Incorporating Model Quality Information in Climate Change Detection and Attribution Studies

B.D. Santer,[a,b] K.E. Taylor,[a] P.J. Gleckler,[a] C. Bonfils,[a] T.P. Barnett,[c] D.W. Pierce,[c] T.M.L. Wigley,[d] C. Mears,[e] F.J. Wentz,[e] W. Brüggemann,[f] N.P. Gillett,[g] S.A. Klein,[a] S. Solomon,[h] P.A. Stott,[i] and M.F. Wehner[j]

[a] Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA 94550, USA; [c] Scripps Institution of Oceanography, La Jolla, CA 92037, USA; [d] National Center for Atmospheric Research, Boulder, CO 80307, USA; [e] Remote Sensing Systems, Santa Rosa, CA 95401, USA; [f] Institut für Unternehmensforschung, Universität Hamburg, 20146 Hamburg, Germany; [g] Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK; [h] National Oceanic and Atmospheric Administration Earth System Research Laboratory, Chemical Sciences Division, Boulder, CO 80305, USA [i] U.K. Met. Office Hadley Centre, Exeter, EX1 3PB, UK; [j] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

Submitted to *Proceedings of the National Academy of Sciences*: February 23, 2009

Revised: June 16, 2009

[b] To whom correspondence should be addressed. Email: santer1@llnl.gov

In a recent multi-model detection and attribution ("D&A") study using the pooled results from 22 different climate models, the simulated "fingerprint" pattern of anthropogenically-caused changes in water vapor was identifiable with high statistical confidence in satellite data. Each model received equal weight in the D&A analysis, despite large differences in the skill with which they simulate key aspects of observed climate. Here, we examine whether water vapor D&A results are sensitive to model quality. The "top ten" and "bottom ten" models are selected with three different sets of skill measures and two different ranking approaches. The entire D&A analysis is then repeated with each of these different sets of more or less skillful models. Our performance metrics include the ability to simulate the mean state, the annual cycle, and the variability associated with El Niño.

We find that estimates of an anthropogenic water vapor fingerprint are insensitive to current model uncertainties, and are governed by basic physical processes that are well-represented in climate models. Because the fingerprint is both robust to current model uncertainties and dissimilar to the dominant noise patterns, our ability to identify an anthropogenic influence on observed multi-decadal changes in water vapor is not affected by "screening" based on model quality.

Since the mid-1990s, pattern-based "fingerprint" studies have been the primary and most rigorous tool for disentangling the complex causes of recent climate change (1–3). Fingerprinting relies on numerical models of the climate system to provide estimates of both the searched-for fingerprint – the pattern of climate response to a change in one or more forcing mechanisms – and the background "noise" of natural internal climate variability. To date, most formal D&A work has used information from only one or two individual models to estimate both the fingerprint and noise (4–6). Relatively few D&A studies have employed climate data from three or more models (7–13).

The availability of large, multi-model archives of climate model output has had important implications for D&A research. A prominent example of such an archive is the CMIP-3 (Coupled Model Intercomparison Project) database, which was a key resource for the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) (14). The CMIP-3 archive enables D&A practitioners to utilize information from two dozen of the world's major climate models, and to examine the robustness of D&A results to current uncertainties in model-based estimates of climate-change signals and natural variability noise (10, 13).

Multi-model databases offer both scientific opportunities and challenges. One challenge is to determine whether the information from each individual model in the database is equally reliable, and should be given equal "weight" in a multi-model D&A

study, or in estimating some "model average" projection of future climate change (15). Previous multi-model D&A investigations with atmospheric water vapor (10) and sea-surface temperatures (SSTs) in hurricane formation regions (13) adopted a "one model, one vote" approach, with no attempt made to weight or screen models based on their performance in simulating aspects of observed climate. An important and hitherto unexplored question, therefore, is whether the findings of such multi-model D&A studies are sensitive to model weighting or screening decisions.

To address this question, objective measures of model performance are required. An obvious difficulty is that model errors are highly complex, and depend on the variable considered, the space and timescale of interest, the statistical metric used to compare modeled and observed climatic fields, the exact property of the fields that is being considered (*e.g.*, mean state, diurnal or annual cycle, amplitude and structure of variability, evolution of patterns, *etc.*), and uncertainties in the observations themselves (16–22). Recent assessments of the overall performance of CMIP-3 models have relied on a variety of statistical metrics, and were primarily focused on how well these models reproduce the observed climatological mean state (23, 24).[1]

In the following, we revisit our multi-model D&A study with atmospheric water

[1] The processes affecting the gradual response of the climate system to long-term anthropogenic forcing need not be the same as those controlling shorter-timescale phenomena. For example, model inadequacies in simulating the diurnal cycle do not necessarily translate to a deficient simulation of long-term responses.

vapor over oceans (10). We calculate a number of different "model quality" metrics, and demonstrate that use of this information to screen models does not affect our ability to identify an externally-forced fingerprint in satellite data.

Observational and Model Water Vapor Data

We rely on observational water vapor data from the satellite-based Special Sensor Microwave Imager (SSM/I). The SSM/I atmospheric moisture retrievals commenced in late 1987, and are based on measurements of microwave emissions from the 22-GHz water vapor absorption line (25–27). Retrievals are unavailable over the highly emissive land surface and sea-ice regions. Our focus is therefore on W, the total column water vapor over oceans for a near-global domain.[2]

As noted above, "fingerprint" studies require estimates of both the climate-change signal in response to external forcing and the noise of internal climate variability. We obtain signal estimates from simulations with historical changes in natural and anthropogenic forcings ("20CEN" runs), and noise information from control integrations with no forcing changes.[3] We use 20CEN and control integrations from 22 different

[2] Our D&A study area encompasses all oceans between 50°N and 50°S. This domain was chosen to minimize the effect of model-versus-SSM/I water vapor differences associated with inaccurate simulation of the latitudinal extent of ice margins.

[3] The external forcings imposed in the 20CEN experiments differed between modeling groups. The most comprehensive experiments included changes in both natural external forcings (solar irradiance and volcanic dust loadings in the atmosphere) and in a wide variety of anthropogenic influences

climate models in the CMIP-3 archive. These are the same models that were employed in our original water vapor D&A study (10).

Strategy for Assessment of Model Quality

Figure 1 illustrates why it may be useful to include model quality information in multi-model D&A studies. The Figure shows the simulated and observed temporal standard deviation of $<W>$, the spatial average of atmospheric water vapor over near-global oceans.[4] Results are given for monthly- and interannual-timescale fluctuations in $<W>$. On both timescales, the simulated variability in 20CEN runs ranges from one-third to two-and-a-half times the amplitude of the observed variability.

Are such variability differences between models and observations of practical importance in multi-model D&A studies? Most D&A studies routinely apply some form of statistical test to check the consistency between observed residual variability (after removal of an estimated externally-forced signal) and model control run variability (4, 7–13), and many studies compare power spectra of the observed and modeled variables being analysed (12, 13). Our focus here is not on formal statistical tests or spectral density comparisons, but instead on calculating metrics which provide more

(such as well-mixed greenhouse gases, ozone, sulfate and black carbon aerosols, and land surface properties). Details of the models, 20CEN experiments, and control integrations are given in the *Supporting Information.*

[4] Here and subsequently, <> denotes a spatial mean.

direct information regarding the fidelity with which models simulate the amplitude and structure of key modes of natural internal variability.

