INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **28**: 1703–1722 (2008)
Published online 10 October 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1756

RMetS
Royal Meteorological Society

Consistency of modelled and observed temperature trends in the tropical troposphere

B. D. Santer,[a]* P. W. Thorne,[b] L. Haimberger,[c] K. E. Taylor,[a] T. M. L. Wigley,[d] J. R. Lanzante,[e] S. Solomon,[f] M. Free,[g] P. J. Gleckler,[a] P. D. Jones,[h] T. R. Karl,[i] S. A. Klein,[a] C. Mears,[j] D. Nychka,[d] G. A. Schmidt,[k] S. C. Sherwood,[l] and F. J. Wentz[j]

[a] *Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA 94550, USA*
[b] *U.K. Meteorological Office Hadley Centre, Exeter, EX1 3PB, UK*
[c] *Department of Meteorology and Geophysics, University of Vienna, Althanstrasse 14, A-1090, Vienna, Austria*
[d] *National Center for Atmospheric Research, Boulder, CO 80307, USA*
[e] *National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, NJ 08542, USA*
[f] *National Oceanic and Atmospheric Administration/Earth System Research Laboratory, Chemical Sciences Division, Boulder, CO 80305, USA*
[g] *National Oceanic and Atmospheric Administration/Air Resources Laboratory, Silver Spring, MD 20910, USA*
[h] *Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK*
[i] *National Oceanic and Atmospheric Administration/National Climatic Data Center, Asheville, NC 28801, USA*
[j] *Remote Sensing Systems, Santa Rosa, CA 95401, USA*
[k] *NASA/Goddard Institute for Space Studies, New York, NY 10025, USA*
[l] *Yale University, New Haven, CT 06520, USA*

ABSTRACT: A recent report of the U.S. Climate Change Science Program (CCSP) identified a '*potentially serious inconsistency*' between modelled and observed trends in tropical lapse rates (Karl *et al*., 2006). Early versions of satellite and radiosonde datasets suggested that the tropical surface had warmed more than the troposphere, while climate models consistently showed tropospheric amplification of surface warming in response to human-caused increases in well-mixed greenhouse gases (GHGs). We revisit such comparisons here using new observational estimates of surface and tropospheric temperature changes. We find that there is no longer a serious discrepancy between modelled and observed trends in tropical lapse rates.

This emerging reconciliation of models and observations has two primary explanations. First, because of changes in the treatment of buoy and satellite information, new surface temperature datasets yield slightly reduced tropical warming relative to earlier versions. Second, recently developed satellite and radiosonde datasets show larger warming of the tropical lower troposphere. In the case of a new satellite dataset from Remote Sensing Systems (RSS), enhanced warming is due to an improved procedure of adjusting for inter-satellite biases. When the RSS-derived tropospheric temperature trend is compared with four different observed estimates of surface temperature change, the surface warming is invariably amplified in the tropical troposphere, consistent with model results. Even if we use data from a second satellite dataset with smaller tropospheric warming than in RSS, observed tropical lapse rate trends are not significantly different from those in all other model simulations.

Our results contradict a recent claim that all simulated temperature trends in the tropical troposphere and in tropical lapse rates are inconsistent with observations. This claim was based on use of older radiosonde and satellite datasets, and on two methodological errors: the neglect of observational trend uncertainties introduced by interannual climate variability, and application of an inappropriate statistical 'consistency test'. Copyright © 2008 Royal Meteorological Society

KEY WORDS tropospheric temperature changes; climate model evaluation; statistical significance of trend differences; tropical lapse rates; differential warming of surface and temperature

Received 25 March 2008; Revised 18 July 2008; Accepted 20 July 2008

1. Introduction

There is now compelling scientific evidence that human activities have influenced global climate over the past century (e.g. Intergovernmental Panel on Climate Change (IPCC), 1996, 2001, 2007; Karl *et al*., 2006). A key line of evidence involves 'fingerprint' studies, which attempt to identify the causes of historical climate change through rigorous statistical comparison of models and observations (e.g. Santer *et al*., 1996; Mitchell *et al*., 2001; Hegerl *et al*., 2007). Fingerprint research consistently finds that natural causes alone cannot explain the recent changes in many different aspects of the climate system – the simplest, most internally consistent explanation of the observations invariably involves a pronounced human effect.

One recurring criticism of such findings is that the climate models employed in fingerprint studies are in fundamental disagreement with observations of

* Correspondence to: B. D. Santer, Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. E-mail: santer1@llnl.gov

tropospheric temperature change (Douglass *et al*., 2004, 2007). In climate model simulations, increases in well-mixed GHGs cause warming of the tropical troposphere relative to the surface (Manabe and Stouffer, 1980). In contrast, some satellite and radiosonde datasets show little or no warming of the tropical troposphere since 1979, and imply that temperature changes aloft are smaller than at the surface.

The 'differential warming' of the surface and troposphere has been the subject of intense scrutiny (NRC, 2000; Santer *et al*., 2005; Karl *et al*., 2006; Trenberth *et al*., 2007). It has raised questions about both model performance and the reliability of observed estimates of surface warming (Singer, 2001). In addressing the latter concern, the first report of the U.S. Climate Change Science Program (CCSP) noted that progress had been made in identifying and correcting for errors in satellite and radiosonde data. At the global scale, newer upper-air datasets showed '*no significant discrepancy*' between surface and tropospheric warming, consistent with model results (Karl *et al*., 2006, p. 3). The Fourth Assessment Report of the IPCC reached similar findings, concluding that '*New analyses of balloon-borne and satellite measurements of lower- and mid-tropospheric temperature show warming rates that are similar to those of the surface temperature record'* (IPCC, 2007, p. 5).

The CCSP report used several of these newer observational datasets in extensive comparisons of simulated and observed temperature changes. For global-mean changes, model estimates of differential warming were consistent with observations. In the tropics, however, it was noted that '*most observational datasets show more warming at the surface than in the troposphere, while most model runs have larger warming aloft than at the surface*' (Karl *et al*., 2006, p. 90). Although the CCSP report did not make a definitive determination of the cause or causes of these tropical discrepancies, it found that 'structural uncertainties' in observations were large enough to encompass the model estimates of temperature change. Residual errors in the satellite and radiosonde data were therefore judged to be the most likely explanation for the remaining discrepancies (Karl *et al*., 2006, p. 3).

Structural uncertainties arise because different groups make different processing choices in the complex procedure of adjusting raw measurements for inhomogeneities (Thorne *et al*., 2005a). In radiosonde temperature records, inhomogeneous behaviour can be caused by changes in site location, measurement time, instrumentation, and the effectiveness of thermal shielding of the temperature sensor (Lanzante *et al*., 2003; Seidel *et al*., 2004; Sherwood *et al*., 2005; Randel and Wu, 2006; Mears *et al*., 2006). Non-physical temperature changes in satellite records can occur through orbital drift or decay, inter-satellite instrumental biases, and drifts in instrumental calibration (Wentz and Schabel, 1998; Christy *et al*., 2000, 2003; Mears *et al*., 2003, 2006; Mears and Wentz, 2005; Trenberth *et al*., 2007). Because of these large uncertainties, neither satellite- nor radiosonde-based atmospheric temperature measurements constitute an unimpeachable gold standard for evaluating model performance (Thorne *et al*., 2007).

A recent study by Douglass, Christy, Pearson, and Singer (Douglass *et al*., 2007; hereinafter DCPS07) revisits its earlier comparisons of simulated and observed tropospheric temperature changes performed by Santer *et al*. (2005, 2006), and concludes that '*models and observations disagree to a statistically significant extent.*' This contradicts the findings of both Santer *et al*. (2005) and the previously mentioned CCSP and IPCC reports (Karl *et al*., 2006; IPCC, 2007). As DCPS07 note, their conclusions were reached '*based on essentially the same data*' used in earlier work.

DCPS07 interpret their results as evidence that models are seriously flawed, and that model-based projections of future climate change are unreliable. Singer (2008) makes an additional and even stronger assertion: that the information presented in DCPS07 '*clearly falsifies the hypothesis of anthropogenic greenhouse warming*'.

If such claims were correct, they would have significant scientific implications. It is therefore of interest to examine (as we do here) the *'robust statistical test'* that DCPS07 rely on in order to reach the conclusion that models are inconsistent with observations. We also evaluate other formal statistical tests of the significance of modelled and observed temperature trend differences. We use a variety of different observational datasets, which enables us to explore the sensitivity of our results to current 'structural uncertainties' in observed estimates of surface and tropospheric temperature change.

The structure of our article is as follows. In Section 2, we introduce the observational and model tropospheric temperature datasets analysed here. Section 3 covers basic statistical issues that arise in comparisons of modelled and observed trends. Section 4 describes various tests (among them the DCPS07 test) of the formal statistical significance of trend differences. Results obtained after applying these tests to model and observational data are discussed in Section 5. Test behaviour with synthetic data is considered in Section 6. This is followed by a comparison of vertical profiles of temperature change in climate models and radiosonde data in Section 7. A summary and the conclusions are given in Section 8. Appendix 1 summarizes the statistical notation used in the article, and Appendix 2 provides detailed technical notes on various aspects of the data used, analysis methods, and results.

2. Observational and model temperature data

2.1. Observational data

2.1.1. Satellite data

Since late 1978, atmospheric temperatures have been monitored routinely from space by the Microwave Sounding Units (MSU) and Advanced Microwave Sounding Units (AMSU) flown on NOAA polar-orbiting satellites. Both instruments measure the microwave emissions of

DOI: 10.1002/joc

oxygen molecules, which are roughly proportional to atmospheric temperature (Spencer and Christy, 1990). By measuring emissions at different frequencies, it is possible to retrieve the temperatures of different atmospheric layers. Most scientific attention has focused on MSU-derived temperatures for the lower stratosphere (T_4), the mid-troposphere to lower stratosphere (T_2), and the lower to mid-troposphere (T_{2LT}). The bulk (90%) of the emissions contributing to these temperatures occurs between roughly 14–29 km for T_4, the surface to 18 km for T_2, and the surface to 8 km for T_{2LT} (Karl *et al*., 2006).

To date, four different groups have been actively involved in the development of multi-decadal temperature records from MSU data. These groups are based at the University of Alabama at Huntsville (UAH; Spencer and Christy, 1990; Christy *et al*., 2007), Remote Sensing Systems in Santa Rosa, California (RSS; Mears *et al*., 2003; Mears and Wentz, 2005), the University of Maryland (UMd; Vinnikov and Grody, 2003; Vinnikov *et al*., 2006), and the NOAA National Environmental Satellite, Data, and Information Service (NOAA/NESDIS; Zou *et al*., 2006). All four groups have made different choices in the complex process of adjusting raw MSU and AMSU data for inhomogeneities. This leads to structural uncertainties in tropical tropospheric temperature trends that are at least as large as 0.14 °C/decade for T_2 and 0.10 °C/decade for T_{2LT} (Lanzante *et al*., 2006).[1]

Our interest here is primarily in the T_2 and T_{2LT} data produced by UAH and RSS.[2] Data from both groups are employed in the DCPS07 consistency test between modelled and observed trends. We use results from version 3.0 of the RSS data and versions 5.1 and 5.2 (respectively) of the UAH T_2 and T_{2LT} data.[3] Data were available in the form of gridded, monthly mean products for the period January 1979 through December 2007.

2.1.2. Radiosonde data

DCPS07 compared model-simulated profiles of atmospheric temperature change with vertical profiles estimated from radiosondes. We perform a similar comparison in Section 7. Like DCSP07, we rely on radiosonde datasets produced by the U.K. Meteorological Office Hadley Centre (HadAT2; Thorne *et al*., 2005b; McCarthy *et al*., 2008), NOAA (RATPAC-A; 'Radiosonde Atmospheric Temperature Products for Assessing Climate'; Free *et al*., 2005), and the University of Vienna (RAOBCORE version 1.2; 'Radiosonde Observation Correction using Reanalysis'; Haimberger, 2007). For the latter dataset, information from the ERA-40 reanalysis (Uppala *et al*., 2005) was used to identify and adjust for inhomogeneities in the radiosonde data assimilated by the reanalysis model. HadAT2 and RATPAC-A do not utilize reanalysis information in adjusting for inhomogeneities.

We also analyse four newly-developed radiosonde datasets that were not considered by DCPS07. The first two (RAOBCORE v1.3 and v1.4; Haimberger *et al*., 2008) are more recent versions of the RAOBCORE dataset used by DCPS07. The third (RICH; 'Radiosonde Innovation Composite Homogenization') uses a new automatic data homogenization method involving information from both reanalysis and composites of neighbouring radiosonde stations (Haimberger *et al*., 2008). The fourth (IUK; 'Iterative Universal Kriging') employs an iterative approach to fit the raw radiosonde data to a statistical model of natural climate variability plus step changes associated with instrumental biases (Sherwood, 2007; Sherwood *et al*., 2008). As will be shown later, all four newer radiosonde datasets exhibit larger warming of the tropical lower troposphere than the datasets selected by DCPS07.

2.1.3. Surface data

Comparisons of surface and tropospheric warming trends provide a simple measure of the changes in temperature lapse rates (Gaffen *et al*., 2000). Here, we use four different surface temperature datasets to estimate changes in lower tropospheric lapse rates in the deep tropics. The first three datasets contain information on sea surface temperatures (SST) only (T_{SST}), while the fourth dataset is a blend of 2-m temperatures over Land plus Ocean SSTs (T_{L+O}). The three SST datasets are more appropriate to analyse in order to determine whether observed lower tropospheric temperature changes follow a moist adiabatic lapse rate (Wentz and Schabel, 2000).

The three SST datasets are spatially complete, and rely on statistical procedures to 'infill' SST information in data-sparse regions. The first dataset, HadISST1, was developed at the U.K. Meteorological Office Hadley Centre (Rayner *et al*., 2003). SSTs were reconstructed from *in situ* observations using an optimal interpolation procedure, with subsequent '*superposition of quality-improved gridded observations onto the reconstructions to restore local detail*' (see http://www.hadobs.org/). The other two SST products are versions 2 and 3 of the NOAA ERSST ('Extended Reconstructed SST') dataset developed at the National Climatic Data Center (NCDC; Smith and Reynolds, 2005; Smith *et al*., 2008). Differences between ERSST-v2 and ERSST-v3 are primarily related to differences in treatment of low-frequency variability and to the inclusion of bias-adjusted satellite infrared data in ERSST-v3. The newer dataset is regarded as '*an improved extended reconstruction over version 2*' (see http://www.ncdc.noaa.gov/oa/climate/research/sst/ersstv3.php).

The fourth dataset, HadCRUT3v, consists of a blend of land 2-m temperatures from the Climatic Research Unit's CRUTEM3 dataset (Brohan *et al*., 2006) and SSTs from the Hadley Centre HadSST2 product (Rayner *et al*., 2006). Unlike the SST datasets described above, HadCRUT3v is not spatially complete. Calculation of lapse-rate changes with HadCRUT3v facilitates comparison with previous work by Santer *et al*. (2005, 2006) and DCPS07, which also relied on surface datasets comprised of combined SSTs and land 2-m temperatures.

DOI: 10.1002/joc

2.2. Model data

A number of different climate model experiments were performed in support of the IPCC Fourth Assessment Report (IPCC, 2007). In the experiment of most interest here, nearly two dozen different climate models were forced with estimates of historical changes in both anthropogenic and natural external factors.[4]

These so-called twentieth-century (20CEN) simulations are the most appropriate runs for direct comparison with satellite and radiosonde data, and provide valuable information on current structural and statistical uncertainties in model-based estimates of historical climate change. Inter-model differences in 20CEN results reflect differences in model physics, dynamics, parameterizations of sub-grid scale processes, horizontal and vertical resolution, and the applied forcings (Santer *et al*., 2005, 2006).

