Supporting Information

1 Observational data

Observational data for W, the column-integrated atmospheric moisture content over oceans, were provided by Remote Sensing Systems (RSS) in Santa Rosa, California (S1, S2). All analyses reported on here rely on version 6.6 of the SSM/I-derived W dataset produced by RSS. Data were available as monthly means on a $2.5° \times 2.5°$ latitude/longitude grid, and span the period July 1987 through December 2008.

We used version 3 of the NOAA Extended Reconstructed Sea Surface Temperature dataset (ERSST) (S3) for the SST-based model quality metrics. ERSST data were available from January 1854 to December 2006 in the form of monthly means on a regular $2° \times 2°$ latitude/longitude grid. Reconstruction of high-frequency SST anomalies involved use of empirically-derived spatial modes of variability to interpolate observations in times of sparse coverage. Further details of the ERSST dataset are available online (S4).

2 Modeling groups contributing to IPCC database

At the time this research was conducted, 15 modeling groups had performed a wide range of simulations in support of the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR4). Climate data from these simulations were made available to the scientific community through the U.S. Dept. of Energy's Program for Climate Model Diagnosis and Intercomparion (PCMDI). Six modeling groups provided column-integrated water vapor and SST results for at least two different model configurations. Results from a total of 22 different climate models were analyzed.

We considered two sets of simulations here: pre-industrial control runs, and 20CEN experiments with historical changes in a number of different anthropogenic and natural forcings. In IPCC terminology, these integrations are referred to as "picntrl" and "20c3m" (respectively).

Official designations of the 15 modeling groups that supplied W data are listed below (with model acronyms in brackets):

1. Bjerknes Center for Climate Research, Norway [BCCR-BCM2.0].
2. Canadian Centre for Climate Modelling and Analysis, Canada [CCCma-CGCM3.1(T47) and CCCma-CGCM3.1(T63)].
3. National Center for Atmospheric Research, U.S.A. [CCSM3 and PCM].

4. Météo-France/Centre National de Recherches Météorologiques, France [CNRM-CM3].

5. Commonwealth Scientific and Industrial Research Organization (CSIRO) Atmospheric Research, Australia [CSIRO-Mk3.0].

6. Max-Planck Institute for Meteorology, Germany [ECHAM5/MPI-OM].

7. Meteorological Institute of the University of Bonn, Meteorological Research Institute of the Korean Meteorological Agency, and Model and Data group, Germany/Korea [MIUB/ECHO-G].

8. Institute for Atmospheric Physics, China [FGOALS-g1.0].

9. Geophysical Fluid Dynamics Laboratory, U.S.A. [GFDL-CM2.0 and GFDL-CM2.1].

10. Goddard Institute for Space Studies, U.S.A. [GISS-AOM, GISS-EH, and GISS-ER].

11. Institute for Numerical Mathematics, Russia [INM-CM3.0].

12. Institute Pierre Simon Laplace, France [IPSL-CM4].

13. Center for Climate System Research, National Institute for Environmental Studies, and Frontier Research Center for Global Change, Japan [MIROC-CGCM2.3.2(medres) and MIROC-CGCM2.3.2(hires)].

14. Meteorological Research Institute, Japan [MRI-CGCM2.3.2].

15. Hadley Centre for Climate Prediction and Research, U.K. [UKMO-HadCM3 and UKMO-HadGEM1].

3 Forcings used in 20CEN runs

Details of the natural and anthropogenic forcings used by differing modeling groups in their IPCC 20CEN simulations are given in Table S1. This Table was compiled using information that participating modeling centers provided to PCMDI.[1] All model acronyms used in the Table are defined in the previous Section.

A total of 11 different forcings are listed in Table S1. A letter 'Y' denotes inclusion of a specific forcing. As used here, 'inclusion' signifies the specification of time-varying forcings, with changes on interannual and longer timescales. Forcings that were varied over the annual cycle only, or not at all, are identified with a dash. A question mark indicates a case where there is uncertainty regarding inclusion of the forcing.

Results in Table S1 are stratified by inclusion or omission of volcanic forcing (V or No-V, respectively). Ten of the twelve V models explicitly incorporated volcanic aerosols. Two V models – MRI-CGCM2.3.2 and MIUB/ECHO-G – represented volcanic effects in a more indirect manner, using estimated volcanic forcing data from (S5) and (S6) (respectively) to adjust the solar irradiance at the top of the model atmosphere. The V versus No-V partitioning also separates models with 'total' external forcing (natural plus anthropogenic) from models with primarily anthropogenic forcing.

[1]See http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php.

While all 15 modeling groups used very similar changes in well-mixed greenhouse gases, the changes in other forcings were not prescribed as part of the experimental design. In practice, each group employed different combinations of 20th century forcings, and often used different datasets for specifying individual forcings. End dates for the experiment varied among groups, and ranged from 1999 to 2003.

4 Calculation of Model Quality Metrics

Model quality metrics were calculated after regridding model water vapor and SST data to the target grids of the SSM/I and ERSST observational data. This regridding step also involves masking (see Section 5.1).

Comparisons between modeled and observed quantities are based on model data from the 20CEN experiments rather than the pre-industrial control runs. This choice was made because model-versus-observed variability comparisons can be influenced by the neglect of historical changes in external forcings, particularly the volcanic aerosol forcing. Such forcing changes are included in many of the 20CEN simulations, but are not incorporated in the control runs.[2]

[2]Our results suggest that the inclusion of combined natural and anthropogenic external forcings in 20CEN runs leads to closer agreement with observed water vapor and SST data (relative to the agreement obtained in 20CEN runs with anthropogenic forcing only). In all six sets of rankings shown in Fig. 4, at least 7 of the top 10 models included volcanic forcing. These results are

In the following, we provide a brief introduction to the statistical notation used in the discussion of metrics.

4.1 Statistical notation

Subscripts

m	Subscript denoting model data
o	Subscript denoting observational data
w	Subscript denoting metric computed with water vapor data
T	Subscript denoting metric computed with SST data

Indices

i	Index over number of 20CEN realizations for j^{th} model
j	Index over number of models
k	Index over number of regions
l	Index over number of timescales in variability analysis
x	Index over number of grid-points
t	Index over time (months or years)

---

consistent with previous work that has demonstrated the existence of volcanically-induced signals in the temporal variability of ocean heat content, SSTs, and precipitation (S7–S9). It is more difficult to interpret why the inclusion of volcanic forcing information appears to enhance model performance in simulating the observed mean state and annual cycle (Fig. 4A). It is unclear whether this is a real physical effect, or a reflection of systematic differences in the quality of the V and No-V models.

Summation limits

$N_r(j)$	Number of 20CEN realizations for j^{th} model (varies from 1-9)
N_m	Number of models (22)
N_k	Number of regions (5)
N_l	Number of timescales for variability analysis (3 for SST, 2 for W)
$N_x(k)$	Number of grid-points (varies with region k)
$N_p(k)$	Number of grid-points $\times 12$ ($N_p(k) = N_x(k) \times 12$)
N_t	Number of time points (months or years)

Averaging notation

$<>$	Spatial average
$\overline{}$	Average over 20CEN realizations (single overbar)
$\overline{\overline{}}$	Average over 20CEN realizations and models (double overbar)
$\widehat{}$	Average over statistics (hat)
$\bullet$	Average over time

Anomalies

$'$	Monthly-mean anomalies w.r.t. climatological annual means (prime)
$''$	Monthly-mean anomalies w.r.t. climatological monthly means (double prime)

Metrics

α	Bias metric
β	Annual cycle metric
ϕ	Variability amplitude metric

φ	Variability pattern metric
$\widehat{\alpha}$	Average of 10 different bias metrics
$\widehat{\beta}$	Average of 10 different annual cycle metrics
$\widehat{\phi}$	Average of 25 different variability amplitude metrics
$\widehat{\varphi}$	Average of 25 different variability pattern metrics
$\widehat{Q}_1$	Average of 20 different mean state and annual cycle metrics
$\widehat{Q}_2$	Average of 50 different variability amplitude and variability pattern metrics
$\widehat{Q}_3$	Average of 70 different mean state, annual cycle, and variability metrics

4.2 Mean State Metrics

As described in the main text, we calculated 10 mean state metrics (two variables × five regions). Each mean state metric is a normalized measure of the absolute value of the model bias. We refer to these bias metrics subsequently as $\alpha_{_W}$ (for water vapor) and $\alpha_{_T}$ (for SST). Here, we limit the discussion to $\alpha_{_W}$, and note that α_T is calculated in an analogous way.