Although our D&A study involves water vapor only, we compute performance metrics both for water vapor and SST. We examine SST data because observed SST datasets are 130 to 150 years in length, and therefore provide a better constraint on model-based estimates of decadal variability than the short (21-year) SSM/I record. Information on low-frequency variability is crucial for D&A applications, since it constitutes the background noise against which we attempt to identify a slowly-evolving anthropogenic signal. All SST-based model quality metrics were calculated using observations from the NOAA Extended Reconstructed SST dataset (ERSST) (28).

We evaluate model performance in simulating W and SST in five different regions. The first is the 50°N-50°S ocean domain employed in our previous water vapor D&A work. The next three regions were chosen because they provide information on model errors in simulating three characteristic modes of natural climate variability: the El Niño/Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO), and the Atlantic Multidecadal Oscillation (AMO).[5] The final region comprises tropical oceans (30°N-30°S), and is of interest because of claims that modeled and observed atmospheric temperature changes differ significantly in the tropics.

[5] ENSO variability can be characterized in a number of different ways. We analyze water vapor and SST changes over the Niño 3.4 region (5°N-5°S; 170°W-120°W). The PDO and AMO regions used here are 20°N-60°N; 115°W-115°E and 20°N-60°N; 75°W-0°, respectively.

We analyze model performance in simulating the mean state, annual cycle, and amplitude and structure of variability.[6] There are 10 mean state diagnostics (two variables × five regions). Each mean state metric is simply a measure of the absolute value of the climatological annual-mean model bias. The 10 annual cycle diagnostics involve the correlations between the simulated and observed climatological mean annual cycle patterns. The 50 variability metrics[7] are measures of model skill in simulating the amplitude and pattern of observed variability on monthly, interannual, and decadal timescales. The rationale for examining model performance on different timescales is that model variability errors are complex and frequency-dependent (29).

All 70 metrics are normalized by some measure of the inter-model standard deviation of the statistical property being considered. This normalization allows us to combine information from the mean, annual cycle, and variability metrics. Details regarding the definition and calculation of our model performance metrics are given

[6] We do not calculate metrics that gauge model performance in simulating observed water vapor and SST trends. Results could be biased towards identification of an anthropogenic fingerprint by first selecting a subset of models with greater skill in replicating observed trends, and then using the same subset in a D&A analysis that compares modeled and observed trend behavior.

[7] For the higher-frequency variability comparisons, there are a total of 40 metrics: two variables (SST and W) × five regions (oceans 50°N-50°S, ENSO, PDO, and AMO regions, and tropical oceans) × two statistical attributes (variability amplitude and pattern) × two timescales (monthly and interannual). For comparisons of decadal variability, there are only 10 diagnostics, since these are meaningful to compute for SST only (see above). All variability pattern metrics are centered correlations, with removal of the spatial means of the two fields being compared.

in the *Supporting Information* (*SI*) *Text*.

Results from Model Quality Assessment

Results for 40 of the 70 individual metrics are shown in Fig. 2. To illustrate the complexity of model errors, we use the example of the UKMO-HadCM3 model (developed at the U.K. Meteorological Office/Hadley Centre). Consider first the results for the absolute bias in the climatological mean state (Fig. 2A). HadCM3 has relatively small bias values for both water vapor and SST, except for SSTs in the PDO region. When models are ranked parametrically on the basis of the "average error" results in Fig. 2A, HadCM3 has the lowest bias values, and is therefore ranked first.

In terms of its simulation of the climatological annual cycle pattern (Fig. 2B) and the amplitude of monthly variability (Fig. 2C), HadCM3 also performs well relative to its peers, and is ranked 7th and 5th (respectively). For the monthly variability pattern, however, HadCM3 has a large error for water vapor in the PDO region (Fig. 2D). This one component has a marked influence on HadCM3's low overall ranking (18th) for the monthly variability pattern. For interannual and decadal variability (not shown), HadCM3 ranks 10th and 1st in terms of its variability amplitude and 15th and 14th in terms of its variability pattern. As is clear from the HadCM3 example and the other model results in Fig. 2, assessments of the relative skill of the CMIP-3 models are sensitive to a variety of analyst choices.

This message is reinforced in Fig. 3, which shows that for our selected variables, regions, and diagnostics, there are no statistically significant relationships between model skill in simulating the climatological mean state and model skill in capturing the observed annual cycle, amplitude, and pattern of monthly variability. Similar findings have been obtained in related studies (19, 20, 22, 23). One possible interpretation of this result is that the spatial averages of observed climatological annual means provide a relatively weak constraint on overall model performance. Modeling groups attempt to reduce biases in these large-scale climatological averages by adjusting poorly known physical parameters (and by flux correction, which still is used in several of the CMIP-3 models). Observed annual cycle and variability patterns offer more stringent tests of model performance. Reliable reproduction of these more challenging observational targets is difficult to achieve through tuning alone – accurate representation of the underlying physics is of greater importance.

The final stage in our model quality assessment is to combine information from different performance metrics. We do this in three different ways. The three combinations involve the 10 mean state and 10 annual cycle diagnostics ("M+AC"), the 25 variability amplitude and 25 variability pattern metrics ("VA+VP"), and the 70 mean state, annual cycle, and variability diagnostics ("ALL"). Individual values of these metrics are averaged, yielding the $\widehat{Q}_1$, $\widehat{Q}_2$, and $\widehat{Q}_3$ statistics, which are used for the parametric ranking of the CMIP-3 models (see *SI Text*). The non-parametric rank is simply the average of the individual ranks rather than the average of individual metric values.

The overall ranking results are shown in Fig. 4. A number of interesting features are evident. First, only three models (MRI-CGCM2.3.2, UKMO-HadGEM1, and IPSL-CM4) are consistently ranked within the top 10 CMIP-3 models based on both ranking approaches and all three sets of performance criteria (M+AC, VA+VP, and ALL). None of the top four models determined with the M+AC metrics (Fig. 4A) is also in the top four based on the VA+VP metrics (Fig. 4B). These results support our previous finding that assessments of model quality are sensitive to the choice of statistical properties used in model evaluation.

Second, there is also some sensitivity to the choice of ranking procedure, particularly for the VA+VP and ALL statistics (Fig. 4B, C). In each of these two cases, the non-parametric and parametric ranking approaches identify slightly different sets of "top 10" models. Only 8 models are in the intersection of these sets.

Third, higher horizontal resolution does not invariably lead to improved model performance. The CMIP-3 archive contains two models (the Canadian Climate Centre's CGCM3.1 and the Japanese MIROC3.2) that were run in both higher- and lower-resolution configurations. The lower-resolution version of CGCM3.1 outperforms the higher-resolution version in terms of the M+AC diagnostics, but not for the VA+VP metrics. The reverse applies to the MIROC3.2 model. The lack of a consistent benefit of higher resolution is partly due to our focus on temperature and moisture changes over oceans. The performance improvement related to higher resolution is more evident over land areas with complex topography (30).