Santer *et al*. (2005) examined a set of 49 simulations of twentieth century climate performed with 19 different models. The same suite of runs is analysed here.[5] Santer *et al*. (2005) were primarily concerned with comparisons of modelled and observed amplification of surface warming in the tropical troposphere,[6] while the focus of the present work is on testing the significance of trend differences.

To facilitate the comparison of simulated and observed tropospheric temperature trends, we calculate synthetic MSU T_2 and T_{2LT} temperatures from gridded, monthly-mean model data using a static global-mean weighting function. For temperature changes averaged over large areas, this procedure yields results similar to those estimated with a full radiative transfer code (Santer *et al*., 1999). Since most of the 20CEN experiments end in 1999, our trend comparisons primarily cover the 252-month period from January 1979 to December 1999, which is the period of maximum overlap between the observed MSU data and the model simulations.

3. Basic statistical issues

We assume a simulated tropospheric temperature time series $y_m(t)$ of the form:

$$y_m(t) = \phi_m(t) + \eta_m(t) \tag{1}$$

where $\phi_m(t)$ is the underlying signal in response to external forcing, $\eta_m(t)$ is a specific realization of natural internal climate variability superimposed on $\phi_m(t)$, t is a nominal index of time in months, and the subscript m denotes model data. The corresponding observed time series $y_o(t)$ is given by:

$$y_o(t) = \phi_o(t) + \eta_o(t) \tag{2}$$

The slopes of the least-squares linear trends in these time series (b_m and b_o) provide one measure of overall change in temperature. Estimates of b_m and b_o are sensitive to the behaviour of both signal and noise components in the time series.

In the tropics, the El Niño/Southern Oscillation (ENSO) phenomenon explains most of the year-to-year variability in observed tropospheric temperatures. The real world provides only one sample of how ENSO and other modes of internal climate variability influence atmospheric temperature. This makes it difficult to achieve an unambiguous separation of signal from noise in observational data. Models, however, can be run many times to generate many different realizations of historical climate change,[7] thus facilitating the separation of $\phi_m(t)$ from $\eta_m(t)$. Since $\eta_m(t)$ is uncorrelated from one realization to the next, averaging over many realizations reduces noise levels and improves estimates of any overall trend in $\phi_m(t)$.

This is clearly illustrated in Figure 1A–E, which shows tropical T_{2LT} changes over 1979–1999 in five 20CEN realizations performed with the Japanese Meteorological Research Institute (MRI) model. The character of $\eta_m(t)$ is different in each realization, resulting in a large range of trends in $y_m(t)$ (from 0.042 to 0.371 °C/decade). The small overall trend in the first realization is partly due to the chance occurrence of El Niños near the beginning and middle of the time series, and the presence of a La Niña at the end. Averaging over these five realizations reduces the amplitude of $\eta_m(t)$, and improves the estimate of the true forced change in $y_m(t)$ (Figure 1F). The key point to note is that the same MRI model, with exactly the same physics and forcings, produces a range of self-consistent estimates of tropical T_{2LT} trends over a particular time interval, not a single discrete value. Many other models with ensembles of 20CEN runs also show substantial inter-realization trend differences (see Section 5.1.1).

A number of factors may contribute to differences between modelled and observed temperature trends. These include:

1. Missing or inaccurately specified values of the external forcings applied in the model 20CEN run.
2. Errors in $\phi_m(t)$, the model's response to the imposed forcing changes.
3. Errors in the variability and other statistical properties of $\eta_m(t)$.
4. The irreproducibility of the specific, essentially random sequence of observed noise, even by a model which correctly simulates the statistical properties of $\eta_o(t)$.
5. The number of 20CEN realizations for any given model, which influences how well we can estimate $\phi_m(t)$. If many individual realizations of $y_m(t)$ were available, the model's ensemble-mean trend would provide an accurate estimate of the forced component of change in $y_m(t)$.
6. Residual inhomogeneities in $y_o(t)$.

Even in a model with no errors in forcing, response, or internally generated variability, there could by chance be

DOI: 10.1002/joc

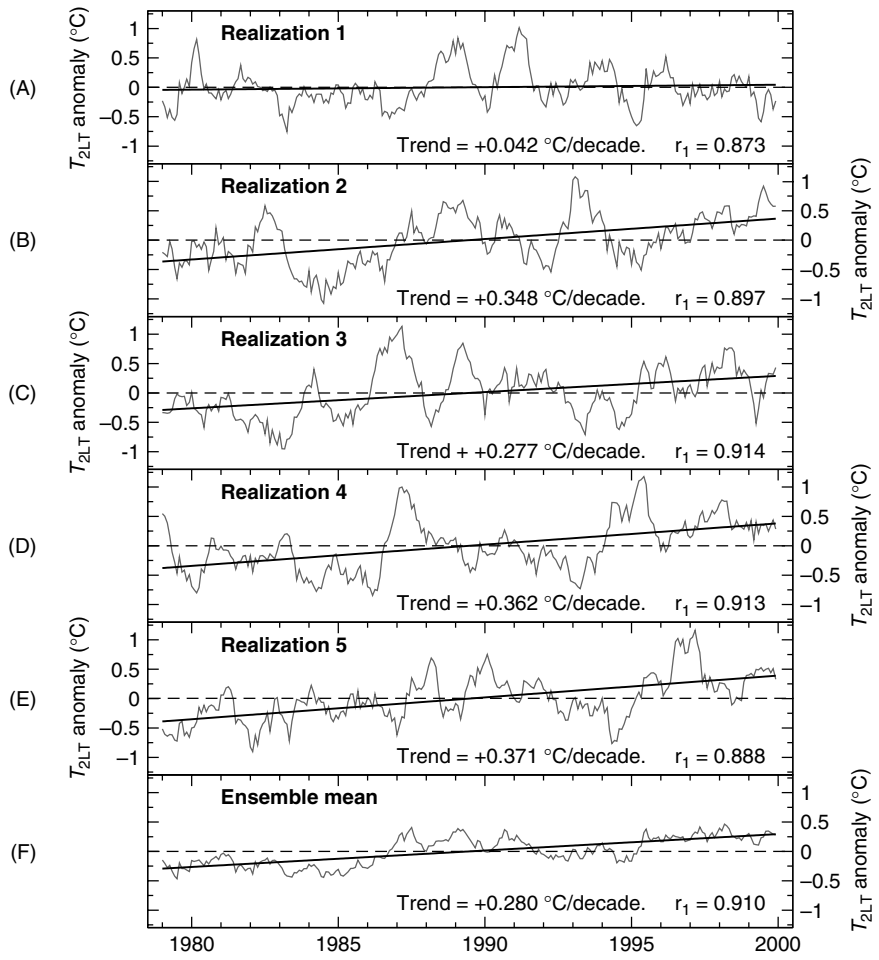

Figure 1. Anomaly time series of monthly-mean T_{2LT}, the spatial average of lower tropospheric temperature over tropical (20 °N–20 °S) land and ocean areas. Results are for five different realizations of 20CEN climate change performed with a coupled A/OGCM (the MRI-CGCM2.3.2). Each of the five realizations (panels A–E) was generated with the same model and the same external forcings, but with initialization from a different state of the coupled atmosphere-ocean system. This yields five different realizations of internally generated variability, $\eta_m(t)$, which are superimposed on the true response to the applied external forcings. The ensemble-mean T_{2LT} change is shown in panel F. Least-squares linear trends were fitted to all time series; values of the trend and lag-1 autocorrelation of the regression residuals (r_1) are given in each panel. Anomalies are defined relative to climatological monthly means over January 1979 to December 1999, and synthetic T_{2LT} temperatures were calculated as described in Santer *et al.* (1999).

realizations of noise that differed markedly from that in the real world, leading to a large difference between modelled and observed trends that was completely unrelated to model error. Any procedure for testing the significance of differences between simulated and observed trends must therefore account for the (potentially different) effects of internally generated variability on b_m and b_o.

4. Significance tests

Our significance testing strategy addresses two different questions. The first is whether models can simulate individual temperature trends that are consistent with the single observed trend. The second question is whether our current best estimate of the model response to external forcing is consistent with our estimate of the externally forced temperature trend in observations.

Each question involves testing a different hypothesis. In the first question, we are testing hypothesis H_1 that the trend in any given realization of $y_m(t)$ is consistent with the trend in $y_o(t)$. As noted previously, interannual climate noise makes it difficult to obtain reliable estimates of the forced components of temperature change [$\phi_o(t)$ and $\phi_m(t)$] from the single $y_o(t)$ time series and from any individual realization of $y_m(t)$. Under hypothesis H_1, therefore, we are comparing trends arising from a combination of forced and unforced temperature changes.

The hypothesis H_2 tested in the second question involves the multi-model ensemble-mean trend. Averaging over realizations and models reduces noise and provides a better estimate of the true model signal in response to external forcing. Under H_2, we seek to determine whether the model-average signal is consistent with the trend in $\phi_o(t)$ (the signal contained in the observations).

4.1. Tests with individual model realizations

To examine H_1, we apply a 'paired trends' test (Santer *et al.*, 2000b; Lanzante, 2005), in which b_o is tested

DOI: 10.1002/joc

against each of the 49 individual b_m trends considered here. The test statistic is of the form:

$$d = (b_m - b_o) / \sqrt{s\{b_m\}^2 + s\{b_o\}^2} \quad (3)$$

where d is the normalized difference between the trends in any two modelled and observed time series, and $s\{b_m\}$ and $s\{b_o\}$ are (respectively) the standard errors of b_m and b_o. The standard errors are measures of the inherent statistical uncertainty in fitting a linear trend to noisy data. For the model data, $s\{b_m\}$ is defined as:

$$s\{b_m\} = \left[s_e^2 / \sum_{t=1}^{n_t} (t - \bar{t})^2 \right]^{\frac{1}{2}} \quad (4)$$

where t is the time index, $\bar{t}$ is the average time index, n_t is the total number of time samples (252 here), and s_e^2 is the variance of the regression residuals, given by:

$$s_e^2 = \frac{1}{n_t - 2} \sum_{t=1}^{n_t} e(t)^2 \quad (5)$$

(see Wilks, 1995). Note that the observed standard error, $s\{b_o\}$, is calculated similarly, but using observational rather than model data.

Assuming that d has a Normal distribution, we can compute its associated p-value and test whether the trend in $y_m(t)$ is consistent with the trend in $y_o(t)$. This test is two-tailed, since we have no expectation *a priori* regarding the direction of the trend difference.

In the case of most atmospheric temperature series, the regression residuals $e(t)$ are not statistically independent. For RSS tropical T_{2LT} data, for example (Figure 2A), values of $e(t)$ have pronounced month-to-month and year-to-year persistence, with a lag-1 temporal autocorrelation coefficient of $r_1 = 0.884$ (Table I). This persistence reduces the number of statistically independent time samples. Following Santer *et al.* (2000a), we account for the non-independence of $e(t)$ values by calculating an effective sample size n_e:

$$n_e = n_t \frac{1 - r_1}{1 + r_1} \quad (6)$$

By substituting $n_e - 2$ for $n_t - 2$ in Equation (5), the standard error can be adjusted for the effects of temporal autocorrelation (see Supporting Information). In the RSS example in Figure 2A, $n_e \approx 16$, and the adjusted standard error is over four times larger than the unadjusted standard error (Figure 2C). The unadjusted standard error should only be used if the regression residuals are uncorrelated. In the case of the synthetic data in Figure 2B, for example, r_1 is close to zero, n_e and n_t are of similar size (236 and 252), and the adjusted and unadjusted standard errors are small and virtually identical (Figure 2C). Our subsequent discussion of the paired trend test (Section 5) deals exclusively with results computed correctly with adjusted standard errors rather than with unadjusted standard errors.

The underlying assumption in our method of adjusting standard errors is that the temporal persistence of $e(t)$ can be well represented by a lag-1 autoregressive (AR)

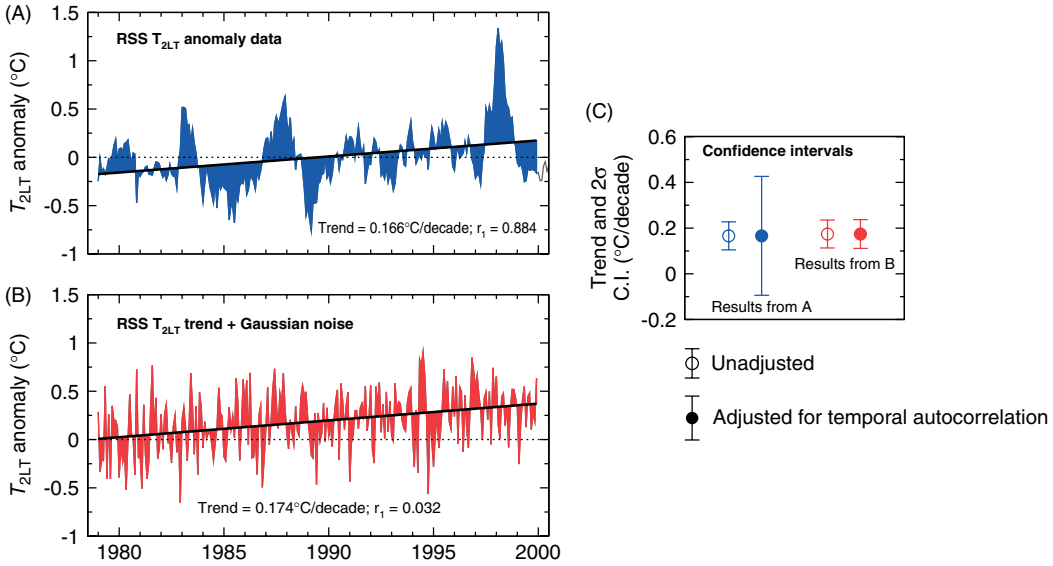

Figure 2. Calculation of unadjusted and adjusted standard errors for least-squares linear trends. The standard error $s\{b_o\}$ of the least-squares linear trend b_o (see Section 4.1) is a measure of the uncertainty inherent in fitting a linear trend to noisy data. Two examples are given here. Panel A shows observed tropical T_{2LT} anomalies from the RSS group (Mears and Wentz, 2005). The regression residuals (shaded blue) are highly autocorrelated ($r_1 = 0.884$). Accounting for this temporal autocorrelation reduces the number of effectively independent time samples from 252 to 16, and inflates $s\{b_o\}$ by a factor of 4 (see 'Results from A' in panel C). The anomalies in panel B were generated by adding Gaussian noise to the RSS tropical T_{2LT} trend, yielding a trend and temporal standard deviation that are very similar to those of the actual RSS data. For this synthetic data series, the regression residuals (shaded red) are uncorrelated and r_1 is close to zero, so that the actual number of time samples is similar to the effective sample size, and the unadjusted and adjusted standard errors are small and virtually identical (see 'Results from B' in panel C). All results in panel C are 2σ confidence intervals (CI). The analysis period is January 1979 to December 1999.

DOI: 10.1002/joc

Table I. Statistics for observed and simulated time series of land and ocean surface temperatures, SST, and tropospheric temperatures.