For the i^{th} 20CEN realization of the j^{th} model, the absolute bias in water vapor is defined as:

$$\delta_{_W}(i,j,k) \;=\; | <W_m(i,j,k)> - <W_o(k)> | \qquad (1)$$

$$i \;=\; 1,\ldots,N_r(j);\; j=1,\ldots,N_m;\; k=1,\ldots,N_k$$

where $< W_m(i,j,k) >$ and $< W_o(k) >$ are values of modeled and observed climatological annual mean water vapor, spatially-averaged over the k^{th} region. The period used for calculating climatological annual means is January 1988 to December 1999 for water vapor[3] and January 1961 to December 1990 for SST.[4] We then compute (for each model for which multiple 20CEN realizations are available) the ensemble-mean absolute bias:

$$\overline{\delta_{_W}}(j,k) = \frac{1}{N_r(j)} \sum_{i=1}^{N_r(j)} \delta_{_W}(i,j,k) \qquad (2)$$

$$j = 1,\ldots,N_m;\ k = 1,\ldots,N_k$$

The multi-model average bias, $\overline{\overline{\delta_{_W}}}(k)$, is defined as:

$$\overline{\overline{\delta_{_W}}}(k) = \frac{1}{N_m} \sum_{j=1}^{N_m} \overline{\delta_{_W}}(j,k) \qquad (3)$$

$$k = 1,\ldots,N_k$$

The inter-model standard deviation of the ensemble-mean absolute bias is given by:

$$s\{\overline{\delta_{_W}}(k)\} = \left[\frac{1}{N_m - 1} \sum_{j=1}^{N_m} \left(\overline{\delta_{_W}}(j,k) - \overline{\overline{\delta_{_W}}}(k) \right)^2 \right]^{1/2} \qquad (4)$$

$$k = 1,\ldots,N_k$$

[3]This is the period of maximum overlap between the 20CEN simulations and the SSM/I observational dataset.

[4]This is a frequently-used observational reference period, and a time of relatively stable observational coverage.

Next, we normalize the absolute bias for the j^{th} model and k^{th} region by the inter-model standard deviation of the bias:

$$\alpha_{_W}(j,k) \;=\; \overline{\delta_{_W}}(j,k) \;/\; s\{\overline{\delta_{_W}}(k)\} \qquad (5)$$

$$j \;=\; 1,\ldots,N_m;\; k=1,\ldots,N_k$$

This normalization step enables us to combine information from two different climate variables (water vapor and SSTs), and will later allow us to combine different types of statistical information (on model performance in simulating the mean state, annual cycle, and variability).

Finally, for each model, we compute the average normalized bias over the five regions and two variables:

$$\widehat{\alpha}(j) \;=\; \frac{1}{N_k \times 2}\left[\sum_{k=1}^{N_k}\alpha_{_W}(j,k) \;+\; \sum_{k=1}^{N_k}\alpha_{_T}(j,k)\right] \qquad (6)$$

$$j \;=\; 1,\ldots,N_m$$

where the $\widehat{\;}$ indicates an average over statistics. By definition, $\widehat{\alpha}(j)$ does not provide information about the direction of model biases.

4.3 Annual Cycle Metrics

As in the case of the mean state, there are 10 annual cycle metrics, one for each variable (W and SST) and region. At each grid-point in the k^{th} region, climatolog-

ical monthly means are computed over the same time periods used to estimate the absolute biases (*i.e.*, over 1988-1999 for W and 1961-1990 for SST). This yields 12 climatological monthly-mean patterns. The climatological annual-mean pattern is subtracted from each of the 12 climatological monthly-mean patterns, and the resulting 12 anomaly fields are then concatenated. To simplify the notation, the model and observed concatenated anomaly patterns are represented by $W_m^{'}(i,j,k,x)$ and $W_o^{'}(k,x)$, where the prime denotes monthly anomalies relative to the climatological annual mean, the index x runs from 1 to $N_p(k)$, and $N_p(k) = N_x(k) \times 12$, where $N_x(k)$ is the number of grid-points for the k^{th} region.

Next, we compute the pattern correlation $r_{_W}(i,j,k)$ between the modeled and observed anomaly fields:

$$r_{_W}(i,j,k) = \frac{\sum_{x=1}^{N_p(k)} W_m^{'}(i,j,k,x)\, W_o^{'}(k,x)}{[\sum_{x=1}^{N_p(k)} W_m^{'}(i,j,k,x)^2]^{1/2}\, [\sum_{x=1}^{N_p(k)} W_o^{'}(k,x)^2]^{1/2}} \qquad (7)$$

$$i = 1,\ldots,N_r(j);\ j = 1,\ldots,N_m;\ k = 1,\ldots,N_k$$

Note that since the (local) climatological annual mean has been subtracted at each grid-point, the overall spatio-temporal mean of the concatenated monthly-mean anomaly field is zero for both the model and observational fields.

As in the case of the absolute bias, we calculate the ensemble-mean value of the pattern correlation statistic for each individual model, $\overline{r_{_W}}(j,k)$, the multi-model average pattern correlation statistic, $\overline{\overline{r_{_W}}}(k)$, and the inter-model standard deviation of the

ensemble-mean pattern correlation, $s\{\overline{r_{_W}}(k)\}$ [see equations (2)-(4)].

The normalized pattern correlation statistic for the j^{th} model and k^{th} region is given by:

$$\beta_{_W}(j,k) = [\,\overline{\overline{r_{_W}}}(k) - \overline{r_{_W}}(j,k)] \,/\, s\{\overline{r_{_W}}(k)\} \qquad (8)$$

$$j = 1,\ldots,N_m;\; k = 1,\ldots,N_k$$

Unlike $\alpha_{_W}$, $\beta_{_W}$ provides directional information – *i.e.*, a negative (positive) value of $\beta_{_W}$ indicates that the pattern correlation between the simulated and observed spatial anomaly fields is larger (smaller) than $\overline{\overline{r_{_W}}}(k)$, the multi-model average value of the pattern correlation.

The average normalized pattern correlation over the five regions and two variables, $\widehat{\beta}_{_W}(j,k)$, is then defined in a similar way to the average of the normalized bias statistic [see equation (6)].

4.4 Variability Amplitude Metrics

Variability amplitude metrics are calculated with monthly-mean values of modeled and observed water vapor and SST, spatially-averaged over the k^{th} region. We define anomalies relative to climatological monthly means over January 1988 to December 1999 for water vapor and over January 1961 to December 1990 for SST. The raw anomalies provide information on the monthly variability of W and SST. We also

smooth the raw anomalies with a digital filter frequently used in data assimilation studies (S10). The selected half-power points for the filter were at two and ten years, which allows us to obtain information on model errors in simulating the observed interannual and decadal timescale variability.