Detection and Attribution Analysis

We now apply the same multi-model D&A method used by Santer *et al.* (10). Instead of employing all 22 CMIP-3 models in the D&A analysis, we restrict our attention to 10-member subsets of the 22 models. These subsets are determined by ranking models on the basis of the three different sets of metrics (M+AC, VA+VP, and ALL) and two different ranking approaches (parametric and non-parametric). From each of these six ranking sets, we select the "top ten" and "bottom ten" models, yielding 12 groups of 10 models.

Fingerprints are calculated in the following way. For each set of ten models, we determine the multi-model average of the atmospheric moisture changes over 1900 to 1999.[8] The fingerprint is simply the first Empirical Orthogonal Function (EOF) of the multi-model average changes in water vapor.

Since 10 modeling groups used anthropogenic forcings only, while the other 12

[8] This involves averaging the ensemble-mean water vapor changes of each model – *i.e.*, averaging the 20CEN realizations of an individual model before averaging over models (see *SI Text*). Note that use of water vapor data for the entire 20th century (rather than simply the period of overlap with SSM/I) provides a less-noisy estimate of the true water vapor response to slowly-varying external forcings, and a response that is more similar across models.

applied a combination of anthropogenic and natural external forcings (see *SI Text*), we expect the multi-model fingerprint to down-weight the contribution of natural external forcing to the fingerprint. However, previous work has found that the fingerprints estimated from combined historical changes in anthropogenic and natural external forcing are very similar to those obtained from "anthropogenic only" forcing (10). We infer from this that anthropogenic forcing is the dominant influence on the changes in atmospheric moisture over the 20th century, and that the multi-model fingerprint patterns are not distorted by the absence of solar and volcanic forcing in 10 of the 22 models analyzed here.[9]

There is pronounced similarity between the fingerprint patterns estimated from the 12 subsets of CMIP-3 models (Fig. 5). All 12 patterns show spatially-coherent water vapor increases, with largest increases over the warmest ocean areas. There are no systematic differences between the fingerprints estimated from different sets of diagnostics, different ranking procedures, or from the top ten or bottom ten models. This indicates that the structure of the water vapor fingerprint is primarily dictated by the zero-order physics governing the relationship between surface temperature and column-integrated water vapor (25, 31).

For each of our 12 subsets of CMIP-3 models, estimates of natural internal vari-

[9] Since volcanic effects on climate have pronounced structure in space and time, they can and have been identified in D&A studies which include information on the spatio-temporal evolution of signal and noise (12).

ability are obtained by concatenating the 10 individual control runs of that subset, after first removing residual drift from each control (Fig. S1, *SI*). The leading EOF patterns estimated from the concatenated control runs are remarkably similar: each displays the horseshoe-shaped pattern characteristic of the effects of ENSO variability on atmospheric moisture (Fig. S2, *SI*). Unlike the fingerprints, the leading noise modes have both positive and negative changes in water vapor.

The similarity of the noise modes in Fig. S2 occurs despite the fact that individual models can have noticeable differences in the spatial structure of their leading mode of water vapor variability (Fig. S3, *SI*). One possible explanation for this result is that errors in the pattern of the dominant noise mode in individual models are quasi-random; these random error components are reduced when the leading noise mode is estimated from a sufficiently large number of concatenated model control runs (22).

The final step was to repeat the multi-model D&A analysis of Santer et al. (10) with updated SSM/I observations, 12 different fingerprints (Fig. 5), and 12 model-based noise estimates (Fig. S2, *SI*). The D&A analysis was performed 144 times, using each possible combination of fingerprint and noise (Fig. 6). We do not employ any form of fingerprint optimization to enhance signal-to-noise (S/N) ratios (3–7, 10). Our D&A method is fully described in the *SI Text*.

In each of the 144 cases, the model-predicted fingerprint in response to external forcing can be positively identified in the observed water vapor data (Fig. 6). S/N ratios for signals calculated over the 21-year period 1988 to 2008 are always above the nominal 5% significance threshold, and exceed the 1% threshold in 62 out of 144 cases. This illustrates that our ability to identify externally-forced changes in water vapor is not affected by the "model screening" choices we have made.

Note that there are systematic differences between S/N ratios estimated with "top ten" and "bottom ten" models, with ratios for the latter larger in all 6 cases (Fig. 6). This result occurs because many of the models ranked in the bottom ten underestimate the observed variability of water vapor, thereby spuriously inflating S/N ratios (Fig. S4, *SI*). In models with more realistic representations of the mean state, annual cycle, and variability, S/N ratios are smaller, but consistently remain above the stipulated 5% significance threshold.

Conclusions

We have shown that the positive identification of an externally-forced fingerprint in satellite estimates of atmospheric water vapor changes is robust to current model uncertainties. Our ability to identify this fingerprint is not affected by restricting our original multi-model D&A study (10) to smaller subsets of models with superior performance in simulating certain aspects of observed water vapor and SST behavior. In fact, we find that even models with noticeable errors in water vapor and SST yield positive detection of an externally-forced fingerprint.

The ubiquitous detection of an externally-forced fingerprint is due to several factors. First, the structure of the water vapor fingerprint is governed by very basic physics, and is highly similar in all 12 of our sensitivity tests (Fig. 5). Second, the fingerprint is characterized by spatially-coherent water vapor increases, while the dominant noise modes in the model control runs are ENSO-like in structure, and do not show coherent water vapor increases over the entire global ocean (Fig. S2, *SI*). Although the structural details of the dominant noise mode differ from model to model (Fig. S3, *SI*), the dissimilarity of the water vapor fingerprint and the leading noise patterns does not. This dissimilarity is the main explanation for the robustness of our D&A results.

The water vapor feedback mechanism is of primary importance in determining the sensitivity of the climate system to external forcing (31, 32). Since our fingerprint estimates are robust across models and relatively insensitive to the model quality metrics calculated here, the contribution of water vapor feedback to projected future climate changes may be similarly insensitive to model skill.[10]

Our study also demonstrates that it is not easy to make an unambiguous iden-

[10] We note, however, that upper tropospheric water vapor is a key component of the water vapor feedback. Our skill measures address only total column water vapor, which is dominated by water vapor in the lower troposphere. Metrics focusing on model performance in simulating the present-day vertical distribution of water vapor may yield stronger relationships between model skill and the component of climate change projections arising from water vapor feedback.

tification of “superior” models, even for a very specific application. Model performance assessments are sensitive to the choice of climate variables, analysis regions and timescales, the physical properties of the fields being compared, the comparison metrics, the way in which individual metrics are normalized and combined, and the ranking approaches (see *SI Text*). There is considerable subjectivity in all of these choices. Different sets of choices would yield different model rankings.

In our analysis of water vapor and SST data, we find that model performance in simulating the mean state is virtually uncorrelated with model performance in reproducing the observed annual cycle or the observed amplitude or pattern of variability. This has implications for attempts to use model performance metrics to weight projections of future climate change. To date, most of these attempts have relied on mean state metrics. Our results imply that different projection weights would be obtained with annual cycle and variability metrics. Whether different weighting approaches lead to important differences in climate-change projections is currently unclear, and may depend on the region, climate variable, and timescale of interest (20, 22). Identification of the ‘best’ models for making projections of future climate change will likely require metrics that can better constrain current uncertainties in feedback mechanisms (33).