Dataset	Trend	1σ S.E.	S.D.	r_1	n_e
HadCRUT3v T_{L+O}	0.119	0.117	0.197	0.934	8.6
Multi-model mean T_{L+O}	0.146	0.214	0.274	0.915	11.7
Inter-model S.D. T_{L+O}	0.066	0.163	0.093	0.087	13.9
HadISST1 T_{SST}	0.108	0.133	0.197	0.944	7.3
ERSST-v2 T_{SST}	0.100	0.131	0.186	0.947	6.9
ERSST-v3 T_{SST}	0.077	0.121	0.190	0.936	8.3
Multi-model mean T_{SST}	0.130	0.333	0.243	0.959	5.3
Inter-model S.D. T_{SST}	0.062	0.336	0.084	0.024	3.2
UAH T_{2LT}	0.060	0.138	0.299	0.891	14.5
RSS T_{2LT}	0.166	0.132	0.312	0.884	15.6
Multi-model mean T_{2LT}	0.215	0.198	0.376	0.876	17.2
Inter-model S.D. T_{2LT}	0.092	0.133	0.127	0.080	12.2
UAH T_2	0.043	0.129	0.306	0.873	17.1
RSS T_2	0.142	0.129	0.319	0.871	17.3
Multi-model mean T_2	0.199	0.181	0.370	0.855	20.3
Inter-model S.D. T_2	0.098	0.133	0.132	0.085	13.0

Results are for time series of monthly mean anomalies in land and ocean surface temperature (T_{L+O}), sea surface temperature (T_{SST}), and tropospheric temperature (T_{2LT}, T_2). Analyses are over the 252-month period from January 1979 through December 1999 (the period of maximum overlap between the observations and most model 20CEN experiments). Gridded anomaly data were spatially averaged over 20°N–20°S. The time series statistics are the least-squares linear trend (b_o, b_m; °C/decade); the standard error of the linear trend, adjusted for temporal autocorrelation effects ($s\{b_o\}$, $s\{b_m\}$; °C/decade); the temporal standard deviation of the anomaly data ($s\{y_o(t)\}$, $s\{y_m(t)\}$; °C); the lag-1 autocorrelation of the regression residuals (r_1); and the effective number of independent time samples (n_e). The multi-model mean and inter-model standard deviation were calculated using the ensemble-mean values of the time series statistics for the 19 models [see Equations (7)–(9)]. Anomalies were defined relative to climatological monthly means computed over the analysis period. For sources of model and observed data, see Section 2.

statistical model. This assumption is not uncommon in meteorological applications (e.g. Wilks, 1995; Lanzante *et al.*, 2006). If the autocorrelation structure is more complex and exhibits long-range dependence, it may be more appropriate to use higher-order AR models for estimating n_e (Thiébaux and Zwiers, 1984). However, it is difficult to reliably estimate the parameters of such statistical models given the relatively short length (20–30 years) and high temporal autocorrelation of the temperature data available here.

Experiments with synthetic data reveal that the use of an AR-1 model for calculating n_e tends to overestimate the true effective sample size (Zwiers and von Storch, 1995). This means that our d test is too liberal, and is more likely to indicate that there are significant differences between modelled and observed trends, even when significant differences do not actually exist.[8] It should therefore be easier for us to confirm DCPS07's finding that modelled and observed trends are inconsistent. As described in Section 5, however, our results do not confirm DCPS07's findings. DCPS07's conclusions are erroneous, and are primarily due to the neglect of observed trend uncertainties in their statistical test (see Section 4.2).

4.2. Tests with multi-model ensemble-mean trend

Here we examine two different tests of the hypothesis H_2 (see Section 4). Both rely on the multi-model ensemble-mean trend,[9] $\ll b_m \gg$:

$$\ll b_m \gg = \frac{1}{n_m} \sum_{i=1}^{n_m} < b_m(i) > \tag{7}$$

where $< b_m(i) >$ is the ensemble-mean trend in the i^{th} model:

$$< b_m(i) > = \frac{1}{n_r(i)} \sum_{j=1}^{n_r(i)} b_m(i, j) \; ; \; i = 1, \ldots, n_m \tag{8}$$

The indices i and j are over model number and realization number (respectively). The total number of models is n_m (19 here), and $n_r(i)$ is the total number of 20CEN realizations for the i^{th} model (which varies from 1 to 5). The standard deviation of ensemble-mean trends, $s\{< b_m >\}$, is given by

$$s\{< b_m >\} = \left[\frac{1}{n_m - 1} \sum_{i=1}^{n_m} (< b_m(i) > - \ll b_m \gg)^2 \right]^{1/2} \tag{9}$$

In the DCPS07 'consistency test', the difference between $\ll b_m \gg$ and b_o is compared with σ_{SE}, '*an estimate of the uncertainty of the* (multi-model) *mean* (trend)'. DCPS07 do not consider any uncertainty in b_o, and σ_{SE} is based solely on the inter-model variability of trends:

$$\sigma_{SE} = s\{< b_m >\} / \sqrt{n_m} \tag{10}$$

To evaluate the performance of the DCPS07 test, we define the test statistic d^*:

$$d^* = (\ll b_m \gg - b_o) / \sigma_{SE} \tag{11}$$

If the DCPS07 test were valid, a large value of d^* would imply a significant difference between $\ll b_m \gg$ and b_o. However, the test is not valid. There are a number of reasons for this:

1. DCPS07 ignore the pronounced influence of interannual variability on the observed trend (see Figure 2A). They make the implicit (and incorrect) assumption that the externally forced component in the observations is perfectly known (i.e. the observed record consists only of $\phi_o(t)$, and $\eta_o(t) = 0$).
2. DCPS07 ignore the effects of interannual variability on model trends – an effect which we consider in our 'paired trends' test [see Equation (3)]. They incorrectly assume that the forced component of temperature change is perfectly known in each individual

DOI: 10.1002/joc

model (*i.e.* each individual 20CEN realization consists only of $\phi_m(t)$, and $\eta_m(t) = 0$).[10]

3. DCPS07's use of σ_{SE} is incorrect. While σ_{SE} is an appropriate measure of how well the multi-model mean trend can be estimated from a finite sample of model results, it is not an appropriate measure for deciding whether this trend is consistent with a single observed trend.

Practical consequences of these problems are discussed later in Sections 5 and 6.

We can easily modify the DCPS07 d^* test to account for the factor neglected by DCPS07 – the effects of interannual variability on the 'trend signal' in $y_o(t)$. The resulting d_1^* test is similar in form to a t-test of the difference in means:

$$d_1^* = (\ll b_m \gg - b_o) \Big/ \sqrt{\frac{1}{n_m} s\{<b_m>\}^2 + s\{b_o\}^2} \tag{12}$$

where the term $\frac{1}{n_m} s\{<b_m>\}^2$ is a standard estimate of the variance of the mean (in this case, the variance of the model-average trend $\ll b_m \gg$; see Storch and Zwiers, 1999), and $s\{b_o\}^2$ is an estimate of the variance of the observed trend b_o [see Equations (4)–(6)].

There are three underlying assumptions in the d_1^* test. The first assumption (which was also made by DCPS07) is that the uncertainty in $\ll b_m \gg$ is entirely due to inter-model differences in forcing and response, and not due to differences in variability and ensemble size. The second assumption is that the uncertainties in the observed trend are due solely to the effects of interannual variability – i.e. there are no residual errors in the observations being tested. The third assumption is that d_1^* has a Student's t distribution, and that the number of degrees of freedom associated with the estimated variances of $\ll b_m \gg$ and b_o are $n_m - 1$ and $n_e - 2$, respectively.

As noted above, the variances of $\ll b_m \gg$ and b_o are influenced by very different factors, and are unlikely to be identical. In this case, the degrees of freedom for the $\{d_1^*\}$ test, DOF$\{d_1{}^*\}$ are approximated by:

$$\text{DOF}\{d_1^*\} = \left[1/n_m\, s\{<b_m>\}^2 + s\{b_o\}^2\right]^2 \Big/ \frac{\left[1/n_m\, s\{<b_m>\}^2\right]^2}{n_m - 1} + \frac{\left[s\{b_o\}^2\right]^2}{n_e - 2} \tag{13}$$

(see Storch and Zwiers, 1999). We will demonstrate in Section 6 that d_1^* and the DCPS07 d^* test exhibit very different behaviour when applied to synthetic data.

5. Results of significance tests

5.1. Tropospheric temperature trends

5.1.1. Tests with individual model realizations

Figure 3A shows trends in tropical T_{2LT} in the two satellite datasets (RSS and UAH) and in 49 realizations of the 20CEN experiment, together with their adjusted 2σ confidence intervals. Values of b_m vary substantially, not only between models but also within the different 20CEN realizations of individual models. The adjusted 2σ confidence interval on the RSS T_{2LT} trend includes 47 of the 49 simulated trends. This strongly suggests that there is no fundamental inconsistency between modelled and observed trends.[11]

Results from the paired trends test [see Equation (3)] are summarized in Table II. For each of the two layer-averaged temperatures considered here (T_{2LT} and T_2), UAH and RSS trends were tested against trends from the 49 individual model simulations. Calculated p-values for the d statistic were compared with stipulated p-values of 0.05, 0.10, and 0.20. We then determined the number of tests in which hypothesis H_1 (see Section 4) is rejected at the 5, 10, and 20% significance levels.

If model and observed trends were in perfect agreement, we would still expect (for a very large number of tests) p% of the tests to show significant trend differences at the p% significance level. Our rejection rates are invariably lower than the theoretical expectation (Table II). There are at least four possible explanations for this:

1. Not all 49 tests are statistically independent.
2. Tests are affected by differences between modelled and observed variability.
3. Results are influenced by the sampling variability arising from the relatively small number of tests performed.
4. Our method of adjusting standard errors for temporal autocorrelation effects is not reliable.[12]

Overall, however, our paired test results show broad agreement between tropospheric temperature trends estimated from models and satellite data. This consistency

Table II. Significance of differences between modelled and observed tropospheric temperature trends: Results for paired trends tests.

Sig. level (%)	RSS T_{2LT} (%)	UAH T_{2LT} (%)	RSS T_2 (%)	UAH T_2 (%)
5	0 (0.0)	1 (2.0)	1 (2.0)	1 (2.0)
10	1 (2.0)	1 (2.0)	1 (2.0)	3 (6.1)
20	1 (2.0)	4 (8.2)	1 (2.0)	6 (12.2)

Results are for the paired trends test described in Section 4.1. Model data employed in the test are tropical T_{2LT} and T_2 trends from 49 realizations of twentieth-century climate change performed with 19 different A/OGCMs (together with their associated adjusted standard errors). Observational trends and adjusted standard errors were estimated from RSS and UAH satellite data. There are 49 tests for each tropospheric layer and each observational dataset. Results are expressed as the number of rejections of hypothesis H_1 (see Section 4) at stipulated significance levels of 5, 10, and 20%. Percentage rejection rates of H_1 (out of 49 tests) are given in parentheses. All trends and standard errors were calculated over the period January 1979 to December 1999 from time series of spatially averaged (20°N–20°S) anomaly data.

DOI: 10.1002/joc

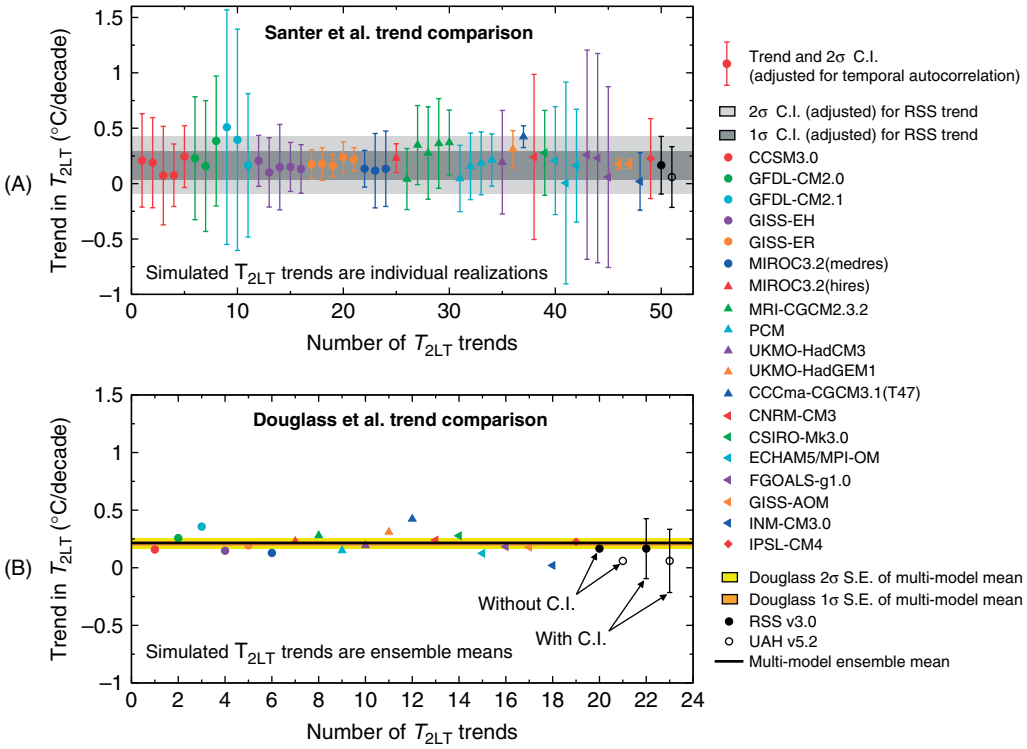

Figure 3. Comparisons of simulated and observed trends in tropical T_{2LT} over January 1979 to December 1999. Model results in panel A are from 49 individual realizations of experiments with 20CEN external forcings, performed with 19 different A/OGCMs. Observational estimates of T_{2LT} trends are from Mears and Wentz (2005) and Christy *et al.* (2007) for RSS and UAH data, respectively. The dark and light grey bands in panel A are the 1σ and 2σ confidence intervals for the RSS T_{2LT} trend, adjusted for temporal autocorrelation effects. In the paired trends test applied here, each individual model T_{2LT} trend is tested against each observational T_{2LT} trend (Section 4.1). Panel B shows the three elements of the DCPS07 'consistency test': the multi-model ensemble mean T_{2LT} trend, $\ll b_m \gg$ (represented by the horizontal black line in panel B); σ_{SE}, DCPS07's estimate of the uncertainty in $\ll b_m \gg$; and b_o, the individual RSS and UAH T_{2LT} trends (with and without their 2σ confidence intervals from panel A). The 1σ and 2σ values of σ_{SE} are indicated by orange and yellow bands, respectively. The coloured dots in panel B are either the ensemble-mean T_{2LT} trends for individual models or the trend in an individual 20CEN realization (for models that did not perform multiple 20CEN realizations). Statistical uncertainties in the observed trends are neglected in the DCSP07 test. If these uncertainties are accounted for, $\ll b_m \gg$ is well within the 2σ confidence intervals on the RSS and UAH T_{2LT} trends (Section 5.1.2).

holds even if we account for errors in model variability (see Supporting Information).

5.1.2. *Tests with multi-model ensemble-mean trend*

We now seek to understand why DCPS07 concluded that the multi-model ensemble-mean trend was inconsistent with observed trends, despite the fact that almost all the individual b_m trends are consistent with observations (see Section 5.1.1).

Application of the DCPS07 test yields values of the test statistic d^* [see Equation (11)] ranging from 2.25 for RSS T_{2LT} trends to 7.16 for UAH T_{2LT} trends (Table III). In all four d^* tests,[13] hypothesis H_2 is rejected at the 5% level or better. This is why DCPS07 concluded that the multi-model ensemble-mean trend is inconsistent with observed T_{2LT} and T_2 trends. As will be shown below, this conclusion is erroneous.

It is obvious from Figure 3B and Table I that for T_{2LT} data, $\ll b_m \gg$ lies within the adjusted 2σ confidence intervals for the RSS and UAH trends. As was noted in Section 4.2, however, DCPS07 ignore trend uncertainties arising from interannual variability, both for observational and model trends. If DCPS07 had accounted for these trend uncertainties, they would have obtained very different results.

Table III. Significance of differences between modelled and observed tropospheric temperature trends: Results for tests involving multi-model ensemble-mean trend.