In the following, the index l denotes the timescale of the variability analysis, with $l = 1, 2$ for water vapor (1 = monthly, 2 = interannual) and $l = 1, 2, 3$ for SST (3 = decadal). We compare simulated and observed decadal-timescale variability for SST data only, since the SSM/I water vapor data are of insufficient length to obtain a meaningful estimate of the observed decadal variability.

The temporal standard deviation of the raw or digitally filtered model anomalies is given by:

$$s\{<W''_m(i,j,k,l)>\} = \left[\frac{1}{N_t - 1} \sum_{t=1}^{N_t} \left(<W''_m(i,j,k,l,t)> - <W''_m(i,j,k,l,\bullet)>\right)^2 \right]^{1/2} \tag{9}$$

$$i = 1, \ldots, N_r(j);\ j = 1, \ldots, N_m;\ k = 1, \ldots, N_k;\ l = 1, \ldots, N_l$$

where the $\bullet$ denotes an average over time, and the double primes indicate anomalies relative to climatological monthly means. Here, N_t (the total number of months used for calculating temporal standard deviations) is 144 for water vapor (January 1988 to December 1999) and 1200 for SST (January 1900 to December 1999). The temporal standard deviation of the observed water vapor data, $s\{< W''_o(k,l) >\}$, is defined similarly.

As in the study by *Gleckler et al.* (S11), we calculate a 'symmetric' variability statistic, which has the same numeric values for a model that simulates half and twice the observed variability:

$$\phi_{_W}(i,j,k,l) \;=\; \left[\frac{s\{<W_m''(i,j,k,l)>\}}{s\{<W_o''(k,l)>\}} - \frac{s\{<W_o''(k,l)>\}}{s\{<W_m''(i,j,k,l)>\}} \right]^2 \qquad (10)$$

$$i \;=\; 1,\ldots,N_r(j);\; j=1,\ldots,N_m;\; k=1,\ldots,N_k;\; l=1,\ldots,N_l$$

This property is particularly desirable in the context of detection and attribution studies, where systematic model underestimation of the observed variability is of more concern than overestimation of observed variability (since underestimation enhances the likelihood of incorrect identification of an anthropogenic fingerprint). If we had applied a more traditional variance ratio statistic (such as an F ratio), the fractional error in the variance for a model with twice the observed variability would be twice as large as the fractional error for a model with half the observed variability. Use of such a statistic would have resulted in a different ranking of model performance. Note that $\phi_{_W}$ does not provide directional information on model variability errors, and has a value of zero if the model and observed temporal standard deviations are identical.

Calculation of the ensemble-mean, the multi-model average, and the inter-model standard deviation of the statistic values ($\overline{\phi_{_W}}(j,k,l)$, $\overline{\overline{\phi_{_W}}}(k,l)$, and $s\{\overline{\phi_{_W}}(k,l)\}$, respectively) proceeds as described for the absolute bias [see equations (2)-(4)]. Normal-

ization of the variability amplitude statistic is also as in the case of the bias statistic [see equation (5)]:

$$\phi_{_W}(j,k,l) \;=\; \overline{\phi_{_W}}(j,k,l) \;/\; s\{\overline{\phi_{_W}}(k,l)\} \qquad (11)$$

$$j \;=\; 1,\ldots,N_m;\; k=1,\ldots,N_k;\; l=1,\ldots,N_l$$

For each model, values of the variability amplitude statistic are then averaged over variables, regions, and timescales:

$$\widehat{\phi}(j) \;=\; \frac{1}{[N_k \times N_l(W)] + [N_k \times N_l(T)]} \left[\sum_{k=1}^{N_k} \sum_{l=1}^{N_l(W)} \phi_{_W}(j,k,l) \;+\; \sum_{k=1}^{N_k} \sum_{l=1}^{N_l(T)} \phi_{_T}(j,k,l) \right] \qquad (12)$$

$$j \;=\; 1,\ldots,N_m$$

As noted above, the number of timescales considered is not the same for water vapor and SST, so that $N_l(W) = 2$ and $N_l(T) = 3$. Since N_k (the number of regions considered) is 5, $\widehat{\phi}(j)$ represents an average over 25 different sets of variability amplitude statistics.

4.5 Variability Pattern Metrics

These metrics provide information on the similarity between modeled and observed spatial fields of temporal variability. The anomalies used for calculating temporal standard deviation fields are defined exactly as for the variability amplitude metrics.

In the case of the variability pattern metrics, however, anomalies are defined at individual grid-points (relative to local climatological monthly means) rather than for the spatial average. For metrics involving patterns of interannual and decadal variability, the digital filtering of anomaly data was performed as described in Section 4.4.

At each grid-point in the k^{th} region, we compute the temporal standard deviation of the raw or filtered anomaly data [see equation (9)]. The resulting standard deviation patterns are 'centered' on the spatial means of the two fields being compared, and the pattern correlation between these fields is calculated as in equation (7).[5] As for the bias, annual cycle, and variability amplitude statistics, we normalize each model's variability pattern correlation metric by a measure of the inter-model variability in the metric values:

$$\varphi_{_W}(j,k,l) = [\,\overline{\overline{r_{_W}}}(k,l) - \overline{r_{_W}}(j,k,l)] \;/\; s\{\overline{r_{_W}}(k,l)\} \qquad (13)$$

$$j = 1,\ldots,N_m;\; k = 1,\ldots,N_k;\; l = 1,\ldots,N_l$$

where the ensemble-mean, multi-model average, and inter-model standard deviation of the pattern correlation statistic are given by $\overline{r_{_W}}(j,k,l)$, $\overline{\overline{r_{_W}}}(k,l)$, and $s\{\overline{r_{_W}}(k,l)\}$, respectively [see equations (2)-(4)]. This is the same normalization that was used for the correlation between the simulated and observed annual cycle patterns [see equation (8)]. For each model, the average of the 25 sets of normalized variability

[5] With the exception that the summation is now over $N_x(k)$ rather than over $N_p(k)$ spatial points, since we are no longer dealing with 12 concatenated monthly-mean fields.

pattern statistics, $\widehat{\varphi}(j)$, is then determined as in equation (12).

4.6 Combining Metrics

As noted in the main text, our detection and attribution analysis is performed with different subsets of the full 22 models used by *Santer et al.* in their original D&A study (S12). These subsets of 'top ten' and 'bottom ten' models are determined on the basis of three different sets of model quality metrics (and on two different ranking approaches).

The first metric used in our overall ranking of models, $\widehat{Q}_1$, is based on the mean state and annual cycle metrics:

$$\widehat{Q}_1(j) = \frac{1}{2}\left[\widehat{\alpha}(j) + \widehat{\beta}(j)\right] \tag{14}$$

$$j = 1, \ldots, N_m$$

where $\widehat{\alpha}(j)$ and $\widehat{\beta}(j)$ are (respectively) the averages of the normalized statistic values of the 10 mean state and 10 annual cycle metrics for the j^{th} model (see Sections 4.2 and 4.3).

Our second ranking metric, $\widehat{Q}_2$, is a measure of overall model performance in simulating the observed amplitude and pattern of variability:

$$\widehat{Q}_2(j) = \frac{1}{2}\left[\widehat{\phi}(j) + \widehat{\varphi}(j)\right] \tag{15}$$

$$j \;=\; 1, \ldots, N_m$$

where $\widehat{\phi}(j)$ and $\widehat{\varphi}(j)$ are (respectively) the averages of the normalized statistic values of the 25 variability amplitude and 25 variability pattern metrics (see Sections 4.4 and 4.5).