Although we find that incorporating model quality information has little impact on our ability to identify an externally-forced water vapor fingerprint, this does not mean that model quality assessment will be of limited value in D&A studies with other variables (8, 11). In the case of water vapor, S/N ratios are invariably above stipulated significance thresholds. If S/N ratios are closer to these thresholds, it may become more important to screen or down-weight models that are deficient in their simulation of the amplitude and structure of natural variability. As we show here, such variability errors can systematically bias D&A results.

In summary, future multi-model D&A studies must deal with the fundamental challenge of how to make appropriate use of the information from a large collection of models of varying complexity and performance levels. Inevitably, model quality assessment will be an integral component of multi-model D&A studies. While a democratic "one model, one vote" approach was successful for the water vapor D&A problem, this approach may not be adequate in all cases.

Acknowledgments

We acknowledge the modeling groups for providing their simulation output for analysis, PCMDI for collecting and archiving this data, and the World Climate Research Programme's Working Group on Coupled Modelling for organizing the model data analysis activity. The CMIP-3 multi-model dataset is supported by the Office of Science, U.S. Department of Energy. NOAA ERSST data were provided by Dick Reynolds at the National Climatic Data Center. PAS was supported by the joint DECC, Defra and MoD Integrated Climate Programme - DECC/Defra (GA01101), MoD (CBC/2B/0417_Annex C5). We thank Gabi Hegerl (University of Edinburgh) and an anonymous reviewer for their constructive comments on the paper.

References

1. Santer BD, Taylor KE, Wigley TML, Johns TC, Jones PD, Karoly DJ, Mitchell JFB, Oort AH, Penner JE, Ramaswamy V *et al.* (1996) A search for human influences on the thermal structure of the atmosphere. *Nature* 382:39-46.

2. Tett SFB, Mitchell JFB, Parker DE, Allen MR (1996) Human influence on the atmospheric vertical temperature structure: detection and observations. *Science* 274:1170-1173.

3. Hegerl GC, von Storch H, Hasselmann K, Santer BD, Cubasch U, Jones PD (1996) Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *J Clim* 9:2281-2306.

4. Stott PA, Tett SFB, Jones GS, Allen MR, Mitchell JFB, Jenkins GJ (2000) External control of 20th century temperature by natural and anthropogenic forcings. *Science* 290:2133-2137.

5. Santer BD, Wehner MF, Wigley TML, Sausen R, Meehl GA, Taylor KE, Ammann C, Arblaster J, Washington WM, Boyle JS *et al.* (2003) Contributions of anthropogenic and natural forcing to recent tropopause height changes. *Science* 301:479-483.

6. Barnett TP, Pierce DW, AchutaRao KM, Gleckler PJ, Santer BD, Gregory JM, Washington WM (2005) Penetration of human-induced warming into the

world's oceans. *Science* 309:284-287.

7. Gillett NP, Zwiers FW, Weaver AJ, Hegerl GC, Allen MR, Stott PA (2002) Detecting anthropogenic influence with a multi-model ensemble. *Geophys Res Lett* 29, 1970, doi:10.1029/2002GL015836.

8. Gillett NP, Zwiers FW, Weaver AJ, Stott PA (2003) Detection of human influence on sea level pressure. *Nature* 422:292-294.

9. Huntingford C, Stott PA, Allen MR, Lambert FH (2006) Incorporating model uncertainty into attribution of observed temperature change. *Geophys Res Lett* 33, L05710, doi:10.1029/2005GL024831.

10. 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.

11. Zhang X, Zwiers FW, Hegerl GC, Lambert FH, Gillett NP, Solomon S, Stott PA, Nozawa T (2007) Detection of human influence on twentieth-century precipitation trends. *Nature* 448:461-465.

12. Hegerl GC, Zwiers FW, Braconnot P, Gillett NP, Luo Y, Marengo Orsini JA, Nicholls N, Penner JE, Stott PA (2007) Understanding and attributing climate change. in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental*

Panel on Climate Change. [Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

13. Gillett NP, Stott PA, Santer BD (2008) Attribution of cyclogenesis region sea surface temperature change to anthropogenic influence. *Geophys Res Lett* 35, L09707, doi: 10.1029/2008GL033670.

14. IPCC (Intergovernmental Panel on Climate Change). Summary for Policymakers. in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* [Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

15. Giorgi F, Mearns LO (2002) Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the reliability ensemble averaging (REA) method. *J Clim* 15:1141-1158.

16. Preisendorfer RW, Barnett TP (1983) Numerical model-reality intercomparison tests using small-sample statistics. *J Atmos Sci* 40:1884-1896.

17. Wigley TML, Santer BD (1990) Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *J Geophys Res* 95:851-865.

18. Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res* 106:7183-7192.

19. Diffenbaugh NS (2005) Response of large-scale eastern boundary current forcing in the 21st century. *Geophys Res Lett* 32, L19718, doi:10.1029/2005GL023905.

20. Brekke LD, Dettinger MD, Maurer EP, Anderson M (2008) Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. *Clim Change* 89:371-394.

21. Waugh DW, Eyring V (2008) Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmos Chem Phys* 8:5699-5713.

22. Pierce DW, Barnett TP, Santer BD, Gleckler PJ (2009) Selecting global climate models for regional climate change studies. *Proc Nat Acad Sci* 106:8441-8446.

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

24. Reichler T, Kim J (2008) How well do coupled models simulate todays climate? *Bull Amer Met Soc*, doi:10.1175/BAMS-89-3-303.

25. Wentz FJ, Schabel M (2000) Precise climate monitoring using complementary data sets. *Nature* 403:414-416.

26. 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.

27. Trenberth KE, Fasullo J, Smith L (2005) Trends and variability in column-integrated atmospheric water vapor. *Clim Dyn* 24:741-758.

28. 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.

29. AchutaRao K, Sperber KR (2006) ENSO simulation in coupled atmosphere-ocean models: are the current models better? *Cli Dyn* 27:1-15.

30. Duffy PB, Govindasamy B, Milovich J, Taylor KE, Thompson S (2003) High Resolution Simulations of Global Climate, Part 1: Present Climate. *Cli Dyn* 21:371-390.

31. Soden BJ, Held IM (2006) An assessment of climate feedbacks in coupled ocean-atmosphere models. *J Clim* 19:3354-3360.

32. Held IM, Soden BJ (2006) Robust responses of the hydrological cycle to global warming. *J Clim* 19:5686-5699.

33. Hall A, Qu X (2006) Using the current seasonal cycle to constrain snow albedo feedback in future climate change.
Geophys Res Lett 33, L03502, doi:10.1029/2005GL025127.