Statistic type	RSS T_{2LT}	UAH T_{2LT}	RSS T_2	UAH T_2
d^*	2.25**	7.16***	2.48**	6.78***
d_1^*	0.37	1.11	0.44	1.19

Results are the actual test statistic values for two different tests of the hypothesis H_2: the original DCPS07 'consistency test' [d^*; see Equation (11)] and a modified version of the DCPS07 test [d_1^*; see Equation (12)]. Both d^* and d_1^* involve the model-average signal trend. The T_{2LT} and T_2 data used in the tests are described in Table II. One, two, and three asterisks indicate model-versus-observed trend differences that are significant at the 10, 5, and 1% levels respectively; (two-tailed tests).

This is evident when we apply our modified version of the DCPS07 test, which accounts for uncertainties in both the observational and model trend signals. For all four tests with d_1^*, hypothesis H_2 cannot be rejected at the

DOI: 10.1002/joc

nominal 5% level (Table III). These findings differ radically from those obtained with DCPS07's 'consistency test'. We conclude, therefore, that when uncertainties in both observational and model trend signals are accounted for, there is no statistically significant difference between the model-average trend signal and the observed trend in $\phi_o(t)$.

5.2. Trends in lower tropospheric lapse rates

5.2.1. Tests with individual model realizations

Tests involving trends in the surface-minus-T_{2LT} difference series are more stringent than tests of trend differences in T_{L+O}, T_{SST}, or T_{2LT} alone. This is because differencing removes much of the common variability in surface and tropospheric temperatures, thus decreasing both the variance and lag-1 autocorrelation of the regression residuals (Wigley, 2006). In turn, these twin effects increase the effective sample size and decrease the adjusted standard error of the trend, making it easier to identify significant trend differences between models and observations.

Despite these decreases in $s\{b_m\}$ and $s\{b_o\}$, however, 45 of 49 trends in the simulated T_{SST} minus T_{2LT} difference series are still within the $\pm 2\sigma$ confidence intervals of the ERSST-v3 minus RSS difference series trend (Figure 4A). Irrespective of which observational dataset is used for estimating surface temperature changes, each of the three T_{SST} minus T_{2LT} pairs involving RSS data (and the single T_{L+O} minus T_{2LT} pair) has a negative trend in the difference series, indicating larger warming aloft than at the surface, consistent with the model results (Table IV). Application of the paired trends test [Equation (3)] reveals that there are very few statistically significant differences between the model difference series trends and observed lapse-rate trends computed using RSS T_{2LT} data (Table V).

For all four difference series 'pairs' involving UAH T_{2LT} data, the warming aloft is smaller than the warming of the tropical surface, leading to a positive trend in the surface minus T_{2LT} time series – i.e. a trend of opposite sign to virtually all model results (Table IV and Figure 4A). Even in the UAH cases, however, not all models are inconsistent with the observed estimates of 'differential warming' (despite DCPS07's claim to the contrary). Rejection rates for paired trend tests with a stipulated 5% significance level range from 31 to 88%, depending on the choice of observed surface record (Table V). The highest rejection rates are for lapse-rate trends computed with the HadCRUT3v surface data, which has the largest surface warming.

5.2.2. Tests with the multi-model ensemble-mean trend

Figure 4B shows that the multi-model ensemble-mean difference series trend is very close to the trend in the ERSST-v3 minus RSS difference series. In this specific case, even the incorrect, unmodified DCPS07 test yields a non-significant value of d^* (0.49; see Table VI). In seven of the other eight difference series pairs, however,

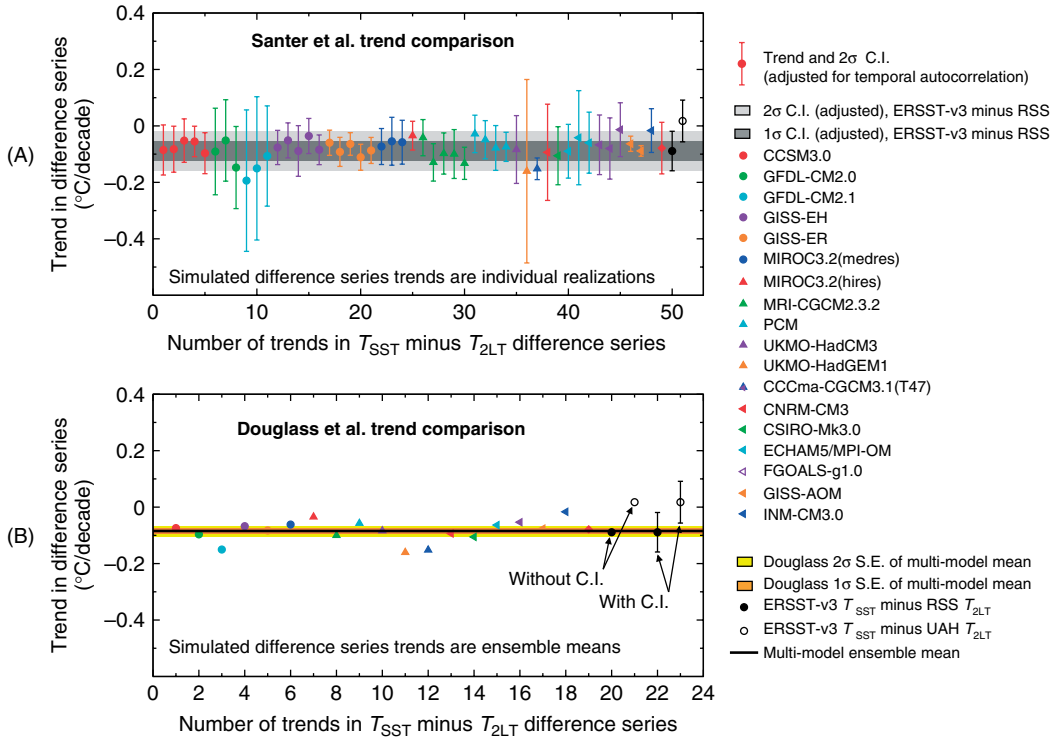

Figure 4. Same as Figure 3 but for the comparisons of simulated and observed trends in the time series of differences between tropical T_{SST} and T_{2LT}. The observed T_{SST} data are from NOAA ERSST-v3 (Smith *et al.*, 2008). For trends and confidence intervals from other observed pairs of surface and T_{2LT} data, refer to Table IV.

DOI: 10.1002/joc

Table IV. Statistics for observed and simulated time series of differences between tropical surface temperature and lower tropospheric temperature.

Dataset	Trend	1σ S.E.	Std. dev.	r_1	n_e
HadCRUT3v T_{L+O} minus UAH T_{2LT}	0.061	0.036	0.165	0.642	55.0
HadCRUT3v T_{L+O} minus RSS T_{2LT}	−0.046	0.034	0.162	0.608	61.5
Multi-model mean T_{L+O} minus T_{2LT}	−0.069	0.040	0.164	0.614	62.5
Inter-model S.D. T_{L+O} minus T_{2LT}	0.032	0.031	0.057	0.137	27.3
HadISST1 T_{SST} minus UAH T_{2LT}	0.049	0.037	0.170	0.630	57.2
ERSST-v2 T_{SST} minus UAH T_{2LT}	0.041	0.040	0.172	0.665	50.7
ERSST-v3 T_{SST} minus UAH T_{2LT}	0.018	0.037	0.167	0.633	56.6
HadISST1 T_{SST} minus RSS T_{2LT}	−0.058	0.035	0.170	0.595	64.0
ERSST-v2 T_{SST} minus RSS T_{2LT}	−0.066	0.038	0.175	0.637	56.0
ERSST-v3 T_{SST} minus RSS T_{2LT}	−0.089	0.035	0.174	0.601	62.7
Multi-model mean T_{SST} minus T_{2LT}	−0.085	0.053	0.197	0.654	55.3
Inter-model S.D. T_{SST} minus T_{2LT}	0.038	0.036	0.064	0.146	28.4

Same as Table I but for basic statistical properties of observed and simulated time series of differences between tropical surface and lower tropospheric temperatures. We use three datasets (HadISST1, ERSST-v2, and ERSST-v3) to characterize observed changes in T_{SST}, one dataset (HadCRUT3v) to describe changes in T_{L+O}, and two datasets (RSS and UAH) to estimate observed changes in tropical T_{2LT}. This yields eight different combinations of observed surface minus T_{2LT} difference series.

Table V. Significance of differences between modelled and observed trends in lower tropospheric lapse rates: Results for paired trends tests.

Dataset pair	5% sig. level (%)	10% sig. level (%)	20% sig. level (%)
HadCRUT3v T_{L+O} minus UAH T_{2LT}	43 (87.8)	45 (91.8)	47 (95.9)
HadISST1 T_{SST} minus UAH T_{2LT}	28 (57.1)	39 (79.6)	44 (89.8)
ERSST-v2 T_{SST} minus UAH T_{2LT}	25 (51.0)	33 (67.4)	44 (89.8)
ERSST-v3 T_{SST} minus UAH T_{2LT}	15 (30.6)	24 (49.0)	35 (71.4)
HadCRUT3v T_{L+O} minus RSS T_{2LT}	1 (2.0)	1 (2.0)	3 (6.1)
HadISST1 T_{SST} minus RSS T_{2LT}	1 (2.0)	2 (4.1)	3 (6.1)
ERSST-v2 T_{SST} minus RSS T_{2LT}	1 (2.0)	1 (2.0)	2 (4.1)
ERSST-v3 T_{SST} minus RSS T_{2LT}	0 (0.0)	0 (0.0)	2 (4.1)

Same as Table II, but for paired tests involving trends in modelled and observed time series of differences between surface and lower tropospheric temperatures in the deep tropics. Trends in T_{SST} minus T_{2LT} and T_{L+O} minus T_{2LT} provide simple measures of changes in lower tropospheric lapse rates. For sources of data, refer to Table IV. Each of the eight observed difference series trends is tested against each of the 49 simulated difference series trends. Results are the number of rejections of hypothesis H_1 and the percentage rejection rates (in parentheses) for three stipulated significance levels. The analysis period and anomaly definition are as for the T_{2LT} and T_2 data described in Table II.

use of the original DCPS07 consistency test leads to rejection of the H_2 hypothesis at the nominal 5% level (see Section 4).

The modified DCPS07 test with d_1^* [see Equation (12)] yields strikingly different results: there is no case in which the model-average signal trend differs significantly from the four pairs of observed surface-minus-T_{2LT} trends calculated with RSS T_{2LT} data (Table VI). When the UAH T_{2LT} data are used to estimate lapse-rate trends, however, H_2 is rejected at the nominal 5% level for all four of the observed surface-minus-T_{2LT} trends. This sensitivity of significance test results to the choice of RSS or UAH T_{2LT} data is qualitatively similar to that obtained for 'paired trends' tests of the H_1 hypothesis (see Section 5.2.1).[14]

5.2.3. *Summary of tests with lower tropospheric lapse rates*

On the basis of these new results, we conclude that considerable scientific progress has been made since the CCSP report, which described '*a potentially serious inconsistency*' between recent modelled and observed trends in tropical lapse rates (Karl *et al*., 2006, p. 11). As described in Sections 5.2.1 and 5.2.2, modelled trends in tropical lapse rates are now broadly consistent with results obtained using RSS T_{2LT} data. Why has this progress occurred?

There are at least two contributory factors. First, the new RSS tropical T_{2LT} trend is over 25% larger than the old trend (0.166 *vs* 0.130 °C/decade), primarily due to a change in RSS's procedure of adjusting for inter-satellite biases. Adjustments now incorporate a latitudinal dependence (as in Christy *et al*., 2003), which tends to increase trends in the tropics and decrease trends at mid-latitudes. Second, our work reveals that comparisons of modelled and observed tropical lapse-rate changes are sensitive to structural uncertainties in the observed SST data, and that these uncertainties may be larger than one would infer from the CCSP report. The tropical SST trends estimated here range from 0.077 to

DOI: 10.1002/joc

Table VI. Significance of differences between modelled and observed trends in lower tropospheric lapse rates: Results for tests involving multi-model ensemble-mean trend.

Dataset pair	d^*	d_1^*
HadCRUT3v T_{L+O} minus UAH T_{2LT}	17.05***	3.50***
HadISST1 T_{SST} minus UAH T_{2LT}	14.94***	3.52***
ERSST-v2 T_{SST} minus UAH T_{2LT}	14.01***	3.04***
ERSST-v3 T_{SST} minus UAH T_{2LT}	11.43***	2.68***
HadCRUT3v T_{L+O} minus RSS T_{2LT}	3.05***	0.67
HadISST1 T_{SST} minus RSS T_{2LT}	3.01***	0.75
ERSST-v2 T_{SST} minus RSS T_{2LT}	2.09**	0.48
ERSST-v3 T_{SST} minus RSS T_{2LT}	0.49	0.12

Same as Table III, but for tests of hypothesis H_2 involving trends in modelled and observed time series of differences between surface and lower tropospheric temperatures in the deep tropics.

0.108 °C/decade (see Table I), with differences primarily related to different processing choices in the treatment of satellite and buoy data and in the applied infilling and filtering procedures (Smith and Reynolds, 2005; Brohan *et al*., 2006; Rayner *et al*., 2006; Smith *et al*., 2008). The smaller observed SST changes in the ERSST-v2 and ERSST-v3 data yield lapse-rate trends that are in better accord with model results. These two SST datasets were not examined in DCPS07 or in the study by Santer *et al*. (2005, 2006).

6. Experiments with synthetic data

The following section compares the performance of d, d^*, and d_1^* under controlled conditions, when the test statistics are applied to synthetic data. We use a standard lag-1 AR model to generate the synthetic time series $x(t)$:

$$x(t) = a_1(x(t-1) - a_m) + z(t) + a_m \; ; \; t = 1, \ldots, n_t \tag{14}$$

where a_1 is the coefficient of the AR-1 model, $z(t)$ is randomly generated white noise, and a_m is a mean term. Here, we set a_1 to 0.87 (close to the lag-1 autocorrelation of the monthly-mean UAH and RSS T_{2LT} and T_2 anomaly data; see Table I), and a_m to zero. The noise $z(t)$ is scaled so that $x(t)$ has approximately the same temporal standard deviation as the UAH anomaly data. Each $x(t)$ series has the same length as the observational and model data (252 months), and monthly-mean anomalies were defined as for $y_m(t)$ and $y_o(t)$.

Rejection rate results for these idealized cases are shown in Figure 5 as a function of N, the number of synthetic time series. Consider first the results for our 'paired trends' test of hypothesis H_1 (see Section 4). For each synthetic time series, we calculated the trend b_x and its unadjusted and adjusted standard errors, and then computed the test statistic d for all unique combinations of time series pairs. In the $N = 19$ case, for example (which corresponds to the number of A/OGCMs used in our study), there are 171 unique pairs. Under the assumption that d has a Normal distribution, we determined rejection rates for H_1 at stipulated significance levels of 5, 10, and 20%. This procedure was repeated 1000 times, with 1000 different realizations of 19 synthetic time series, allowing us to obtain estimates of the parameters of the underlying rejection rate distributions. We followed a similar process for all other values of N considered.

The paired trend results obtained with adjusted standard errors are plotted as blue lines in Figure 5A. The percentage rejections of hypothesis H_1 (averaged over all values of N) are close to the theoretical expectations: the 5, 10, and 20% significance tests have rejection rates of *ca.* 6, 11, and 21%, respectively (see Supporting Information).

This bias of roughly 1% between theoretical and empirically estimated rejection rates is very small compared to the bias that occurs if the paired trends test is applied without adjustment for temporal autocorrelation effects. In the latter case, rejection rates for 5, 10, and 20% tests consistently exceed 60, 65, and 72% respectively; (see green lines in Figure 5A). Clearly, ignoring the influence of temporal autocorrelation on the estimated number of independent time samples yields incorrect test results.