The third and final overall ranking metric, $\widehat{Q}_3$, is simply the average of the normalized statistic values of the 70 individual model quality metrics – *i.e.*, the metrics for the mean state (10), annual cycle (10), variability amplitude (25), and variability pattern (25):

$$\begin{aligned}
\widehat{Q}_3(j) = \frac{1}{70} \Bigg[& \sum_{k=1}^{N_k} \alpha_{_W}(j,k) \;+\; \sum_{k=1}^{N_k} \alpha_{_T}(j,k) \;+ \\
& \sum_{k=1}^{N_k} \beta_{_W}(j,k) \;+\; \sum_{k=1}^{N_k} \beta_{_T}(j,k) \;+ \\
& \sum_{k=1}^{N_k} \sum_{l=1}^{N_l(W)} \phi_{_W}(j,k,l) \;+\; \sum_{k=1}^{N_k} \sum_{l=1}^{N_l(T)} \phi_{_T}(j,k,l) \;+ \\
& \sum_{k=1}^{N_k} \sum_{l=1}^{N_l(W)} \varphi_{_W}(j,k,l) \;+\; \sum_{k=1}^{N_k} \sum_{l=1}^{N_l(T)} \varphi_{_T}(j,k,l) \Bigg]
\end{aligned} \tag{16}$$

$$j \;=\; 1, \ldots, N_m$$

In what we refer to as our 'parametric ranking' procedure, the 22 CMIP-3 models are ranked in three different ways, based on their values of $\widehat{Q}_1$, $\widehat{Q}_2$, and $\widehat{Q}_3$. To determine the 'top ten' and 'bottom ten' models in each of the three parametric

ranking cases, values of the ranking statistics are sorted and arranged from smallest to largest. The ten models with the smallest (largest) values of the ranking statistic are designated as the 'top ten' ('bottom ten').[6]

In our non-parametric ranking procedure, models are ranked from 1 to 22 for each of the 70 model quality metrics. Next, we average the ranks (rather than the statistic values) for the same three sets of metrics used in the parametric ranking: *i.e.*, for the 20 mean state and annual cycle diagnostics, the 50 variability amplitude and variability pattern diagnostics, and the 70 combined diagnostics. In each of these three cases, the average rank is used to sort the individual models, and thus to determine the three sets of 'top ten' and 'bottom ten' models.

As discussed in the main text, the parametric and non-parametric ranking approaches yield similar but not identical results. The former is more sensitive to outliers, so that poor model performance in a relatively small number of metrics can have a large impact on the model's overall parametric rank.

Three general points should be made regarding our strategy for ranking models. First, we emphasize that the metrics used in our ranking strategy were selected for a very specific application – determining which models were most skillful in capturing

[6]The bias statistic α and variability amplitude statistic ϕ were defined so that smaller statistic values denote smaller model errors (see Sections 4.2 and 4.4). Similarly, negative values of the annual cycle statistic β and variability pattern statistic φ indicate smaller model errors in simulating the observed annual cycle and variability patterns (see Sections 4.3 and 4.5).

aspects of the observed climate that are likely to be important in a water vapor D&A study. Other sets of metrics would be required for other applications, and would yield different model rankings.

Second, we have not made any explicit judgements about the importance of the individual metrics we have used. All are assigned equal weight. There is, however, an implicit weighting of metrics in the model ranking based on the 70 combined diagnostics, since we are considering more variability diagnostics (50) than mean state (10) or annual cycle diagnostics (10). We believe that this implicit weighting is justifiable, since model errors in the amplitude and/or structure of natural internal variability are of particular concern in D&A studies.

Third, rankings determined with absolute model errors would be different from those shown here. Our decision to base rankings on normalized errors was motivated by the desire to combine information on model performance from different variables, regions, timescales, and statistical quantities. Combining such diverse information would have been much more difficult to achieve (at least in a parametric ranking scheme) with 70 different sets of absolute errors.

5 Fingerprint analysis

5.1 Regridding and masking of data

Model results were available on different grids (Table S2). To calculate fingerprints from the multi-model averages of the water vapor changes in the 20CEN runs, and to obtain 'pooled' noise estimates from the concatenated control integrations, we regridded 20CEN and control run W data from all 22 models to a common $10^\circ \times 10^\circ$ latitude/longitude grid. Regridding to a relatively coarse-resolution grid reduces the spatial dimensionality of the input datasets, which is of benefit in the estimation of Empirical Orthogonal Functions (EOFs) used in the fingerprint analysis. Because changes in W tend to be smoothly varying, regridding does not lead to appreciable loss of information on the spatial structure of the leading signal or noise modes.

Each model has a 'mask', $M_m(j,x)$, of the ocean fraction on the original model grid. We use the same statistical terminology employed in the discussion of metrics: j is an index over the number of models, and x is an index over the total number of model grid-points (see Section 4.1). Since observed W data were available over ocean only, each model's land W values had to be appropriately masked out in the regridding process – *i.e.*, any land grid-points within a given $10^\circ \times 10^\circ$ 'target' grid cell were excluded from the calculation of the ocean W value for the target grid cell.

For each model, we calculated the ocean fraction at every target grid cell. Global-

mean values of these fractions are generally different across models, reflecting differences in the original land/sea masks. Observed W data and their associated ocean fraction were also transformed to the same $10° \times 10°$ target grid.

5.2 Definition of fingerprint

Let $S(i, j, x, t)$ represent annual-mean W data at grid-point x and time t from the i^{th} realization of the j^{th} model's 20CEN experiment. Here, $i = 1, \ldots N_r(j)$, $j = 1, \ldots N_m$, $x = 1, \ldots N_x$, and $t = 1, \ldots N_t$. Data are expressed as anomalies relative to the smoothed initial state (1900-1909) of the experiment.

The total time in years is $N_t = 100$ (since all 20CEN experiments cover the common period 1900 to 1999), and the total number of model grid points for the 50°N-50°S domain used in the D&A analysis is $N_x = 287$ (after regridding to the common $10° \times 10°$ latitude/longitude grid and masking out land points). Because we are dealing with 10-member subsets of the 22 CMIP-3 models, $N_m = 10$.

The multi-model average water vapor change, $\overline{\overline{S}}(x, t)$, was calculated by first averaging over an individual model's 20CEN realizations (where multiple realizations were available; see Table S2), and then averaging over models [see equations (2) and (3)]. Since the individual model land/sea masks are not identical after regridding, the number of models contributing to the multi-model averages varies near coastlines and in the vicinity of islands.

Finally, we calculated the EOFs of $\overline{\overline{S}}(x,t)$. The fingerprint $F(x)$ is simply the first EOF of the multi-model average water vapor change. $F(x)$ explains at least 88% of the overall variance in each of the 12 fingerprints shown in Fig. 5, and primarily captures the large simulated increase in water vapor over the 20th century (not shown).

In calculating the EOFs of $\overline{\overline{S}}(x,t)$, we had to account for inter-model differences in $M_m(j,x)$, the regridded ocean fraction. We did this in the following way. First, the regridded $M_m(j,x)$ values were set to zero at any grid cell with less than 1% ocean coverage. We then computed $\overline{\overline{M}}_m(x)$, the geometrical mean of the ocean fraction for the current 10-member subset of models. Use of the geometrical mean excludes areas in which any model has zero ocean fraction.