Figure 1: Comparison of the simulated and observed temporal variability of atmospheric water vapor. Observations are from the SSM/I dataset (25, 26); model data are from 71 realizations of 20th century climate change performed with 22 different models (see *SI Text*). All variability calculations rely on monthly-mean values of $< W >$, the spatial average of total atmospheric moisture over near-global oceans. Model and observational $< W >$ data were first expressed as anomalies relative to climatological monthly means over 1988 to 1999, and then linearly detrended. We computed temporal standard deviations from both the unfiltered and filtered anomaly data. The latter were smoothed using a filter with a half-power point at *ca.* two years. The raw and filtered standard deviations provide information on monthly- and interannual-timescale variability, respectively. All calculations were over the 144-month period from January 1988 to December 1999 (the period of maximum overlap between the SSM/I data and most 20CEN simulations). The dashed grey lines are centered on the observations.

Figure 2: Results for four different sets of metrics used in the ranking of model performance. The statistics are measures of how well 22 of the models in the CMIP-3 database reproduce key features of observed water vapor and SST behavior in five different geographical regions. The metrics shown here are a subset of the full suite of metrics that we applied for model ranking, and are for the mean state (A), annual cycle pattern (B), amplitude of monthly variability (C), and pattern of monthly variability (D). Full details of how these metrics are defined and calculated are given

in the *SI Text*. For models with multiple 20CEN realizations, values of metrics are averaged over realizations. The black dots labelled "average error" represent the arithmetic average (for each model) of the 10 metric values (2 variables × 5 regions). Metrics are normalized to facilitate the combination of information from different climate variables and geographical regions. Small values of the normalized metrics in panels A and C indicate greater skill in simulating the mean state and the amplitude of monthly variability; negative values of the normalized pattern correlation metrics in panels B and D denote greater skill in simulating the annual cycle and monthly variability patterns.

Figure 3: Relationship between between model skill in simulating the mean state and skill in simulating the annual cycle pattern (A), amplitude of monthly variability (B), and monthly variability pattern (C). Results plotted are the "average errors" shown and described in Fig. 2. The black lines are the fitted least-squares regression lines. Models to the left of the vertical dashed grey line are ranked in the "top ten" based on values of the mean state metric $\widehat{\alpha}$. Models below the horizontal dashed grey line are ranked in the "top ten" based on values of the annual cycle pattern metric $\widehat{\beta}$ (A), the variability amplitude metric $\widehat{\phi}$ (B), and the variability pattern metric $\widehat{\varphi}$ (C). The grey shaded region indicates the intersection of the two sets of "top ten" models plotted in each panel.

Figure 4: Parametric and non-parametric ranking of 22 CMIP-3 models. The parametric ranking is based on the $\widehat{Q}_1$, $\widehat{Q}_2$, and $\widehat{Q}_3$ statistics, which are (respectively) measures of model skill in simulating the observed mean state and annual cycle (A), the amplitude and pattern of variability (B), and the combined mean state, annual cycle, and variability properties (C). The $\widehat{Q}_1$, $\widehat{Q}_2$, and $\widehat{Q}_3$ statistics are averages of the normalized values of 20 mean state and annual cycle metrics ("M+AC"), 50 variability amplitude and variability pattern metrics ("VA+VP"), and 70 combined metrics ("ALL"). In the non-parametric ranking procedure, models are ranked from 1 to 22 for each of the 70 metrics, and the individual ranks are then averaged in each of the three groups of metrics (M+AC, VA+VP, and ALL). Full details of the statistics and ranking procedures are given in the *SI Text*. The grey shaded boxes indicate the intersection of the two sets of "top ten" models (based on the parametric and non-parametric ranking approaches).

Figure 5: Model fingerprints of externally-forced changes in water vapor over near-global oceans. Fingerprints were estimated from 12 different 10-member sets of model 20CEN simulations, as described in Fig. 6 and the *SI Text*. The fingerprint is the leading EOF of the multi-model average change in water vapor over the 20th century. The first four fingerprints (panels A-D) were estimated from the "top ten" ("TT") and "bottom ten" ("BT") models, with non-parametric ("N") and parametric ("P") rankings based on the M+AC metrics (see Fig. 4). The fingerprints in panels E-H were estimated from models ranked with the VA+VP pattern statistics. The final four fingerprints (panels I-L) were calculated from models ranked with a combination

of mean state, annual cycle, and variability metrics (ALL). All fingerprint calculations were performed on a common 10° × 10° latitude/longitude grid. The variance explained by the leading mode ranges from 88.3% to 94.0%.

Figure 6: Sensitivity of signal-to-noise (S/N) ratios to "model quality" information. As described in the main text, 12 different sets of ten models were employed to calculate 12 externally-forced fingerprints and 12 estimates of natural internal variability. The D&A analysis was then performed 144 times, with all possible combinations of fingerprint and noise. In "Test 1", for example, the D&A analysis was run 12 times, each time with the same concatenated control runs (from the top ten models determined with the M+AC metrics and non-parametric ranking), but with a different fingerprint (see Fig. 5). The height of each colored bar is the average of the 12 S/N values for the current test. The black error bars denote the maximum and minimum S/N ratios, and are a measure of the effects of fingerprint uncertainty on S/N. The signal in the S/N ratio is the linear trend over 1988 to 2008 in $Z(t)$, the projection of the SSM/I water vapor data onto the fingerprint estimated from the current 10-member set of 20CEN runs. The noise is the standard deviation of the sampling distribution of 21-year trends. This distribution is estimated by fitting non-overlapping 21-year trends to $N(t)$, the time series of the projection of the current set of 10 concatenated control runs onto the fingerprint. Detection of the externally-forced fingerprint in observed data occurs when the S/N ratio exceeds and remains above a stipulated 5% significance threshold. The 1% significance threshold is also shown.