We now examine tests of hypothesis H_2 with the DCPS07 d^* statistic [Equation (11)] and our d_1^* statistic [Equation (12)]. Consider again the example of the $N = 19$ case. The first time series is designated as the 'observations', from which we calculate the trend $b_x(1)$ and its adjusted standard error. With the remaining 18 time series, we compute the ensemble-mean 'model' trend, $< b_x >$, and DCPS07's σ_{SE}. We then calculate the test statistics d^* and d_1^*. This is repeated with the trend in the second time series as surrogate observations, and with $< b_x >$ and σ_{SE} calculated from time series 1, 3, 4, ... 19, *etc.* For each of the two test statistics, our procedure yields 19 separate tests of hypothesis H_2 (see Section 4). As for the paired trends test with synthetic data, we repeat this procedure 1000 times, generate distributions of rejection rates at the three stipulated significance levels, and then repeat the process for all other values of N.

Application of the unmodified DCPS07 test to synthetic data leads to alarmingly large rejection rates of H_2 (Figure 5B; red lines). Rejection rates are a function of N. For 5% significance tests, rejection rates rise from 65 to 84% (for $N = 19$ and $N = 100$, respectively). Although DCPS07 refer to this as a '*robust statistical test*', it is clearly flawed, and is robust only in its ability to incorrectly reject hypothesis H_2. When our modified version of this test is applied to the same synthetic data, results are strikingly different: rejection rates are within 1-2% of the theoretical expectation values (Figure 5B; black lines).

The lesson from this exercise is that DCPS07's consistency test, when applied to synthetic data generated with the same underlying statistical model, yields incorrect results. It finds a very high proportion of significant differences between 'modelled' and 'observed' trends,

DOI: 10.1002/joc

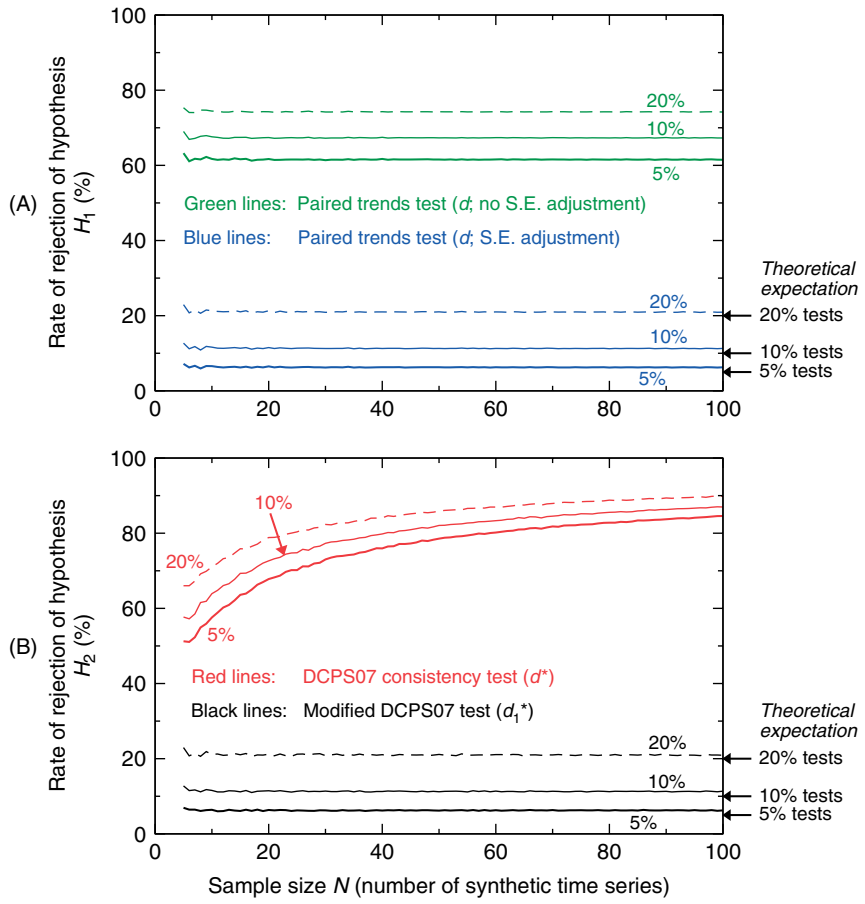

Figure 5. Performance of statistical tests with synthetic data. Results in panel A are for the 'paired trends' test [d; see Equation (3)], in which trends from 'observed' temperature time series are tested against trends from individual realizations of 'model' 20CEN runs. Two versions of the paired trends test are evaluated, with and without adjustment of trend standard errors for temporal autocorrelation effects. Panel B shows results obtained with the DCPS07 'consistency test' [d^*; see Equation (11)] and a modified version of the DCPS07 test [d_1^*; see Equation (12)] which accounts for statistical uncertainties in the observed trend. In the d^* and d_1^* tests, the 'model-average' signal trend is compared with the 'observed' trend. Synthetic $x(t)$ time series were generated using the standard AR-1 model in Equation (14). Rejection rates for hypotheses H_1 (for the 'paired trends' test) and H_2 (for the d^* and d_1^* tests; see Section 4) are given as a function of N, the total number of synthetic time series, for $N = 5, 6, \ldots 100$. Each test is performed for stipulated significance levels of 5, 10, and 20% (denoted by dashed, thin and bold lines, respectively). For each value of N, rejection rates are the mean of the sampling distribution of rejection rates obtained with 1000 realizations of N synthetic time series. The specified value of the lag-1 autocorrelation coefficient in Equation (14) is close to the sample value of r_1 in the UAH and RSS T_{2LT} data (Table I). Similarly, the noise component of the synthetic $x(t)$ data was scaled to ensure $x(t)$ had (on average) approximately the same temporal standard deviation as the observed T_{2LT} anomaly data. See Section 6 for further details..

even in a situation where we know *a priori* that trend differences should occur by chance alone, and that the proportion of tests with significant differences should be small. Although these synthetic data simulations are not an exact analogue of the 'real-world' application of the d^* and d_1^* tests, a test that yields incorrect results under controlled conditions with synthetic data cannot be expected to produce reasonable results in a 'real-world' application.

7. Vertical profiles of atmospheric temperature trends

DCPS07 also use their consistency test to compare simulated vertical profiles of tropical temperature change with results from radiosondes. They conclude that the multi-model ensemble-mean trend profile, $\ll b_m(z) \gg$ (where z is a nominal height coordinate), is inconsistent with the trends inferred from radiosondes. We have shown previously that their test is flawed and yields incorrect results when applied in controlled settings (Sections 5 and 6).

A further concern relates to the observational data used by DCPS07. They rely on radiosonde data from HadAT2 (McCarthy *et al.*, 2008), RATPAC version B (Free *et al.*, 2005),[15] RAOBCORE version 1.2 (Haimberger, 2007), and the Integrated Global Radiosonde Archive ('IGRA'; Durre *et al.*, 2006). DCSP07 claim that these constitute '*the best available updated observations*'. As noted in Section 1, there are large structural uncertainties in radiosonde-based estimates of atmospheric temperature change (see, e.g. Seidel *et al.*, 2004; Thorne *et al.*, 2005b; Mears *et al.*, 2006). An important question, therefore, is whether DCSP07 accurately represented our best currently available estimates of structural uncertainties in radiosonde data.

DOI: 10.1002/joc

To address this question, we first consider the RAOBCORE datasets developed at the University of Vienna (UnV). We use three versions of the RAOBCORE data: v1.2 and v1.3, which were described in Haimberger (2007), and v1.4, which was introduced in Haimberger *et al*. (2008). While RAOBCORE v1.2 shows little net warming of the tropical troposphere over the satellite era, v1.3 and v1.4 exhibit pronounced tropospheric warming, with warming maxima in excess of 0.6 °C/decade at 200 hPa, and cooling of up to 0.1 °C/decade between 700 and 500 hPa (Figure 6A). These large differences in RAOBCORE vertical temperature profiles arise because of different decisions made by the UnV group in the data homogenization process. Although DCPS07 had access to all three RAOBCORE versions, they presented results from v1.2 only.

We also analyse two new radiosonde products, RICH and IUK, which were not available to DCPS07. RICH relies on the same procedure as the RAOBCORE datasets to identify inhomogeneities ('breaks') in radiosonde data. Unlike the RAOBCORE products (which use information from the ERA-40 background forecasts for break adjustment), RICH adjusts for breaks with homogeneous information from nearby radiosonde stations (Haimberger *et al*., 2008). IUK employs a new homogenization procedure in which raw radiosonde data are represented by a model of step-function changes (associated with instrument biases) and natural climate variability (Sherwood, 2007).[16] Both RICH and IUK do not display the prominent lower tropospheric cooling evident in the RAOBCORE, HadAT2, and RATPAC-A products. For comparisons over the period 1979–1999, the multi-model ensemble-mean trend profile in the tropical lower troposphere is closer to the IUK and RICH results than to the changes derived from the other five radiosonde datasets.

The results presented here illustrate that current structural uncertainties in the radiosonde data are substantially larger than one would infer from DCPS07. Different choices in the complex process of dataset construction and homogenization lead to marked differences in both the amplitude and vertical structure of the resulting tropical trends. Temperatures from the most recent homogenization efforts, however, invariably show

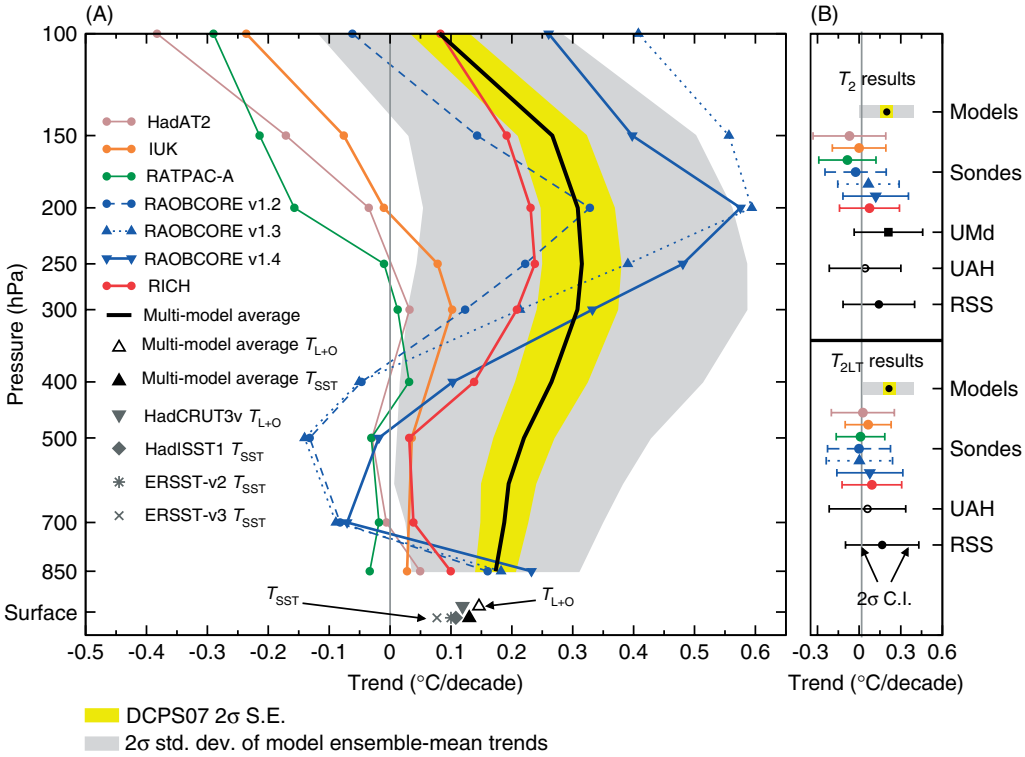

Figure 6. Vertical profiles of trends in atmospheric temperature (panel A) and in actual and synthetic MSU temperatures (panel B). All trends were calculated using monthly-mean anomaly data, spatially averaged over 20 °N–20 °S. Results in panel A are from seven radiosonde datasets (RATPAC-A, RICH, HadAT2, IUK, and three versions of RAOBCORE; see Section 2.1.2) and 19 different climate models. Tropical T_{SST} and T_{L+O} trends from the same climate models and four different observational datasets (Section 2.1.3) are also shown. The multi-model average trend at a discrete pressure level, $\ll b_m(z) \gg$, was calculated from the ensemble-mean trends of individual models [see Equation (7)]. The grey-shaded envelope is $s\{<b_m(z)>\}$, the 2σ standard deviation of the ensemble-mean trends at discrete pressure levels. The yellow envelope represents $2\sigma_{SE}$, DCPS07's estimate of uncertainty in the mean trend. For visual display purposes, T_{L+O} results have been offset vertically to make it easier to discriminate between trends in T_{L+O} and T_{SST}. Satellite and radiosonde trends in panel B are plotted with their respective adjusted 2σ confidence intervals (see Section 4.1). Model results are the multi-model average trend and the standard deviation of the ensemble-mean trends, and grey- and yellow-shaded areas represent the same uncertainty estimates described in panel A (but now for layer-averaged temperatures rather than temperatures at discrete pressure levels). The y-axis in panel B is nominal, and bears no relation to the pressure coordinates in panel A. The analysis period is January 1979 through December 1999, the period of maximum overlap between the observations and most of the model 20CEN simulations. Note that DCPS07 used the same analysis period for model data, but calculated all observed trends over 1979–2004.

DOI: 10.1002/joc

greater warming in the tropical troposphere than is evident in the raw data upon which they are based. Climate model results are in closer agreement with these newer radiosonde datasets, which were not used by DCPS07.

The model-average warming of the tropical surface over 1979–1999 is slightly larger than in the single realization of the observations, both for T_{SST} and T_{L+O} (Figure 6A and Table I). As discussed in Section 3, this small difference in simulated and observed surface warming rates may be due to the random effects of natural internal variability, model error, or some combination thereof.[17] One important consequence of this difference is that we expect the simulated warming in the free troposphere to be generally larger than in observations.

Figure 6B summarizes results from a variety of trend comparisons, and shows trends in tropical T_{2LT} and T_2 from RSS and UAH, in synthetic MSU temperatures from the seven radiosonde products, and in the model-average synthetic MSU temperatures. Results are also given for DCPS07's σ_{SE} and for $s\{<b_m>\}$, the inter-model standard deviation of trends. Application of the DCPS07 consistency test leads to the incorrect conclusion that the model-average T_{2LT} and T_2 signal trends are significantly different from the observed signal trends in all radiosonde products. Modification of the test to account for uncertainties in the observed trends leads to very different conclusions. For T_{2LT}, for example, the d_1^* test statistic [see Equation (12)] indicates that the model-average signal trend is not significantly different (at the 5% level) from the observed signal trends in three of the more recent radiosonde products (RICH, IUK, and RAOBCORE v1.4). Clearly, agreement between models and observations depends on both the observations that are selected and the metric used to assess agreement.

8. Summary and conclusions

Several recent comparisons of modelled and observed atmospheric temperature changes have focused on the tropical troposphere (Santer *et al*., 2006; Douglass *et al*., 2007; Thorne *et al*., 2007). Interest in this region was stimulated by an apparent inconsistency between climate model results and observations. Climate models consistently showed tropospheric amplification of surface warming in response to human-caused increases in well-mixed GHGs. In contrast, early versions of satellite and radiosonde datasets implied that the surface had warmed by more than the tropical troposphere over the satellite era. This apparent discrepancy has been cited as evidence for the absence of a human effect on climate (e.g. Singer, 2008).

A number of national and international assessments have tried to determine whether this discrepancy is real and of practical significance, or simply an artifact of problems with observational data (e.g., NRC, 2000; Karl *et al*., 2006; IPCC, 2007). The general tenor of these assessments is that structural uncertainties in satellite- and radiosonde-based estimates of tropospheric temperature change are currently large: we do not have an unambiguous observational yardstick for gauging true levels of model skill (or lack thereof). The most comprehensive assessment was the first report produced under the auspices of the U.S. Climate Change Science Program (CCSP; Karl *et al*., 2006). This report concluded that advances in identifying and adjusting for inhomogeneities in satellite and radiosonde data had helped to resolve the discrepancies described above, at least at global scales.