Since the regridded ocean fraction for the observations, $M_o(x)$, may differ from that of $\overline{\overline{M}}_m(x)$, we also need to calculate the 'overall' geometrical mean ocean fraction, $M_{tot}(x)$, which is the geometrical mean of $\overline{\overline{M}}_m(x)$ and $M_o(x)$. Use of $M_{tot}(x)$ ensures that all EOF calculations (and all calculations in the subsequent determination of detection time) are performed on a common grid, with a common land/sea mask. Appropriate weights are carried throughout the EOF analysis. For each grid cell, the weight is the product of the 'overall' geometrical mean ocean fraction and the grid cell's area weight.

5.3 Calculation of concatenated noise datasets

As described in the main text, we generated 12 different noise datasets by concatenating W data from individual control runs. For example, consider 'Test 1' in Fig. 6, which involves the models identified as being within the "top ten" using the M+AC metrics and the non-parametric ranking procedure. For each of the 10 model control runs, we first regridded annual-mean W data to the same target $10^\circ \times 10^\circ$ grid used for fingerprint estimation. Inspection of the spatially-averaged W values revealed that a number of control runs show evidence of residual non-physical drift (S12). Since this drift can bias D&A results, its removal is advisable.

Various drift removal strategies are possible. Here, we assume that drift behavior of column-integrated precipitable water can be well-approximated by a least-squares linear trend. This assumption is not unreasonable for most control runs. In the case of the GISS-EH control run, however, the drift is strongly non-linear, with very large precipitable water changes in the first 20 years, and much smaller drift over the remaining 380 years of integration (Fig. S1A). Since this initial drift is clearly unphysical and unrepresentative of real-world natural internal variability, we decided to discard the first 20 years of the GISS-EH control run. A linear trend fitted to the remaining 380 years of control run data then provides a much better representation of overall drift behavior (Fig. S1B).

After removing the overall linear trend at each grid point in each model control

run, we concatenated the regression residuals to form the noise dataset $C(x, t_j)$. The index t_j indicates that there is now a single concatenated time dimension, with time as a function of the model number j. In the 'Test 1' example, there are a total of 4,267 years of control run data.

In calculating the EOFs of $C(x, t_j)$, we used the geometrical mean ocean fraction mask appropriate for this specific subset of 10 models (see Section 5.2). The first EOF of $C(x, t_j)$ (which is displayed in Fig. S2A) explains 35% of the total variance of the concatenated control run data.

5.4 Method for estimating signal-to-noise ratios and detection time

In the following, we assume for illustrative purposes that both the fingerprint and the concatenated noise data have been obtained from the above-described 'Test 1' models (*i.e.*, the "top ten" models selected on the basis of the M+AC metrics and non-parametric ranking; see Fig. 6). Note also that our discussion deals solely with fingerprints that have not been optimized in order to enhance S/N ratios.[7]

We begin with regridded annual-mean observational data, $O(x, t)$ (from SSM/I), and the concatenated noise data, $C(x, t_j)$. Observed data are expressed as anomalies

[7]This is because the focus of our original multi-model water vapor D&A study was on non-optimized fingerprints (S12).

relative to climatological annual means over the entire period for which we have SSM/I data (1988 to 2008); control runs are detrended and concatenated as described in Section 5.3. $O(x,t)$ and $C(x,t_j)$ are then projected onto the fingerprint $F(x)$, yielding (respectively) a test statistic time series $Z(t)$ and a 'signal free' time series $N(t)$. $Z(t)$ has a length of 21 years, while $N(t)$ in 'Test 1' is of length 4,267 years.

This projection step is performed 12 times, each time using the same observations and the same $C(x,t_j)$ noise data from 'Test 1', but with a different estimate of the fingerprint $F(x)$ (see Fig. 5). In the results displayed in Fig. 6, we fit least-squares linear trends of length $L = 21$ years to each of the 12 $Z(t)$ time series, and then compare these with the standard error of the distribution of non-overlapping L-length trends in $N(t)$. This is the S/N ratio. Because there are 12 different $Z(t)$ time series in 'Test 1', there are 12 different values of the S/N ratio. The colored bar in 'Test 1' represents the average of these 12 S/N ratios. The black error bar denotes the range of the maximum and minimum S/N ratio values.

Comparing the size of the error bars with the size of the 12 colored bars provides information on the relative contributions of fingerprint and noise uncertainty to the estimated S/N ratios (Fig. 6). In the case of model ranking with the 'ALLD' performance metrics, the effect of noise uncertainty is much larger than that of fingerprint uncertainty. For ranking with the M+AC and VAVP metrics, however, fingerprint and noise uncertainty are of comparable importance in terms of their impact on S/N.

As in our previous work (S13), we estimate the detection time by fitting least-squares linear trends of increasing length L to $Z(t)$, and then comparing these with the standard error of the distribution of non-overlapping L-length trends in $N(t)$. Detection is stipulated to occur when the trend in $Z(t)$ exceeds and remains above the 5% significance level. The test is one-tailed, and we assume a Gaussian distribution of trends in $N(t)$. The start date for fitting linear trends to $Z(t)$ is 1988, the first complete year of the SSM/I data. We use a minimum trend length of ten years, so the earliest possible detection time is in 1997. Our estimated detection times in the 12 sensitivity studies vary from 1999 to 2003. Full details of the detection method are given elsewhere (S13).

Supporting References

S1. Mears CA, Wentz FJ, Santer BD, Taylor KE, Wehner MF (2007) Relationship between temperature and precipitable water changes over tropical oceans. *Geophys Res Lett* 34, L24709, doi:10.1029/2007GL031936.

S2. Wentz FJ, Ricciardulli L, Hilburn K, Mears C (2007) How Much More Rain Will Global Warming Bring? *Science* 317:233-235.

S3. Smith TM, Reynolds RW, Peterson TC, Lawrimore, J (2008) Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006). *J Clim* 21:2283-2296.

S4. http://www.ncdc.noaa.gov/oa/climate/research/sst/sst.html#grid.

S5. Sato M, Hansen JE, McCormick MP, Pollack JB (1993) Stratospheric aerosol optical depths, 1850-1990. *J Geophys Res* 98:22987-22994.

S6. Crowley TJ (2000) Causes of climate change over the past 1000 years. *Science* 289:270-277.

S7. AchutaRao KM, Ishii M, Santer BD, Gleckler PJ, Taylor KE, Barnett TP, Pierce DW, Stouffer RJ, Wigley TML (2007) Simulated and observed variability in ocean temperature and heat content. *Proc Nat Acad Sci* 104:10768-10773.

S8. Santer BD, Wigley TML, Gleckler PJ, Bonfils C, Wehner MF, AchutaRao K, Barnett TP, Boyle JS, Brüggemann W, Fiorino M *et al.* (2006) Forced and

unforced ocean temperature changes in Atlantic and Pacific tropical cyclogenesis regions. *Proc Nat Acad Sci* 103:13905-13910.

S9. Gillett NP, Weaver AJ, Zwiers FW, Wehner MF (2004) Detection of volcanic influence on global precipitation.
Geophys Res Lett 31, L14201, doi:10.1029/2004GL020044.

S10. Lynch P, Huang X-Y (1992) Initialization of the HIRLAM model with a digital filter. *Mon Weath Rev* 120:1019-1034.

S11. Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. *J Geophys Res* 113, D06104, doi:10.1029/2007JD008972.

S12. Santer BD, Mears C, Wentz FJ, Taylor KE, Gleckler PJ, Wigley TML, Barnett TP, Boyle JS, Brüggemann W, Gillett NP *et al.* (2007) Identification of human-induced changes in atmospheric moisture. *Proc Natl Acad Sci* 104:15248-15253.

S13. Santer BD, Mikolajewicz U, Brüggemann W, Hasselmann K, Höck H, Maier-Reimer E, Wigley TML (1995) Ocean variability and its influence on the detectability of greenhouse warming signals. *J. Geophys. Res.* 100:10693-10725.