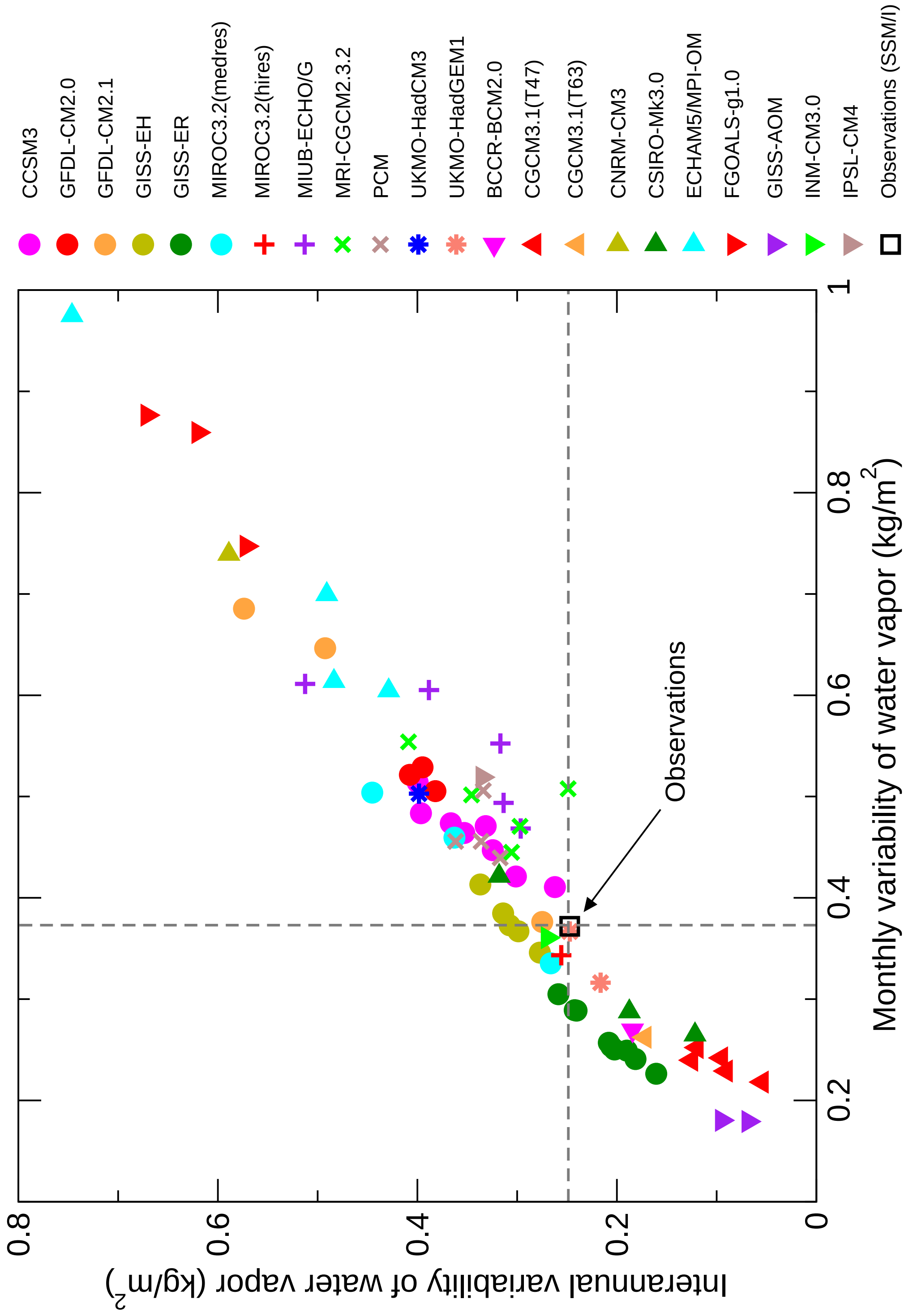

Figure 1: Santer *et al.*

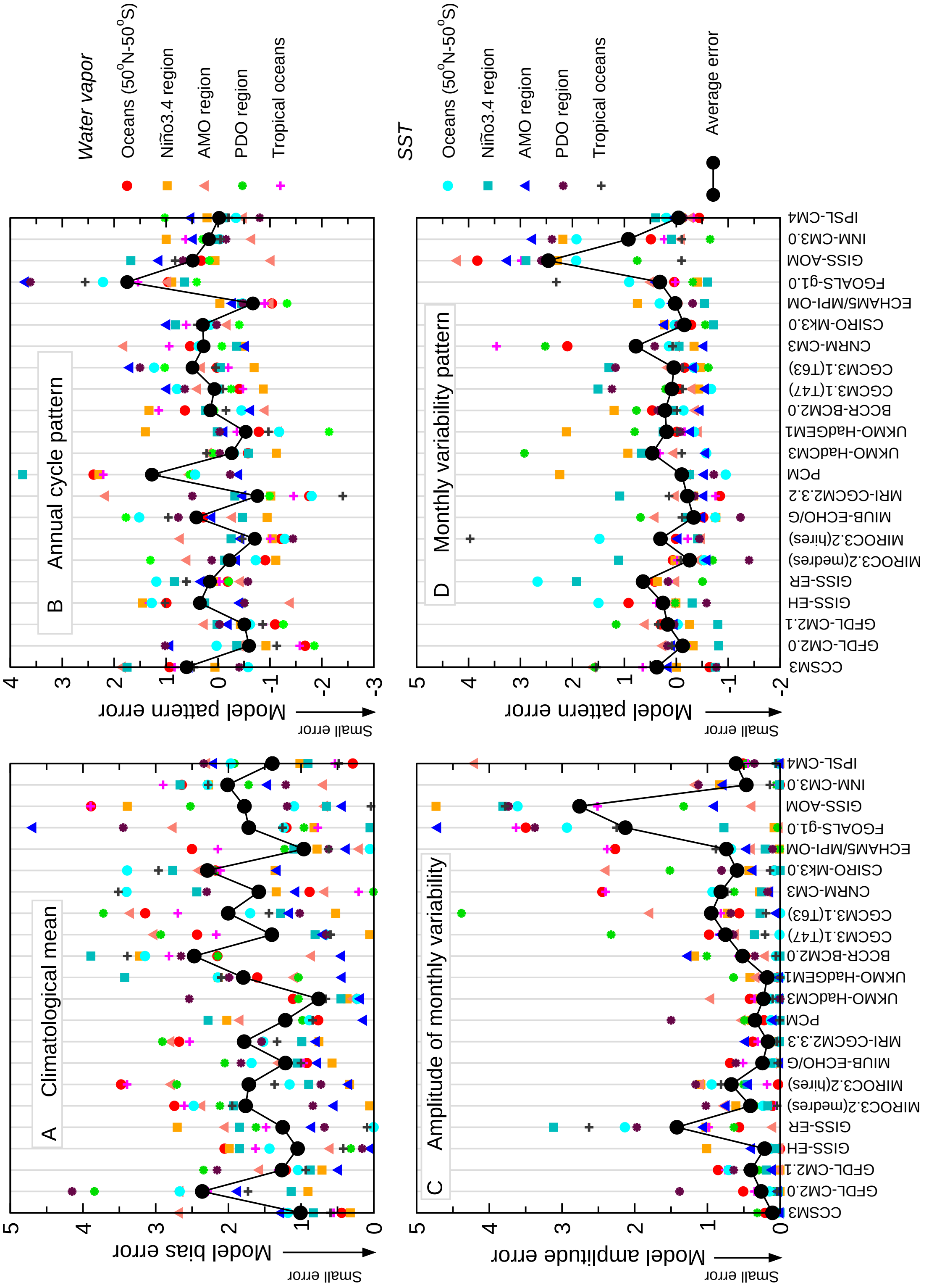

Figure 2: Santer *et al.*