In the tropics, however, important differences remained between the simulated and observed 'differential warming'. In climate models, the tropical lower troposphere warmed by more than the surface. This amplification of surface warming was timescale-invariant, consistent across a range of models, and in accord with basic theoretical considerations (Santer *et al*., 2005, 2006; Thorne *et al*., 2007). For month-to-month and year-to-year temperature changes, all satellite and radiosonde datasets showed amplification behaviour consistent with model results and basic theory. For multi-decadal changes, however, only two of the then-available satellite datasets (and none of the then-available radiosonde datasets) indicated warming of the troposphere exceeding that of the surface (Karl *et al*., 2006).

Karl *et al*. noted that these findings could be interpreted in at least two ways. Under one interpretation, the physical mechanisms controlling real-world amplification behaviour vary with timescale, and models have some common error in representing this timescale-dependence. The second interpretation posited residual errors in many of the satellite and radiosonde datasets used in the CCSP report. In view of the large structural uncertainties in the observations, the consistency of model amplification results across a range of timescales, and independent evidence of substantial tropospheric warming (Santer *et al*., 2003, 2007; Paul *et al*., 2004; Mears *et al*., 2007; Allen and Sherwood, 2007, 2008), this was deemed to be the more plausible explanation.

DCPS07 reached a very different conclusion from that of the CCSP report, and claim to find significant differences between models and observations, both for trends in tropospheric temperatures and for trends in lower tropospheric lapse rates. Their claim is based on the application of a 'consistency test' to essentially the same model and observational data available to Karl *et al*. (2006). Their test has two serious flaws: it neglects statistical uncertainty in observed temperature trends arising from interannual temperature variability, and it uses an inappropriate metric [σ_{SE}; see Equation (10)] to judge the statistical significance of differences between the observed trend and the multi-model ensemble-mean trend, $\ll b_m \gg$.

Consider first the issue of statistical uncertainties. DCPS07 make the implicit assumption that the observed and simulated trends are unaffected by interannual climate variability, and provide perfect information on the true temperature response to external forcing. This assumption is incorrect, as examination of Figures 1 and 2A readily shows: the true response is not perfectly

DOI: 10.1002/joc

known in either observations or the model results. It can only be estimated from a single, noisy observational record and from relatively small ensembles of model results. Any meaningful consistency test must account for the effects of interannual variability, and for the uncertainties it introduces in estimating the underlying (but unknown) 'trend signal' in observations. The DCPS07 test does not do this.

Second, DCPS07's σ_{SE} is not a meaningful basis for testing whether a highly uncertain observed trend signal is consistent with the average of imperfectly-known model signal trends. This is readily apparent when one applies the DCPS07 test to synthetic data with approximately the same statistical properties as satellite T_{2LT} and T_2 data. In this case, we know *a priori* that the same statistical model generated the synthetic 'observed' and synthetic 'simulated' data, and that application of the test should yield (on average) rejection of the hypothesis of 'no significant difference in signal trends' approximately $p\%$ of the time at a stipulated $p\%$ significance level. The DCPS07 test, however, gives rejection rates that are many times higher than values expected by chance alone (see Figure 5B).

In contrast to DCPS07, we explicitly account for the effects of interannual variability on observational trends. We do this using two different significance testing strategies. In the first, we use a 'paired trends' test [with the d statistic; Equation (3)] that compares each observational trend with the trend from each individual realization of each model. With this procedure, we test the hypothesis (H_1) that the trend in an individual model realization of signal plus noise is consistent with the single realization of signal plus noise in the observations. In our second approach, we use a modified version of DCPS07's consistency test [with the d_1^* statistic; Equation (12)], to test the hypothesis (H_2) that the model-average signal trend is consistent with the signal trend estimated from the single realization of the observations. With the d test, very few of the model trends in tropical T_{2LT} and T_2 over 1979 to 1999 are significantly different from RSS or UAH trends (Table II). Similarly, when the d_1^* test is applied to T_{2LT} and T_2 trends, hypothesis H_2 cannot be rejected at the nominal 5% level (Table III).

A more stringent test of model performance involves trends in the time series of differences between surface and lower tropospheric temperature anomalies. Trends in T_{SST} (or T_{L+O}) minus T_{2LT} provide a simple measure of changes in lapse rate. Differencing reduces the amplitude of the (common) unforced variability in surface temperature and T_{2LT}, and makes it easier to identify true model errors in the forced component of lapse-rate trends.

While tests involving trends in T_{2LT} and T_2 time series almost invariably showed non-significant differences between models and satellite data (Section 5.1), results for lapse-rate trends are more sensitive to structural uncertainties in observations (Section 5.2). If RSS T_{2LT} data are used for computing lapse-rate trends, the warming aloft is larger than at the surface (consistent with model results). Very few simulated lapse-rate trends differ significantly from observations in 'paired trends' tests (Table V). When the d_1^* test is applied, there is no case in which hypothesis H_2 can be rejected at the nominal 5% level (Table VI).

When UAH T_{2LT} data are used, the warming aloft is smaller than at the surface. Even in the UAH case, however, hypothesis H_1 is not rejected consistently. Rejection rates for 'paired trends' tests conducted at the 5% significance level range from *ca.* 31 to 88%, depending on the choice of observational surface temperature dataset (Table V). Alternately, our modified version of the DCPS07 test reveals that hypothesis H_2 is rejected at the nominal 5% level in all cases involving UAH-based estimates of lapse-rate changes (Table VI).

Our findings do not bring final resolution to the issue of whether UAH or RSS provide more reliable estimates of temperature changes in the tropical troposphere. We note, however, that the RSS-based estimates of tropical lapse-rate changes are in better accord with satellite datasets developed by the UMd and NOAA/NESDIS groups (Vinnikov *et al*., 2006; Zou *et al*., 2006), with newer radiosonde datasets (e.g. Allen and Sherwood, 2007, 2008; Haimberger *et al*., 2008; Sherwood *et al*., 2008; Titchner *et al*., 2008), and with basic moist adiabatic lapse-rate theory. Furthermore, RSS results show amplification of tropical surface warming across a range of timescales (consistent with model behaviour), whereas UAH T_{2LT} data yield amplification for monthly and annual temperature changes, but not for decadal changes. If the UAH results were correct, the physics controlling the response of the tropical atmosphere to surface warming must vary with timescale. Mechanisms that might govern such behaviour have not been identified.

Model errors in forcing and response must also contribute to remaining differences between simulated and observed lapse-rate trends. For example, only 9 of the 19 models used in our study attempted to represent the climate forcing associated with the eruptions of El Chichòn and Pinatubo (Forster and Taylor, 2006). Statistical comparisons between modelled and observed temperature changes can be sensitive to the inclusion or exclusion of volcanic forcing (Santer *et al*., 2001; Wigley *et al*., 2005; Lanzante, 2007).

Similarly, roughly half the models analysed here exclude stratospheric ozone depletion, which has a pronounced impact on lower stratospheric and upper tropospheric temperatures, and hence on T_2 (Santer *et al*., 2006). Even models which include some form of stratospheric ozone depletion do not correctly represent the observed profile of ozone losses below *ca.* 20 km in the tropics (Forster *et al*., 2007). The latter deficiency may have considerable impact on model-predicted temperature changes above the tropical tropopause and in the uppermost troposphere, and therefore on agreement with observations.

In summary, considerable scientific progress has been made since the first report of the U.S. Climate Change Science Program (Karl *et al*., 2006). There is no longer a

DOI: 10.1002/joc

serious and fundamental discrepancy between modelled and observed trends in tropical lapse rates, despite DCPS07's incorrect claim to the contrary. Progress has been achieved by the development of new T_{SST}, T_{L+O}, and T_{2LT} datasets, better quantification of structural uncertainties in satellite- and radiosonde-based estimates of tropospheric temperature change, and the application of rigorous statistical comparisons of modelled and observed changes.

We may never completely reconcile the divergent observational estimates of temperature changes in the tropical troposphere. We lack the unimpeachable observational records necessary for this task. The large structural uncertainties in observations hamper our ability to determine how well models simulate the tropospheric temperature changes that actually occurred over the satellite era. A truly definitive answer to this question may be difficult to obtain. Nevertheless, if structural uncertainties in observations and models are fully accounted for, a partial resolution of the long-standing 'differential warming' problem has now been achieved. The lessons learned from studying this problem can and should be applied towards the improvement of existing climate monitoring systems, so that future model evaluation studies are less sensitive to observational ambiguity.

Acknowledgements

We acknowledge the modelling groups for providing their simulation output for analysis, the Program for Climate Model Diagnosis and Intercomparison (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. The authors received support from a Distinguished Scientist Fellowship of the U.S. Dept. of Energy, Office of Biological and Environmental Research (BDS); the joint DEFRA and MoD Programme (PWT; contracts GA01101 and CBC/2B/0417_Annex C5, respectively); Grant no. P18120-N10 of the Austrian Science Funds (LH); and the NOAA Office of Climate Programs ('Climate Change, Data and Detection') Grant no. NA87GP0105 (TMLW). We thank Mike MacCracken (Climate Institute), David Parker (U.K. Meteorological Office Hadley Centre), Dick Reynolds (NCDC), Dian Seidel (NOAA Air Resources Laboratory), Francis Zwiers (Environment Canada), and an anonymous reviewer for useful comments and discussion. Dave Easterling and Imke Durre (NCDC) and R. Dobosy and Jenise Swall (NOAA Air Resources Laboratory) provided helpful comments in the course of NOAA internal reviews. Observed MSU data were kindly provided by John Christy (UAH) and Konstantin Vinnikov (UMd). Observed surface temperature data were provided by John Kennedy at the U.K. Meteorological Office Hadley Centre (HadISST1), and by Dick Reynolds at the NCDC (ERSST-v2 and ERSST-v3).

Appendix 1: Statistical notation

Subscripts and indices

m	Subscript denoting model data
o	Subscript denoting observational data
t	Index over time (in months)
i	Index over number of models
j	Index over number of 20CEN realizations
z	Index over number of atmospheric levels

Sample sizes

n_t	Total number of time samples (usually 252)
n_e	Effective number of time samples, adjusted for temporal autocorrelation
n_m	Total number of models (19)
$n_r(i)$	Total number of 20CEN realizations for the i^{th} model
N	Total number of synthetic time series

Time series

$y_m(t)$	Simulated T_{2LT} or T_2 time series
$\phi_m(t)$	Underlying signal in $y_m(t)$ in response to forcing
$\eta_m(t)$	Realization of internally generated noise in $y_m(t)$
$x(t)$	Synthetic AR-1 time series
$z(t)$	Synthetic noise time series

Trends

b_m	Least-squares linear trend in an individual $y_m(t)$ time series
$<b_m(i)>$	Ensemble-mean trend in the i^{th} model
$\ll b_m \gg$	Multi-model ensemble-mean trend
$\ll b_m(z) \gg$	Multi-model ensemble-mean trend profile

Standard errors and standard deviations

$s\{b_m\}$	Standard error of b_m
$s\{y_m(t)\}$	Temporal standard deviation of $y_m(t)$ anomaly time series
$s\{<b_m>\}$	Inter-model standard deviation of ensemble-mean trends
$s\{<b_m(z)>\}$	Inter-model standard deviation of ensemble-mean trends at discrete pressure levels
σ_{SE}	DCPS07 '*estimate of the uncertainty of the mean*'

Other regression terms

$e(t)$	Regression residuals
r_1	Lag-1 autocorrelation of regression residuals

Test statistics

d	Paired trends test statistic [Equation (3)]

DOI: 10.1002/joc

d^* Test statistic for original DCPS07 consistency test [Equation (11)]

d_1^* Test statistic for modified version of DCPS07 consistency test [Equation (12)]

Appendix 2: Technical Notes

[1]See Table 3.4 in Lanzante *et al*., 2006. For the specific period 1979 to 2004, tropical (20 °N–20 °S) T_2 trends range from 0.05 °C/decade (UAH) to 0.19 °C/decade (UMd), while T_{2LT} trends span the range 0.05 °C/decade (UAH) to 0.15 °C/decade (RSS). The most important sources of uncertainty are likely to be '*due to inter-satellite calibration offsets and calibration drifts*' (Mears *et al*., 2006, page 78).

[2]The UMd and NOAA/NESDIS groups do not provide a T_{2LT} product. Because of their calibration procedure, the NOAA/NESDIS T_2 data are only available for a shorter period (1987 to present) than the T_2 products of the three other groups.

[3]A more recent version of the RSS T_2 and T_{2LT} datasets (version 3.1) now exists. RSS versions 3.0 and 3.1 are virtually identical over the primary analysis period considered here (1979 to 1999). For UAH data, a version 5.2 exists for T_{2LT} but not for T_2 data, for which only version 5.1 is available.

[4]All simulations included human-induced changes in well-mixed GHGs and the direct (scattering) effects of sulphate aerosols on incoming solar radiation. Other external forcings (such as changes in ozone, carbonaceous aerosols, indirect effects of aerosols on clouds, land surface properties, solar irradiance, and volcanic dust loadings) were not handled uniformly across different modeling groups. For further details of the applied forcings, see Santer *et al*., 2005, 2006.

[5]DCPS07 used a larger set of 20CEN runs (67 simulations performed with 22 different models) and incorporated model results that were not available at the time of the Santer *et al*. (2005) study. This difference in the number of 20CEN models employed in the two investigations is immaterial for illustrating the statistical problems in the consistency test applied by DCPS07. All 49 simulations employed in our current work were also analyzed by DCSP07.

[6]Amplification occurs due to the non-linear effect of the release of latent heat by moist ascending air in regions experiencing convection.

[7]The 20CEN experiments analyzed here were performed with coupled atmosphere-ocean General Circulation Models (A/OGCMs) driven by estimates of historical changes in external forcing. Due to chaotic variability in the climate system, small differences in the atmospheric or oceanic initial conditions at the start of the 20CEN run (typically in the mid- to late-19th century) rapidly lead to different manifestations of climate noise. Within the space of several months, the state of the atmosphere is essentially uncorrelated with the initial state. This means that even the same model, when run many times with identical external forcings (but each time from slightly different initial conditions), produces many different samples of $\eta_m(t)$, each superimposed on the same underlying signal, $\phi_m(t)$.

[8]Our d_1^* test involving the multi-model ensemble-mean trend [see Equation (12)], also relies on an AR-1 model to estimate n_e and adjust the observed standard error, and is therefore also likely to be too liberal.

[9]We use $<>$ to denote an ensemble average over multiple 20CEN realizations performed with a single model. Double angle brackets, $\ll\gg$, indicate a multi-model ensemble average.

[10]Under this assumption, the total uncertainty in $\ll b_m \gg -b_o$ is determined solely by inter-model trend differences arising from structural differences between the models [see Equations (9)–(11)]. As discussed in Section 3, however, the total uncertainty in the magnitude of $\ll b_m \gg -b_o$ reflects not only these structural differences, but also inter-model differences in internal variability and ensemble size.

[11]Inter-model differences in the size of the confidence intervals in Fig. 3A are due primarily to differences in the amplitude and temporal autocorrelation properties of $\eta_m(t)$, but are also affected by neglect or inclusion of the effects of volcanic forcing (see Santer *et al*., 2005, 2006). Models with large ENSO variability (such as GFDL-CM2.1 and FGOALS-g1.0) have large adjusted confidence intervals, while A/OGCMs with relatively coarse-resolution, diffusive oceans (such as GISS-AOM) have much weaker ENSO variability and smaller values of $s\{b_m\}$.

