements, *Clim. Change*, *31*, 559–600, 1995.

Pierce, D., T. Barnett, and U. Mikolajewicz, On the competing roles of heat and fresh water flux in forcing thermohaline oscillations, *J. Phys. Oceanogr.*, *25*, 2046–2064, 1995.

Räisänen, J., Internal variability as a cause of qualitative intermodel disagreement on anthropogenic climate changes, *Theor. Appl. Clim.*, *64*, 1–13, 1999.

Richman, M. B., Rotation of principal component, *J. Clim.*, *6*, 293–335, 1986.

Santer, B. D., T. M. L. Wigley, and P. D. Jones, Correlation methods in fingerprint detection studies, *Clim. Dyn.*, *8*, 265–276, 1993.

Santer, B. D., K. E. Taylor, T. M. L. Wigley, J. E. Penner, P. D. Jones, and U. Cubasch, Towards the detection and attribution of an anthropogenic effect on climate, *Clim. Dyn.*, *12*(2), 77–100, 1995a.

Santer, B. D., U. Mikolajewicz, W. Brueggemann, U. Cubasch, K. Hasselmann, H. Hoeck, E. Maier-Reimer, and T. M. L. Wigley,

Ocean variability and its influence on the detectability of greenhouse warming signals, *J. Geophys. Res*., *100*, 10,693–10,725, 1995b.

Santer, B. D., et al., Human effect on global climate? (Letter), *Nature*, *384*, 524, 1996.

Santer, B. D., T. M. L. Wigley, T. P. Barnett, and E. Anyamba, Detection of climate change and attribution of causes, in *Climate Change 1995: The IPCC Second Scientific Assessment*, edited by J. T. Houghton and B. A. Callander, Cambridge Univ. Press, New York, in press, 1996.

Senior, C. A., and J. F. B. Mitchell, Carbon dioxide and climate—The impact of cloud parameterization, *J. Clim*., *6*(3), 393–418, 1993.

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

Stott, P. A., S. F. B. Tett, G. S. Jones, M. R. Allen, W. J. Ingram, and J. F. B. Mitchell, Attribution of twentieth century temperature change to natural and anthropogenic causes, *Clim. Dyn*., in press, 1999.

Stouffer, R. J., G. C. Hegerl, and S. F. B. Tett, A comparison of surface air temperature variability in three 1000-year coupled ocean-atmosphere model integrations, *Clim. Dyn*., in press, 1999.

Tett, S. F. B., J. F. B. Mitchell, D. E. Parker, and M. R. Allen, Human influence on the atmospheric vertical temperature structure: Detection and observations, *Science*, *274*, 1170–1173, 1996.

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

Trenberth, K. E., and J. W. Hurrell, Decadal atmosphere-ocean variations in the Pacific, *Clim. Dyn*., *9*, 303–319, 1994.

von Storch, J.-S., Interdecadal variability in a global coupled model, *Tellus*, *Ser. A*, *46*, 419–432, 1994.

Wigley, T. M. L., P. J. Jaumann, B. D. Santer, and K. E. Taylor, Relative detectability of greenhouse-gas and aerosol climate change signals, *Clim. Dyn*., *14*(11), 781–790, 1998.

T. Barnett, Scripps Institution of Oceanography, MC 0224, La Jolla, CA 92037. (tbarnett@ucsd.edu)

G. Hegerl, Texas A&M University, College Station, TX 77843.

T. Knutson, Geophysical Fluid Dynamics Laboratory, Princeton, NJ 08542.

S. Tett, Hadley Centre for Climate Prediction and Research, Meteorological Office, Bracknell, RG12 2SY UK.

(Received July 20, 1999; revised January 13, 2000;
accepted March 3, 2000.)