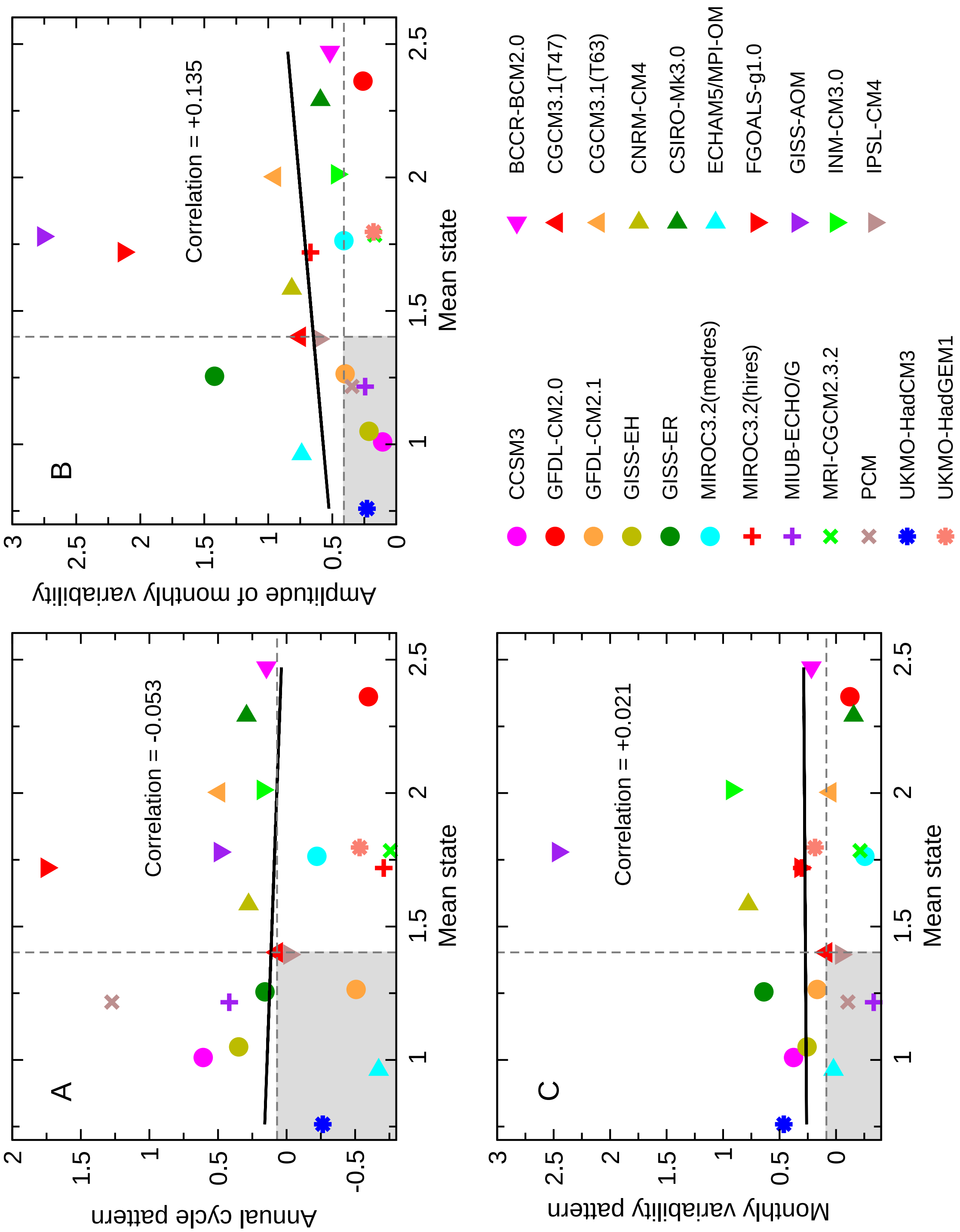

Figure 3: Santer *et al.*

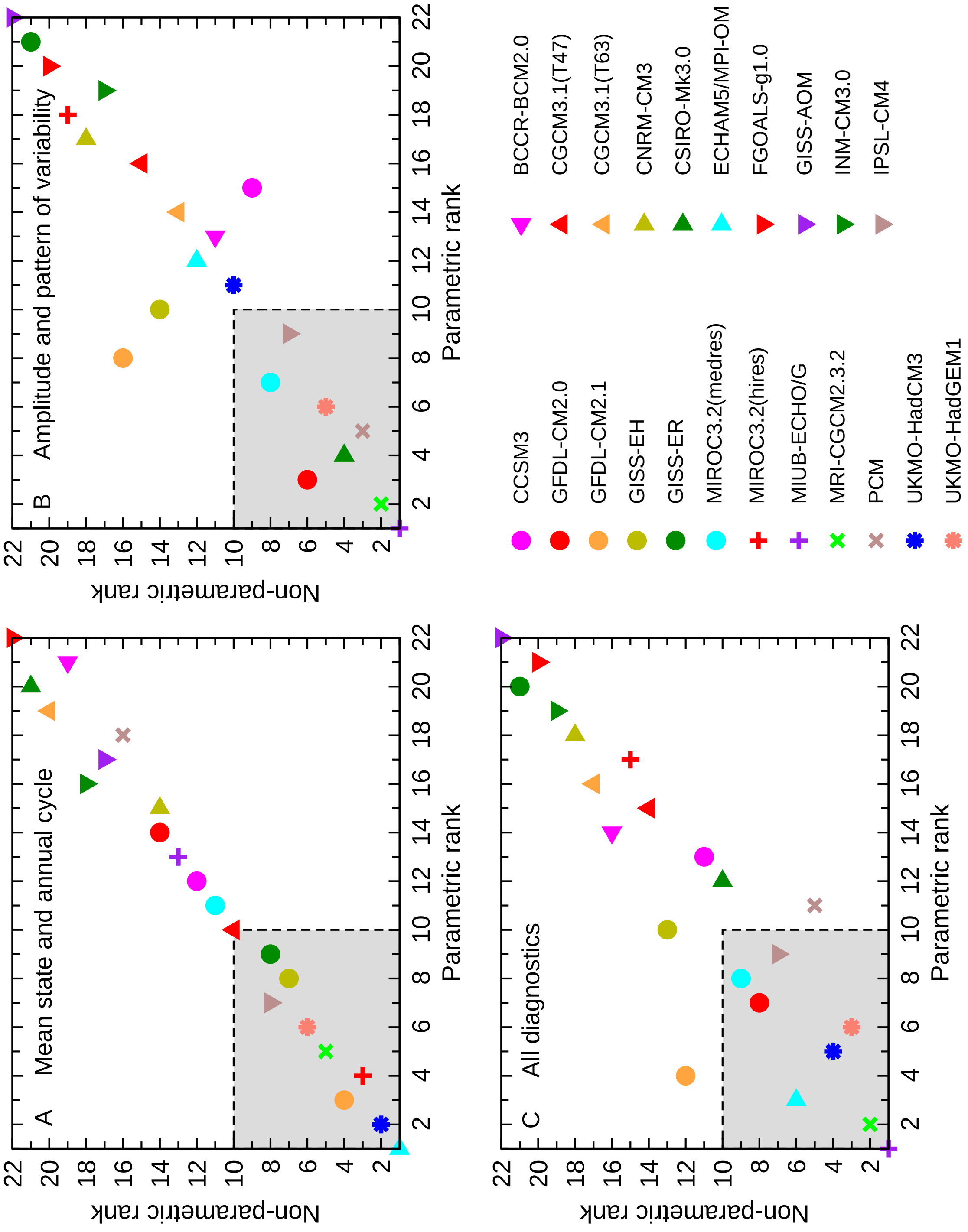

Figure 4: Santer *et al.*