[12]We have explored the sensitivity of our adjusted standard errors and significance test results to choices of averaging period ranging from two to 12 months. These choices span a wide range of temporal autocorrelation behaviour. Results for the d test are relatively insensitive to the selected averaging period, suggesting that our adjustment method is reasonable.

[13]There are four tests because we are using two atmospheric layers (T_{2LT} and T_2) and two observational datasets (RSS and UAH).

[14]One of the assumptions underlying the d_1^* test (and all tests performed here) is that structural uncertainty in the observations is negligible (see Section 4.2). We know this is not the case in the real world (see, *e.g.,* Seidel *et al*., 2004; Thorne *et al*., 2005a; Lanzante *et al*., 2006; Mears *et al*., 2006). In the present study, we have examined the effects of structural uncertainties in satellite and radiosonde data by treating each observational dataset independently, and assessing the robustness of our model-versus-observed trend comparisons to different dataset choices. An alternative approach would be to explicitly include a structural uncertainty term for the observations in the d_1^* test statistic itself.

[15]Note that RATPAC-B is unadjusted after 1997. RATPAC-A, which we use here, accounts for inhomogeneities before and after 1997.

DOI: 10.1002/joc

[16]Sherwood *et al*. (2008) argue that this procedure does not completely homogenize data from stations between 5 °S and 20 °N, since trends at these stations remained highly variable and (on average) unphysically low compared to those at neighbouring latitudes that are much more accurately known. The implication is that gradual (rather than step-like) changes in bias at many tropical stations may not be reliably identified and adjusted by the IUK homogenization procedure. If this is the case, the IUK trends shown here are likely to be underestimates of the true trends.

[17]An error in the model average surface warming is entirely likely given the neglect of indirect aerosol effects in roughly half of the models analyzed here.

References

Allen RJ, Sherwood SC. 2007. Utility of radiosonde wind data in representing climatological variations of tropospheric temperature and baroclinicity in the western tropical Pacific. *Journal of Climate*, **20**: 5229–5243.

Allen RJ, Sherwood SC. 2008. Warming maximum in the tropical upper troposphere deduced from thermal winds. *Nature Geoscience*, **65**: 399–403.

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD. 2006. Uncertainty estimates in regional and global observed temperature changes: A new dataset from 1850. *Journal of Geophysical Research* **111**: D12106, Doi:10.1029/2005JD006548.

Christy JR, Norris WB, Spencer RW, Hnilo JJ. 2007. Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements. *Journal of Geophysical Research* **112**: D06102, Doi:10.1029/2005JD006881.

Christy JR, Spencer RW, Braswell WD. 2000. MSU tropospheric temperatures: Data set construction and radiosonde comparisons. *Journal of Atmospheric and Oceanic Technology* **17**: 1153–1170.

Christy JR, Spencer RW, Norris WB, Braswell WD, Parker DE. 2003. Error estimates of version 5.0 of MSU/AMSU bulk atmospheric temperatures. *Journal of Atmospheric and Oceanic Technology* **20**: 613–629.

Douglass DH, Christy JR, Pearson BD, Singer SF. 2007. A comparison of tropical temperature trends with model predictions. *International Journal of Climatology* **27**: Doi:10.1002/joc.1651.

Douglass DH, Pearson BD, Singer SF. 2004. Altitude dependence of atmospheric temperature trends: Climate models versus observations. *Geophysical Research Letters* **31**: L13208, Doi:10.1029/2004/GL020103.

Durre I, Vose R, Wuertz DB. 2006. Overview of the integrated global radiosonde archive. *Journal of Climate* **19**: 53–68.

Forster PM, Bodeker G, Schofield R, Solomon S. 2007. Effects of ozone cooling in the tropical lower stratosphere and upper troposphere. *Geophysical Research Letters* **34**: L23813, Doi:10.1029/2007GL031994.

Forster PM, Taylor KE. 2006. Climate forcings and climate sensitivities diagnosed from coupled climate model integrations. *Journal of Climate* **19**: 6181–6194.

Free M, Seidel DJ, Angell JK, Lanzante JR, Durre I, Peterson TC. 2005. Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC): A new dataset of large-area anomaly time series. *Journal of Geophysical Research* **110**: D22101, Doi:10.1029/2005JD006169.

Gaffen D, *et al*. 2000. Multi-decadal changes in the vertical temperature structure of the tropical troposphere. *Science* **287**: 1239–1241.

Haimberger L. 2007. Homogenization of radiosonde temperature time series using innovation statistics. *Journal of Climate* **20**: 1377–1403.

Haimberger L, Tavolato C, Sperka S. 2008. Towards elimination of the warm bias in historic radiosonde temperature records – Some new results from a comprehensive intercomparison of upper air data. *Journal of Climate*, (in press).

Hegerl GC, *et al*. 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, New York.

IPCC (Intergovernmental Panel on Climate Change). 1996. Summary for policymakers. In *Climate Change 1995: The Science of Climate Change*, Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change, Houghton JT, Meira Filho LG, Callander BA, Harris N, Kattenberg A, Maskell K (eds). Cambridge University Press: Cambridge, New York.

IPCC (Intergovernmental Panel on Climate Change). 2001. Summary for policymakers. In *Climate Change 2001: The Scientific Basis*, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds). Cambridge University Press: Cambridge, New York.

IPCC (Intergovernmental Panel on Climate Change). 2007. 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, New York.

Karl TR, Hassol SJ, Miller CD, Murray WL (eds). 2006. *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. National Oceanic and Atmospheric Administration, National Climatic Data Center: Asheville, NC; 164.

Lanzante JR. 2005. A cautionary note on the use of error bars. *Journal of Climate* **18**: 3699–3703.

Lanzante JR. 2007. Diagnosis of radiosonde vertical temperature trend profiles: Comparing the influence of data homogenization versus model forcings. *Journal of Climate* **20**(21): 5356–5364.

Lanzante JR, Klein SA, Seidel DJ. 2003. Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities, and MSU comparison. *Journal of Climate* **16**: 241–262.

Lanzante JR, Peterson TC, Wentz FJ, Vinnikov KY. 2006. What do observations indicate about the change of temperatures in the atmosphere and at the surface since the advent of measuring temperatures vertically? In *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*, Karl TR, Hassol SJ, Miller CD, Murray WL (eds). A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research, Washington DC.

Manabe S, Stouffer RJ. 1980. Sensitivity of a global climate model to an increase of CO_2 concentration in the atmosphere. *Journal of Geophysical Research* **85**: 5529–5554.

McCarthy MP, Titchner HA, Thorne PW, Tett SFB, Haimberger L, Parker DE. 2008. Assessing bias and uncertainty in the HadAT adjusted radiosonde climate record. *Journal of Climate* **21**: 817–832.

Mears CA, Schabel MC, Wentz FJ. 2003. A reanalysis of the MSU channel 2 tropospheric temperature record. *Journal of Climate* **16**: 3650–3664.

Mears CA, Forest CE, Spencer RW, Vose RS, Reynolds RW. 2006. What is our understanding of the contributions made by observational or methodological uncertainties to the previously-reported vertical differences in temperature trends? In *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*, Karl TR, Hassol SJ, Miller CD, Murray WL (eds). A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research, Washington DC.

Mears CA, Santer BD, Wentz FJ, Taylor KE, Wehner MF. 2007. Relationship between temperature and precipitable water changes over tropical oceans. *Geophysical Research Letters* **34**: L24709, Doi:10.1029/2007GL031936.

Mears CA, Wentz FJ. 2005. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309**: 1548–1551.

Mitchell JFB, *et al*. 2001. Detection of climate change and attribution of causes. In *Climate Change 2001: The Scientific Basis*, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, Mitchell JFB, Karoly DJ, Hegerl GC, Zwiers FW, Allen MR, Marengo J (eds). Cambridge University Press: Cambridge, New York; 881.

NRC (National Research Council). 2000. *Reconciling Observations of Global Temperature Change*. National Academy Press: Washington, DC; 85.

Paul F, Kaab A, Maisch M, Kellenberger T, Haeberli W. 2004. Rapid disintegration of Alpine glaciers observed with satellite data. *Geophysical Research Letters* **31**: L21402, Doi:10.1029/2004GL020816.

Randel WJ, Wu F. 2006. Biases in stratospheric and tropospheric temperature trends derived from historical radiosonde data. *Journal of Climate* **19**: 2094–2104.

Rayner NA, *et al*. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research* **108**: 4407, Doi:10.1029/2002JD002670, HadISST1 data are available at http://www.hadobs.org/.

Rayner NA, *et al*. 2006. Improved analyses of changes and uncertainties in marine temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. *Journal of Climate* **19**: 446–469.

Santer BD, Penner JE, Thorne PW. 2006. How well can the observed vertical temperature changes be reconciled with our understanding of the causes of these changes? In *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*, Karl TR, Hassol SJ, Miller CD, Murray WL (eds). A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research, Washington DC.

Santer BD, Wigley TML, Barnett TP, Anyamba E. 1996. Detection of climate change and attribution of causes. In *Climate Change 1995: The Science of Climate Change*, Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change, Houghton JT, Meira Filho LG, Callander BA, Harris N, Kattenberg A, Maskell K (eds). Cambridge University Press: Cambridge, New York; 572.

Santer BD, *et al*. 1999. Uncertainties in observationally based estimates of temperature change in the free atmosphere. *Journal of Geophysical Research* **104**: 6305–6333.

Santer BD, *et al*. 2000a. Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *Journal of Geophysical Research* **105**: 7337–7356.

Santer BD, *et al*. 2000b. Interpreting differential temperature trends at the surface and in the lower troposphere. *Science* **287**: 1227–1232.

Santer BD, *et al*. 2001. Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends. *Journal of Geophysical Research* **106**: 28033–28059.

Santer BD, *et al*. 2003. Contributions of anthropogenic and natural forcing to recent tropopause height changes. *Science* **301**: 479–483.

Santer BD, *et al*. 2005. Amplification of surface temperature trends and variability in the tropical atmosphere. *Science* **309**: 1551–1556.

Santer BD, *et al*. 2007. Identification of human-induced changes in atmospheric moisture content. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 15248–15253.

Seidel DJ, *et al*. 2004. Uncertainty in signals of large-scale climate variations in radiosonde and satellite upper-air temperature data sets. *Journal of Climate* **17**: 2225–2240.

Sherwood SC. 2007. Simultaneous detection of climate change and observing biases in a network with incomplete sampling. *Journal of Climate* **20**: 4047–4062.

Sherwood SC, Lanzante JR, Meyer CL. 2005. Radiosonde daytime biases and late-20th century warming. *Science* **309**: 1556–1559.

Sherwood SC, Meyer CL, Allen RJ, Titchner HA. 2008. Robust tropospheric warming revealed by iteratively homogenized radiosonde data. *Journal of Climate*, early online release, Doi:10.1175/2008JCLI2320.1.

Singer SF. 2001. Global warming: An insignificant trend? *Science* **292**: 1063–1064.

Singer SF. 2008. *Nature, Not Human Activity, Rules the Climate: Summary for Policymakers of the Report of the Nongovernmental International Panel on Climate Change*, Singer SF (ed.). The Heartland Institute: Chicago, IL.

Smith TM, Reynolds RW. 2005. A global merged land and sea surface temperature reconstruction based on historical observations (1880–1997). *Journal of Climate* **18**: 2021–2036.

Smith TM, Reynolds RW, Peterson TC, Lawrimore J. 2008. Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). *Journal of Climate*, (in press).

Spencer RW, Christy JR. 1990. Precise monitoring of global temperature trends from satellites. *Science* **247**: 1558–1562.

Storch H, Zwiers FW. 1999. *Statistical Analysis in Climate Research*. Cambridge University Press: Cambridge; 484.

Thiébaux HJ, Zwiers FW. 1984. The interpretation and estimation of effective sample size. *Journal of Meteorology and Applied Climatology* **23**: 800–811.

Thorne PW, *et al*. 2005a. Uncertainties in climate trends: Lessons from upper-air temperature records. *Bulletin of the American Meteorological Society* **86**: 1437–1442.

Thorne PW, *et al*. 2005b. Revisiting radiosonde upper-air temperatures from 1958 to 2002. *Journal of Geophysical Research* **110**: D18105, Doi:10.1029/2004JD005753.

Thorne PW, *et al*. 2007. Tropical vertical temperature trends: A real discrepancy? *Geophysical Research Letters* **34**: L16702, Doi:10.1029/2007GL029875.

Titchner HA, Thorne PW, McCarthy MP, Tett SFB, Haimberger L, Parker DE. 2008. Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments. *Journal of Climate*, early online release, Doi:10.1175/2008JCLI2419.1.

Trenberth KE, *et al*. 2007. Observations: Surface and atmospheric 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, New York.

Uppala SM, *et al*. 2005. The ERA-40 reanalysis. *Quarterly Journal of the Royal Meteorological Society* **131**: 2961–3012.

Vinnikov KY, Grody NC. 2003. Global warming trend of mean tropospheric temperature observed by satellites. *Science* **302**: 269–272.

Vinnikov KY, *et al*. 2006. Temperature trends at the surface and the troposphere. *Journal of Geophysical Research* **111**: D03106, Doi:10.1029/2005jd006392.

Wentz FJ, Schabel M. 1998. Effects of orbital decay on satellite-derived lower-tropospheric temperature trends. *Nature* **394**: 661–664.

Wentz FJ, Schabel M. 2000. Precise climate monitoring using complementary satellite data sets. *Nature* **403**: 414–416.

Wigley TML. 2006. Appendix A: Statistical issues regarding trends. In *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*, Karl TR, Hassol SJ, Miller CD, Murray WL (eds). A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research, Washington DC.

Wigley TML, Ammann CM, Santer BD, Raper SCB. 2005. The effect of climate sensitivity on the response to volcanic forcing. *Journal of Geophysical Research* **110**: D09107, Doi: 10.1029/2004/JD005557.

Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press: San Diego, CA; 467.

Zou C-Z, *et al*. 2006. Recalibration of Microwave Sounding Unit for climate studies using simultaneous nadir overpasses. *Journal of Geophysical Research* **111**: D19114, Doi:10.1029/2005JD006798.

Zwiers FW, von Storch H. 1995. Taking serial correlation into account in tests of the mean. *Journal of Climate* **8**: 336–351.

Supporting Online Material

1 Additional Information Regarding Observational Datasets

1.1 Radiosonde Data

An important component of the structural uncertainty in radiosonde datasets relates to changes over time in thermal shielding of the balloon-borne temperature sensor. Since solar heating of the sensor influences the temperature measurement itself, improvements in the effectiveness of the shielding are likely to introduce spurious cooling (Mears *et al.*, 2006; Trenberth *et. al.*, 2007). This 'solar heating' problem is clearly manifest in comparisons of temperature trends based on daytime and night-time radiosonde ascents (Sherwood *et al.*, 2005; Randel and Wu, 2006). Trends based on night-time ascents (which are not subject to solar heating of the temperature sensor) show consistently larger warming than trends based on daytime ascents. Although there are errors in the nighttime radiosonde data at individual stations (Randel and Wu, 2006; Christy *et al.*, 2007), it is unlikely that such night-time biases can completely offset daytime biases in all tropical radiosonde stations, as DCPS07 imply.

1.2 Surface Temperature Data

Previous model evaluation studies relied on $T_{\mathrm{L+O}}$ datasets for estimating lapse rate changes (Santer *et al.*, 2005, 2006; DCPS07). Changes in tropical surface temperature in merged land and ocean data are larger than in T_{SST} datasets (see Table 1), because the recent warming over land is larger than over ocean. Since model 20CEN results also show larger warming over land than over ocean, tests of differences between modelled and observed lapse-rate trends should be relatively insensitive to the choice of T_{SST} or $T_{\mathrm{L+O}}$ datasets, as long as the SST changes in both are of similar size (as they are in HadISST1 T_{SST} and HadCRUT3v $T_{\mathrm{L+O}}$). The test results for lapse rate trends calculated with the HadISST1-RSS difference series pair are indeed very similar to those obtained with HadCRUT3v-RSS (Table 5). The same does not hold for difference series involving HadISST1-UAH and HadCRUT3v-UAH.