Captions for Figures in Supporting Text

Figure S1: Secular behavior of column-integrated water vapor anomalies in the GISS-EH control run. Results are monthly-mean anomalies of $< W >$, the spatial average of total atmospheric moisture over near-global oceans (50°N-50°S). Anomalies were defined with respect to climatological monthly means over the first 10 years of the integration. The drift over the entire 400 years of control run (panel A) is not well-represented by an overall least-squares linear trend (the black line). After removal of the first 20 years of data, the slow residual drift over the remaining 380 years is well-described by a linear trend (panel B).

Figure S2: As for Fig. 5, but for the leading water vapor noise mode estimated from the 12 different sets of concatenated control runs. The length of concatenated control run used for estimating the leading noise EOF varies from 3,820 to 4,267 years. The variance explained by the leading mode ranges from 30% to 43%.

Figure S3: Leading EOF of water vapor calculated from 10 individual model control runs. Results are for 'Test 12' in Fig. 6, which involves the models identified as being within the "bottom ten" using the "ALLD" metrics and parametric ranking. The explained variance ranges from 11.9% to 69.5%.

Figure S4: Comparison of water vapor principal component time series for "top ten" and "bottom ten" models. Results in panel A (panel B) are for 'Test 9' ('Test

10') in Fig. 6, which involves the models identified as being within the "top ten" ("bottom ten") using the "ALLD" metrics and non-parametric ranking. The principal component time series are the projections of the 'Test 9' and 'Test 10' concatenated control runs onto the multi-model fingerprints in Figs. 5I and 5J, respectively. The time index is nominal. The dashed horizontal lines are measures of the observed variability, and represent the 1σ temporal standard deviation of the projection of the SSM/I water vapor anomaly data onto the multi-model fingerprints. The brown vertical lines and numbers are visual aids to identify the 10 individual control runs that have been concatenated.

Captions for Tables in Supporting Text

Table S1: Forcings used in IPCC 20CEN simulations. Results are partitioned into V and No-V models (first 12 and last ten rows, respectively). A letter 'Y' denotes inclusion of a specific forcing. A question mark indicates a case where there is uncertainty regarding inclusion of the forcing.

Table S2: Technical details of IPCC 20CEN runs and pre-industrial control integrations. The AGCM resolution is given for both spectral models (in terms of the triangular truncation; *e.g.*, T30, T42, *etc.*) and grid-point models (in terms of the latitude-longitude spacing of grid-points). N_r is the number of realizations that were used for calculating 20CEN ensemble means. CTL_1, CTL_N, and L are (respectively)

the first year, last year, and length (in years) of the pre-industrial control runs employed in the D&A analysis. Note that the start date of each control run is arbitrary. As described in Section 5.3, the first 20 years of the GISS-EH control run display large residual drift, and were therefore discarded.