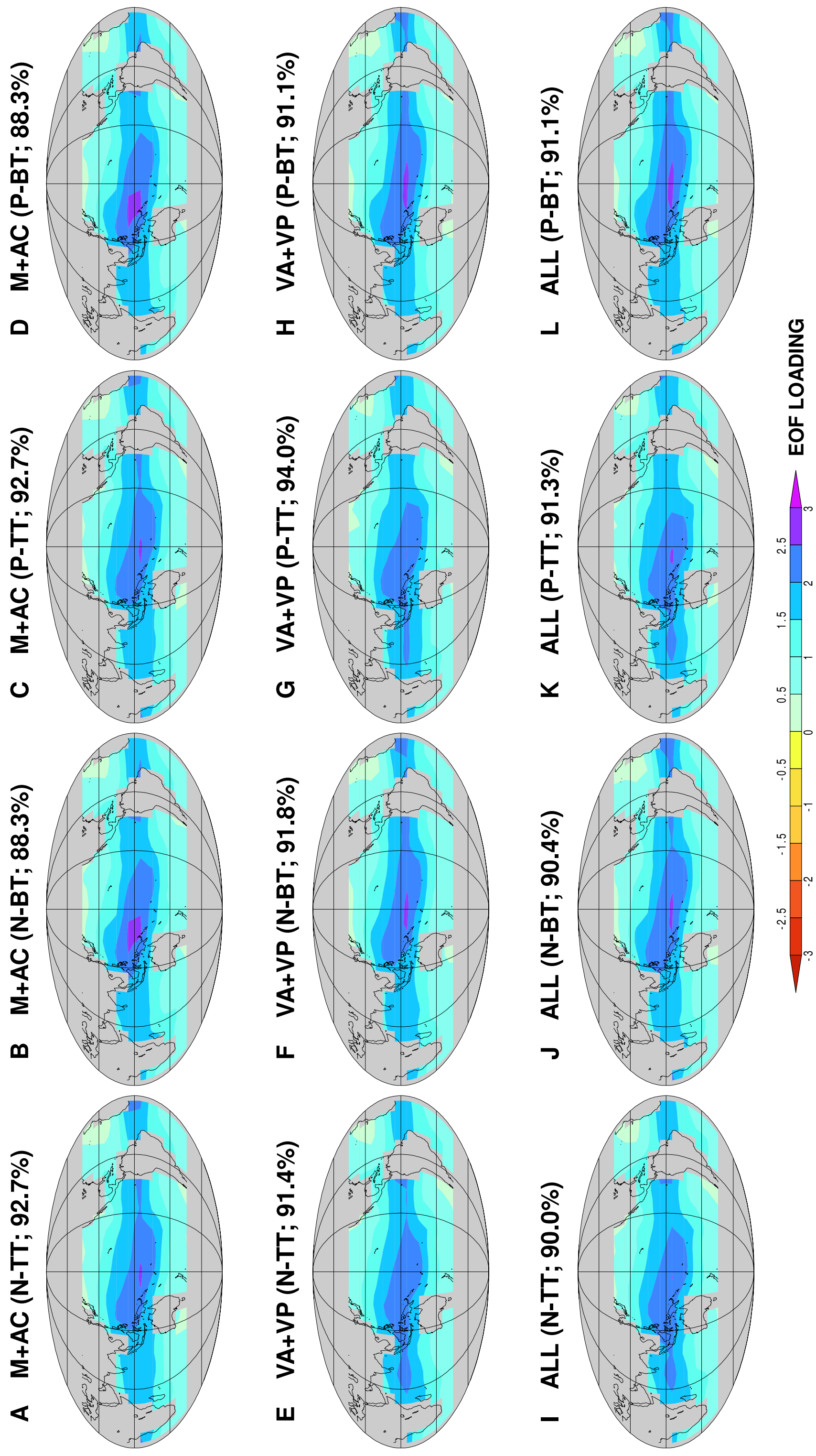

Figure 5: Santer *et al.*

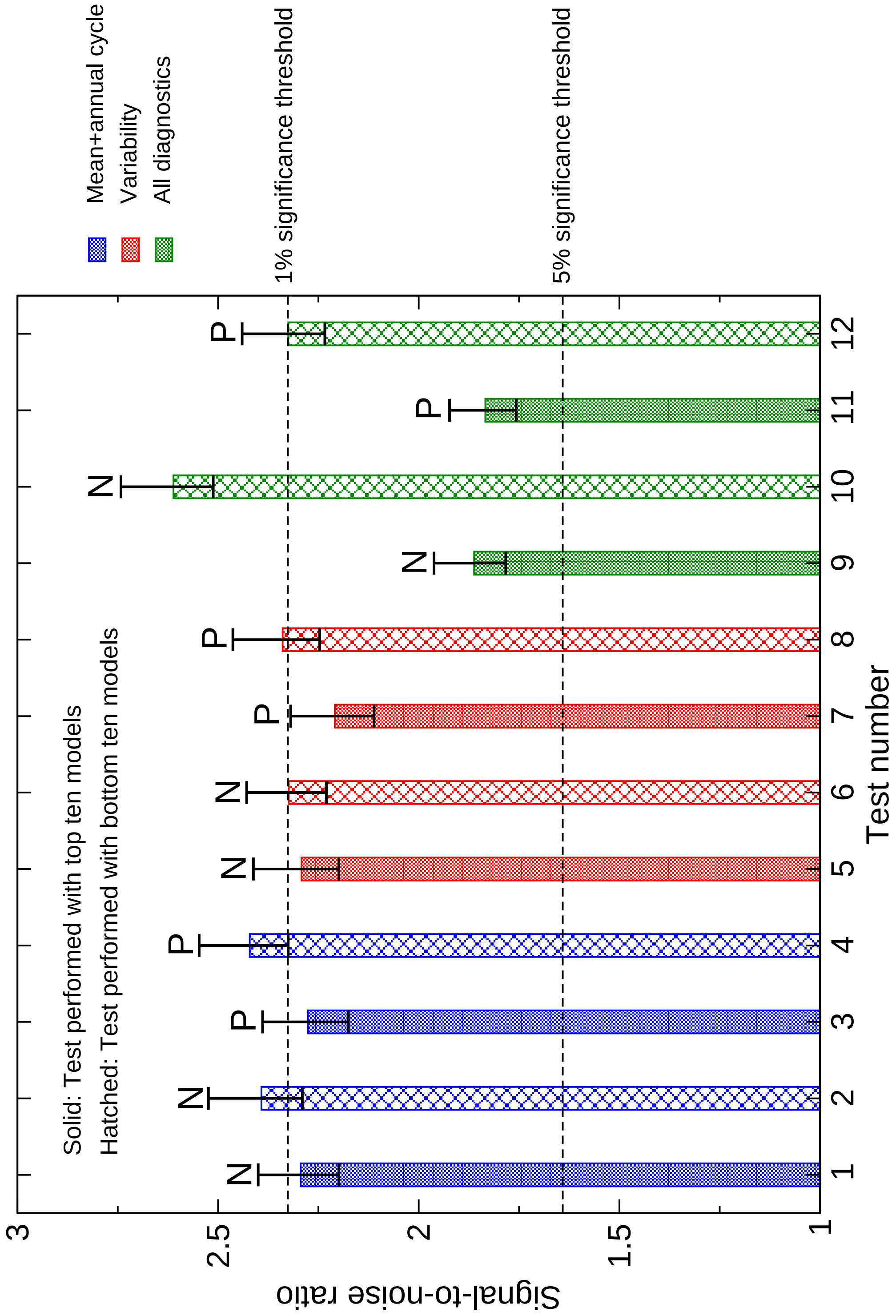

Figure 6: Santer *et al.*