2 Technical Information on Significance Tests

2.1 Degrees of Freedom for Paired Trends Test

The "-2" in equation (5) requires explanation. This loss of two degrees of freedom (d.o.f.) arises because two parameters (the y-axis intercept and the trend b) are being estimated in the linear regression (Wilks, 1995, page 164). A complicating factor is

our use of monthly-mean anomalies with respect to climatological monthly means. For anomalies defined in this way, time-mean values for each of the 12 months are constrained to be zero, and we may therefore lose additional d.o.f. in the regression. It is unclear whether it is most appropriate to subtract these additional d.o.f. from n_t in equation (5) or from n_t in equation (6). Here, we follow previous work (Wilks, 1995) and subtract two d.o.f. from n_t in equation (5). Given the large sample sizes ($n_t = 252$), this choice has little impact on the unadjusted standard errors, but does affect the adjusted standard errors. For the latter case, our subsequent experiments with synthetic data (see Section 6) reveal that use of $n_e - 2$ in equation (5) yields rejection rates that have a small positive bias relative to theoretical expectations (see Fig. 6). This bias would be reduced by accounting for the additional d.o.f. loss related to the anomaly definition.

The key point to note here is that our paired trends test is slightly liberal – *i.e.*, it is more likely to incorrectly reject hypothesis H_1 (see Section 4). This should make it easier for us to obtain results that are in accord with DCPS07's finding of statistically significant differences between modelled and observed trends. Even with a slightly liberal test, however, our results are <u>not</u> in accord with those of DCPS07.

2.2 Tests with Synthetic Data: Explanation of Positive Bias in Rejection Rates

The small positive bias in rejection rates arises in part because r_1, the sample value of the lag-1 autocorrelation coefficient (which is estimated from the regression residuals of the synthetic time series) is a biased estimate of the population value of a_1 used in the AR-1 model (Nychka *et al.*, 2000). On average, $r_1 < a_1$, leading to a slight inflation of the effective sample size n_e in equation (6), which contributes to the small positive bias in the rejection rates. A further contributory factor to the rejection rate bias is the slight skewness of the rejection rate distributions: while minimum rejection rate values are bounded by zero, the maximum is not, and in rare cases can be large and influence the mean[1].

3 Results of Significance Tests

3.1 Sensitivity Tests

We performed several tests to explore the sensitivity of the "paired trends" test results in Table 2 to errors in model variability and to the use of longer observational

[1]In the $N = 19$ case, for example, rejection rates in the 1,000-member distribution range from a minimum of 0 to a maximum of 24% for 5% tests, with a mean value of 6.7%.

records. These results are presented in Table 7. In the first sensitivity test, denoted by "SENS1", we set $s\{b_m\} = s\{b_o\}$ in equation (3), *i.e.*, we replaced each model's adjusted standard error with the observed adjusted standard error. Comparison of SENS1 results with the baseline case reveals that the rejection rates of hypothesis H_1 are only slightly increased. This is in accord with our finding that, even without accounting for statistical uncertainty in b_m, most of the simulated tropospheric trends are already within the 2σ confidence intervals of the RSS and UAH trends (Fig. 3A).

In the second sensitivity test ("SENS2"), we calculated observed trends in $T_{2\mathrm{LT}}$ and T_2 over the 336-month period from January 1979 to December 2006, which is a third longer than the analysis period in the baseline case. As in SENS1, we set $s\{b_m\} = s\{b_o\}$. Since most of the model 20CEN experiments end in 1999, we make the necessary assumption that values of b_m estimated over 1979 to 1999 are representative of the longer-term b_m trends over 1979 over 2006. Examination of the observed data suggests that this assumption is not unreasonable: UAH and RSS $T_{2\mathrm{LT}}$ and T_2 trends over 1979 to 2006 are only 0.01 to 0.02°C/decade larger than trends estimated over 1979 to 1999 (see also Thorne *et al.*, 2007).

The use of longer $y_o(t)$ time series in SENS2 yields larger effective sample sizes and larger values of the denominator in equation (4). Values of $s\{b_o\}$ are therefore smaller (by a factor of *ca.* 1.8) than in BASE. Because of this marked reduction in the size of the observed standard errors, smaller differences between b_m and b_o are

deemed to be statistically significant, and rejection rates are consistently higher than in BASE or SENS1 (see Table 7). Even with longer records, however, no more than 23% of the tests performed lead to rejection of hypothesis H_1 at the nominal 5% significance level.

3.2 Tests Involving Model Data Only

3.2.1 Tests With Individual Model Realizations

It is of interest to consider how well the paired trends test [see equation(3)] performs in a controlled setting, where we know *a priori* that trend differences are due to the effects of natural internal variability alone, and are unrelated to model-versus-observed differences in forcing and/or response. We therefore calculate the test statistic d using results from models with multiple realizations of the 20CEN experiment. These tests rely on different 20CEN realizations performed with the same model, and do not involve inter-model differences. In the following, we show results for tests involving tropical $T_{2\mathrm{LT}}$ data. Very similar results are obtained for tests with tropical T_2 data.

For each model for which $n_r(i)$ is greater than 1, we calculate:

$$d\,(j,k), \quad j = 1, \ldots, n_r(i); \quad k = 1, \ldots, n_r(i); \quad j \neq k$$

where j and k are indices over the number of 20CEN realizations. If $n_r(i) = 5$, for

example, there are 20 different combinations of realizations[2]. Here, 11 of the 19 models have multiple 20CEN realizations, yielding a total of 124 different "intra-ensemble" tests.

Since we are dealing with trend differences arising due to internal variability alone, the sampling distribution of $d\,(j,k)$ must have zero mean (Fig. 7A). The standard deviation of the distribution of $d\,(j,k)$ values is 0.51, with maximum and minimum values of ± 1.64. This distribution is very similar to that obtained from the 49 model-versus-RSS paired trend tests (Fig. 7B), and even shows substantial overlap with the distribution of d values from the 49 model-versus-UAH paired trend tests (Fig. 7C).

We conclude, therefore, that differences between tropospheric temperature trends estimated from models and satellite-based microwave sounders are generally consistent with trend differences that we might expect due to internal variability alone. The situation is more ambiguous for tests of differences between simulated and observed trends in $T_{\rm SST}$ or $T_{\rm L+O}$ minus $T_{\rm 2LT}$ (not shown). In this case, values of d are also of similar magnitude to those obtained from intra-ensemble tests, but only if observed differences in surface minus $T_{\rm 2LT}$ are calculated with RSS $T_{\rm 2LT}$ data.

[2]Note that this yields a symmetrical sampling distribution of d values, *e.g.*, $d\,(5,2) = -d\,(2,5)$. This symmetry is useful given the relatively small ensemble sizes available here.

3.2.2 Tests With Multi-Model Ensemble-Mean Trend

It is less meaningful to perform "intra-ensemble" tests with the DCPS07 test statistic. This is because DCPS07's $\sigma_{\rm SE}$ term requires an estimate of the inter-realization standard deviation of trends, which is difficult to determine reliably from a sample size of five or less. Nevertheless, it is instructive to calculate d^* for the specific example of the MRI $T_{\rm 2LT}$ trends shown in Fig. 1. We assume that the trend in realization 1 (0.042°C/decade) is a surrogate for the observed trend. For realizations 2, 3, 4, and 5, the ensemble-mean model trend is 0.340°C/decade, with a standard deviation of 0.043°C/decade. DCPS07's $\sigma_{\rm SE}$ is therefore $0.043/\sqrt{(4-1)} = 0.025$°C/decade. The difference between the trend in our surrogate observations and the ensemble-mean trend is $0.340 - 0.042 = 0.298$°C/decade. This trend difference is nearly 12 times larger than DCPS07's $\sigma_{\rm SE}$, yielding an infinitesimally small p-value. Based on this test, DCPS07 would have erroneously concluded that the MRI model's first 20CEN realization was virtually certain to have been generated by another model.

In another test with model data only, we considered the model results in Figure 3B. We designated the ensemble-mean $T_{\rm 2LT}$ trend in version 3.0 of the Community Climate System Model (CCSM3.0) as the surrogate "observations", calculated $\ll b_m \gg$ and $\sigma_{\rm SE}$ from the ensemble-mean trends in the remaining 18 models, and then computed the DCPS07 test statistic d^*. In this case, $d^* \approx 3$, and hypothesis H_2 is rejected at the nominal 5% level. This procedure was repeated 18 times, with each of the other

models serving in turn as surrogate "observations". In 11 of the 19 DCPS07 tests, H_2 is rejected at the nominal 5% level: *i.e.*, we would conclude that 11 models had trends that were fundamentally inconsistent with the model average signal trend. A test that rejects greater than 50% of the model population used to estimate $\ll b_m \gg$ and $\sigma_{\rm SE}$ is clearly an incorrect means of judging whether model trends are significantly different from observed trends.

Caption for Supporting Figure

Figure 7: Histograms of d, the test statistic for the paired trends test (Section 4.1). All tests involve tropical $T_{\rm 2LT}$ trends over the 252-month period from January 1979 to December 1999. The histogram in panel A is based on results for 124 intra-ensemble tests. These tests involve trend differences between different 20CEN realizations performed with the same model. They provide information on the size of d values expected to occur due to internal climate variability alone. The d test statistic results in panels B and C are for paired trends tests in which each of the 49 model $T_{\rm 2LT}$ trends in Fig. 3A is compared with RSS and UAH $T_{\rm 2LT}$ trends, respectively.

Caption for Supporting Table

Table 7: Statistical significance of differences between modelled and observed tropospheric temperature trends. Results are for the paired trends test described in

Section 4.1. Model data employed in the test are tropical T_{2LT} and T_2 trends from 49 realizations of 20th century climate change performed with 19 different A/OGCMs (together with their associated adjusted standard errors). Observational trends and adjusted standard errors were estimated from RSS and UAH satellite data. There are 49 tests for each tropospheric layer and each observational dataset. Results are expressed as the number of rejections of hypothesis H_1 at stipulated significance levels of 5%, 10%, and 20%. Percentage rejection rates of H_1 (out of 49 tests) are given in parentheses. In the BASE case, all trends and standard errors were calculated over the period January 1979 to December 1999 from time series of spatially-averaged (20°N-20°S) anomaly data. SENS1 results involve the same analysis period, but with model standard errors replaced by observational results. In SENS2, observational trends and adjusted standard errors were calculated over a longer period (1979 to 2006). Model trends were assumed to be the same as over 1979 to 1999, and model standard errors were set to observational values for the 1979 to 2006 period.

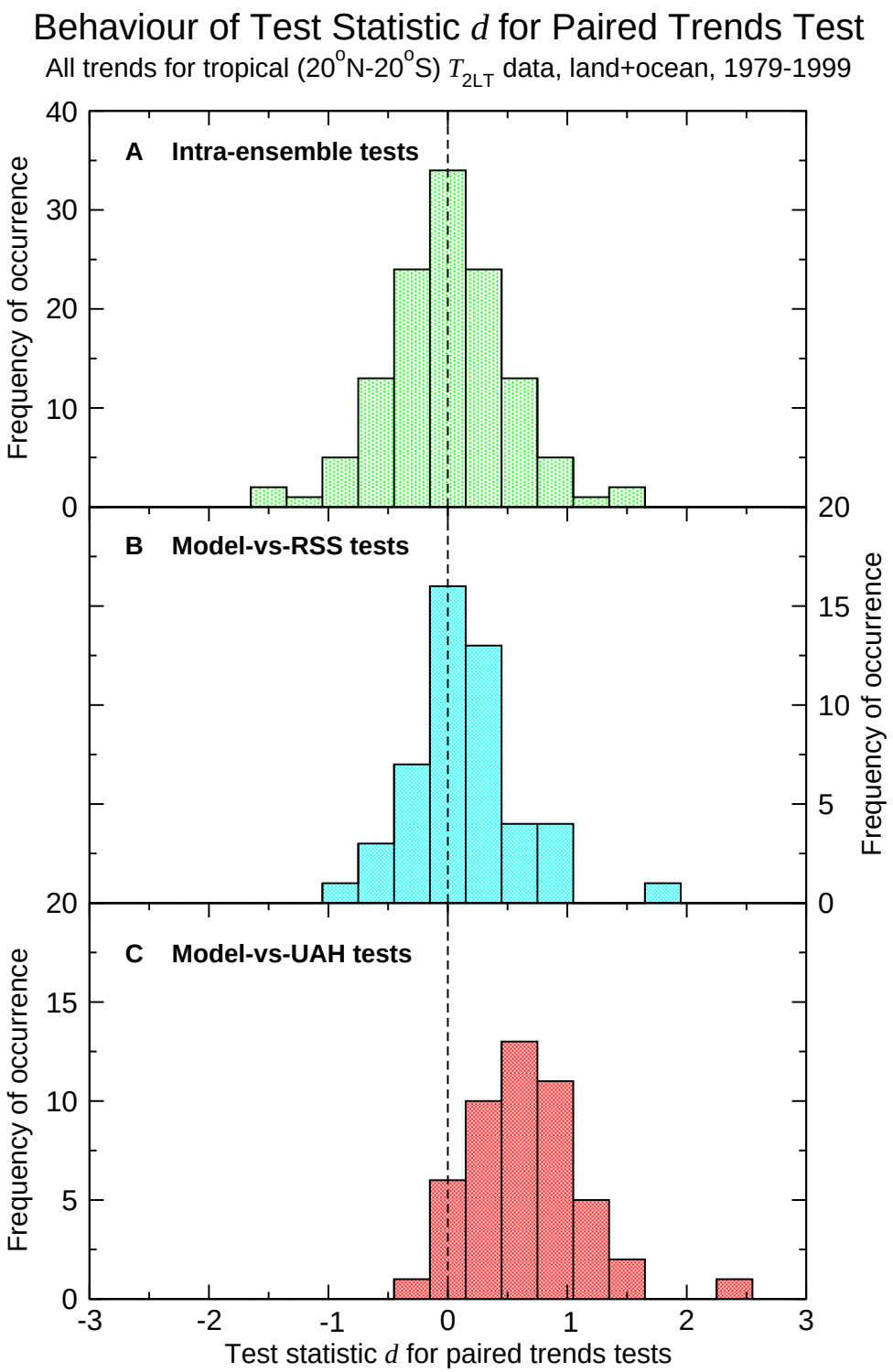

Figure 7: Santer *et al.*

Table 7: Significance of differences between modelled and observed tropospheric temperature trends: Results for paired trends tests

Type	Sig. level	RSS T_{2LT}		UAH T_{2LT}		RSS T_2		UAH T_2	
BASE	5%	0	(0.0%)	1	(2.0%)	1	(2.0%)	1	(2.0%)
BASE	10%	1	(2.0%)	1	(2.0%)	1	(2.0%)	3	(6.1%)
BASE	20%	1	(2.0%)	4	(8.2%)	1	(2.0%)	6	(12.2%)
SENS1	5%	0	(0.0%)	1	(2.0%)	0	(0.0%)	2	(4.1%)
SENS1	10%	1	(2.0%)	4	(8.2%)	1	(2.0%)	6	(12.2%)
SENS1	20%	2	(4.1%)	7	(14.3%)	2	(4.1%)	9	(18.4%)
SENS2	5%	4	(8.2%)	8	(16.3%)	6	(12.2%)	11	(22.5%)
SENS2	10%	7	(14.3%)	11	(22.5%)	8	(16.3%)	16	(32.7%)
SENS2	20%	11	(22.5%)	18	(36.7%)	11	(22.5%)	25	(51.0%)