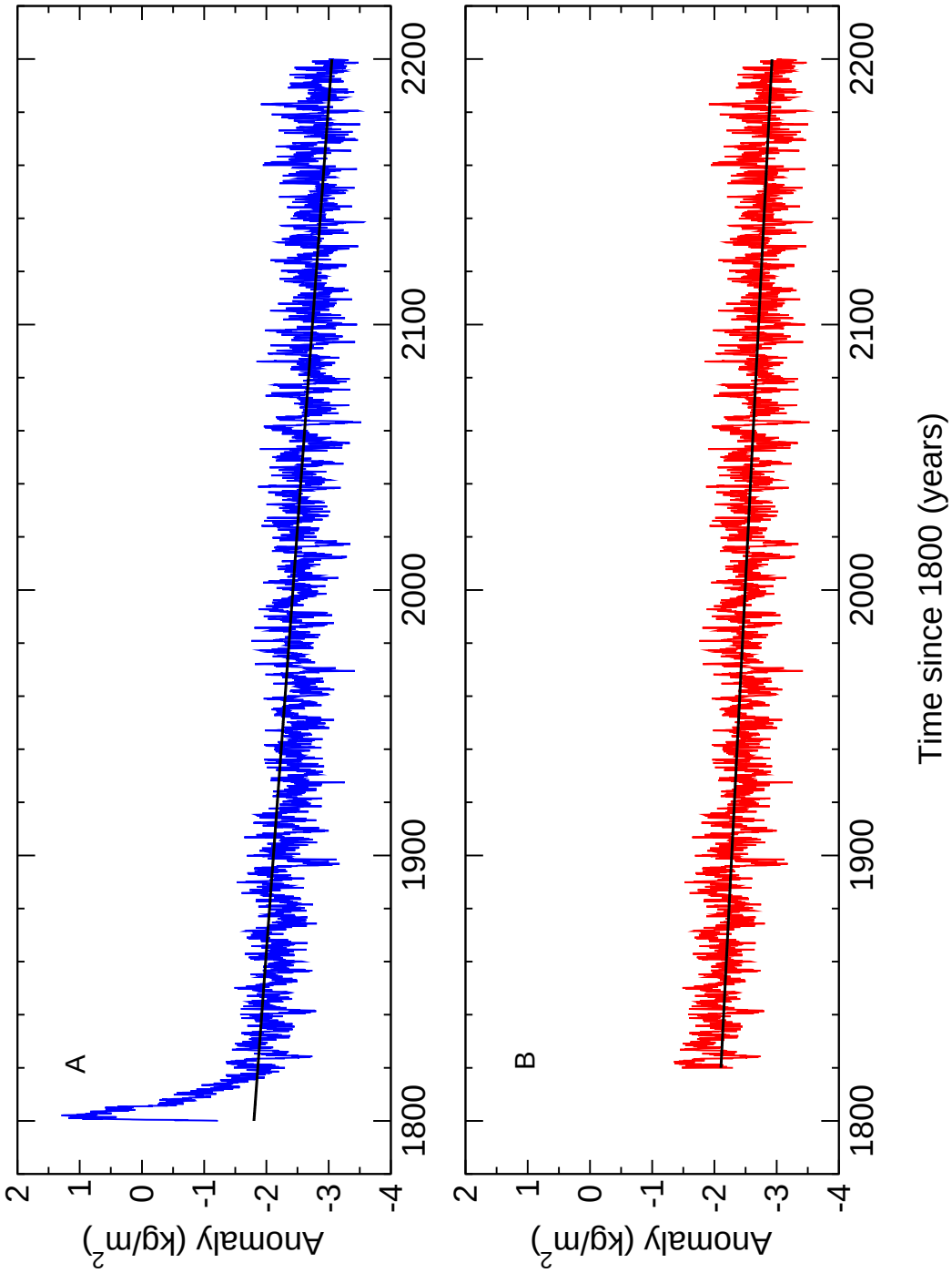

Figure 1: Santer *et al.* Figure S1

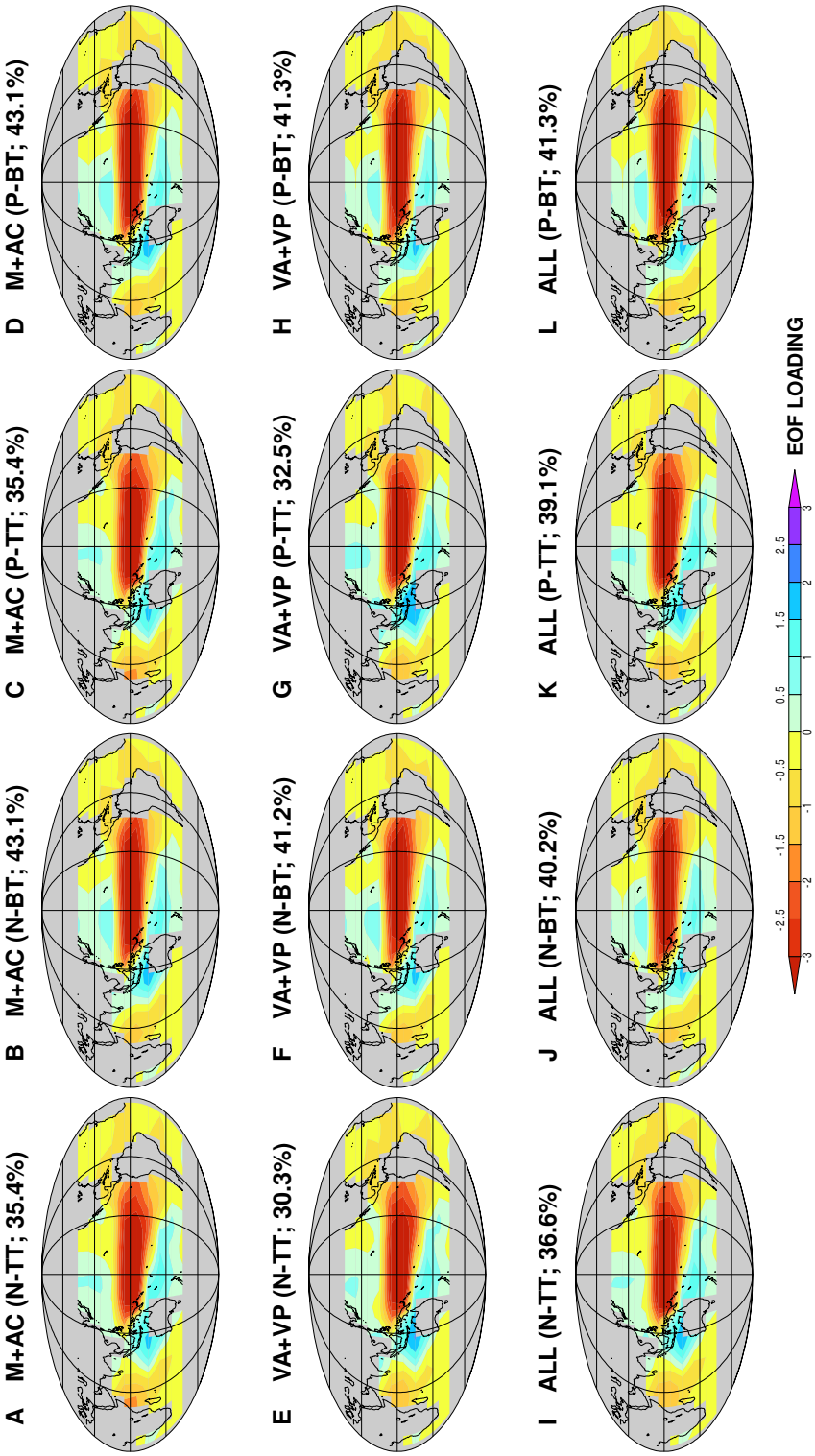

Figure 2: Santer *et al.* Figure S2

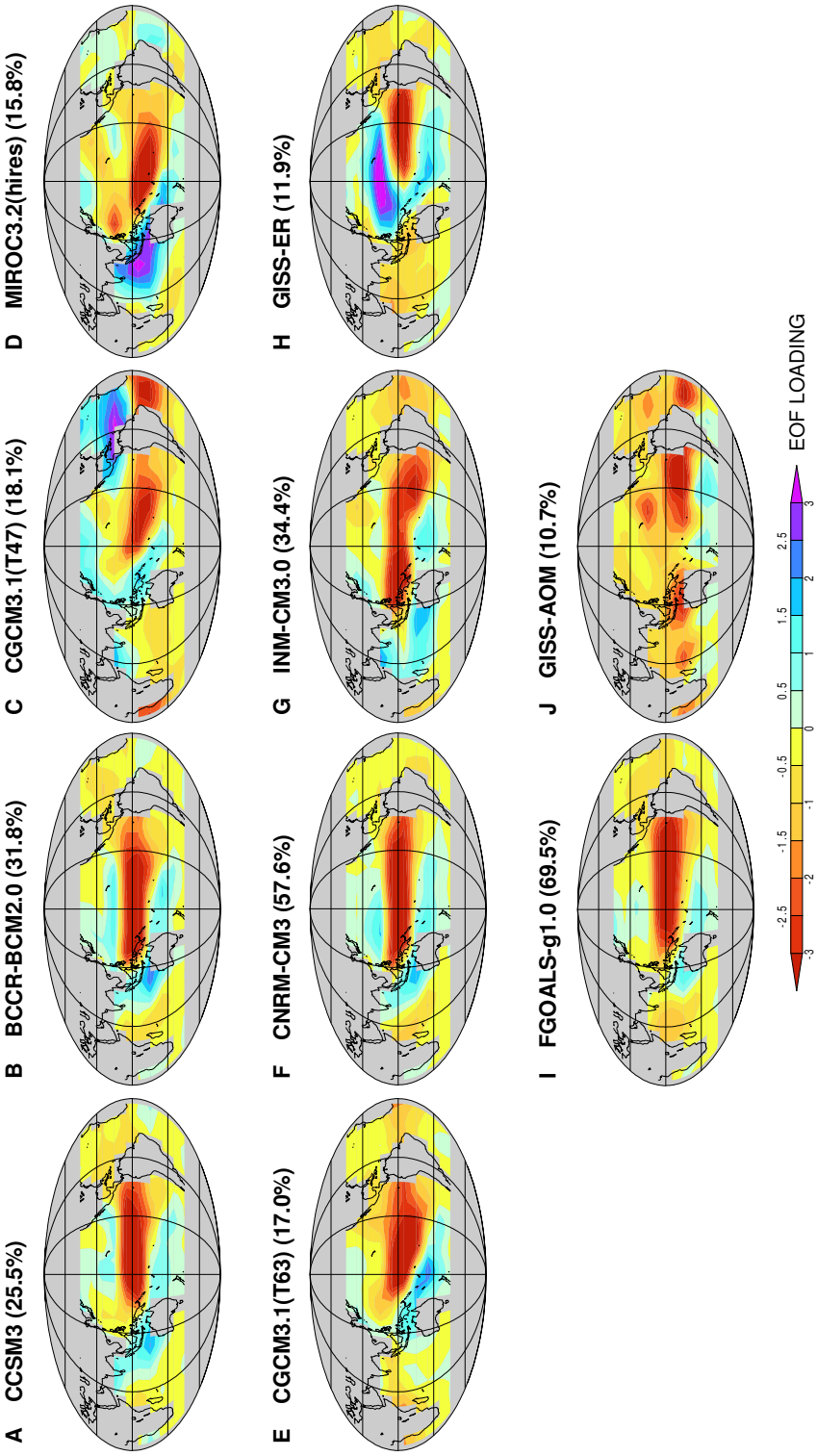

Figure 3: Santer *et al.* Figure S3

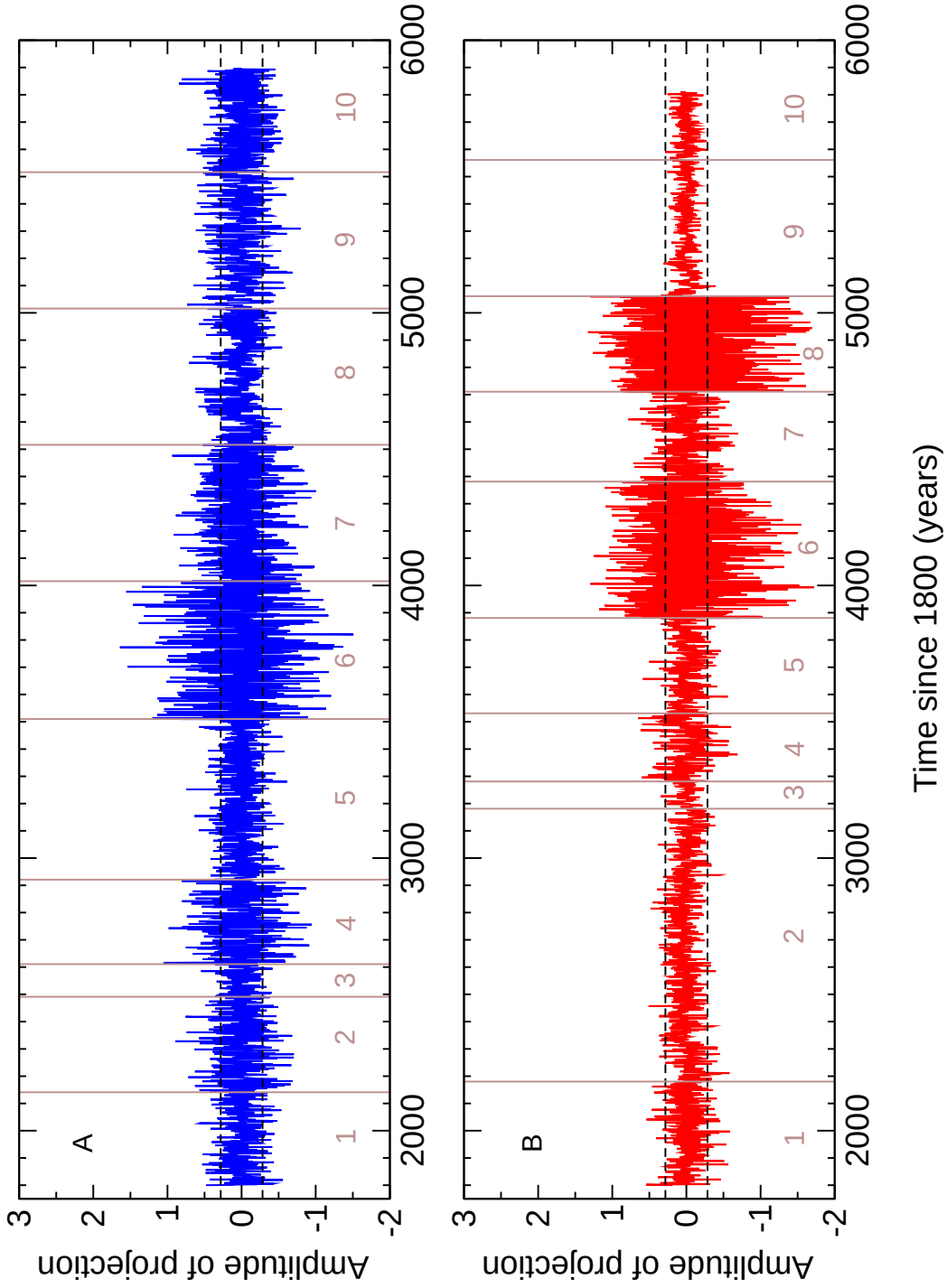

Figure 4: Santer *et al.* Figure S4

Table S1: Forcings used in IPCC simulations of 20th century climate change.

	Model	G	O	SD	SI	BC	OC	MD	SS	LU	SO	VL
1	CCSM3	Y	Y	Y	-	Y	Y	-	-	-	Y	Y
2	GFDL-CM2.0	Y	Y	Y	-	Y	Y	-	-	Y	Y	Y
3	GFDL-CM2.1	Y	Y	Y	-	Y	Y	-	-	Y	Y	Y
4	GISS-EH	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
5	GISS-ER	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
6	MIROC3.2(medres)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
7	MIROC3.2(hires)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
8	MIUB/ECHO-G	Y	-	Y	Y	-	-	-	-	-	Y	Y
9	MRI-CGCM2.3.2	Y	-	Y	-	-	-	-	-	-	Y	Y
10	PCM	Y	Y	Y	-	-	-	-	-	-	Y	Y
11	UKMO-HadCM3	Y	Y	Y	Y	-	-	-	-	-	Y	Y
12	UKMO-HadGEM1	Y	Y	Y	Y	Y	Y	-	-	Y	Y	Y
1	BCCR-BCM2.0	Y	-	Y	-	-	-	-	-	-	-	-
2	CCCma-CGCM3.1(T47)	Y	-	Y	-	-	-	-	-	-	-	-
3	CCCma-CGCM3.1(T63)	Y	-	Y	-	-	-	-	-	-	-	-
4	CNRM-CM3	Y	Y	Y	-	Y	-	-	-	-	-	-
5	CSIRO-Mk3.0	Y	-	Y	-	?	?	?	?	?	?	-
6	ECHAM5/MPI-OM	Y	Y	Y	Y	-	-	-	-	-	-	-
7	FGOALS-g1.0	Y	-	Y	?	-	-	-	-	-	-	-
8	GISS-AOM	Y	-	Y	-	-	-	-	Y	-	-	-
9	INM-CM3.0	Y	-	Y	-	-	-	-	-	-	Y	-
10	IPSL-CM4	Y	-	Y	Y	-	-	-	-	-	-	-

G = Well-mixed greenhouse gases

O = Tropospheric and stratospheric ozone

SD = Sulfate aerosol direct effects

SI = Sulfate aerosol indirect effects

BC = Black carbon

OC = Organic carbon

MD = Mineral dust

SS = Sea salt

LU = Land use change

SO = Solar irradiance

VL = Volcanic aerosols.

Table S2: Technical details of IPCC 20CEN runs and pre-industrial control integrations.

	Model	AGCM resolution	N_r	CTL_1	CTL_N	L
1	CCSM3	T85	8	280	509	230
2	GFDL-CM2.0	$2.0° \times 2.5°$	3	1	500	500
3	GFDL-CM2.1	$2.0° \times 2.5°$	3	1	500	500
4	GISS-EH	$4.0° \times 5.0°$	5	1900	2279	380
5	GISS-ER	$4.0° \times 5.0°$	9	1901	2400	500
6	MIROC3.2(medres)	T42	3	2300	2799	500
7	MIROC3.2(hires)	T106	1	1	100	100
8	MIUB/ECHO-G	T30	5	1860	2200	341
9	MRI-CGCM2.3.2	T42	5	1851	2200	350
10	PCM	T42	4	451	1079	589
11	UKMO-HadCM3	$2.5° \times 3.75°$	1	1800	2109	310
12	UKMO-HadGEM1	$1.25° \times 1.875°$	2	1800	1919	120
1	BCCR-BCM2.0	T63	1	1850	2099	250
2	CCCma-CGCM3.1(T47)	T47	5	1850	2850	1001
3	CCCma-CGCM3.1(T63)	T63	1	1850	2199	350
4	CNRM-CM3	T63	1	1930	2429	500
5	CSIRO-Mk3.0	T63	3	1871	2250	380
6	ECHAM5/MPI-OM	T63	4	2150	2655	506
7	FGOALS-g1.0	T42	3	1850	2199	350
8	GISS-AOM	$3.0° \times 4.0°$	2	1850	2100	251
9	INM-CM3.0	$4.0° \times 5.0°$	1	1871	2200	330
10	IPSL-CM4	$2.5° \times 3.75°$	1	1860	2359	500
	TOTAL	-	71	-	-	8838