Climate Dynamics (2003) 21: 573–591
DOI 10.1007/s00382-003-0353-1

**P. W. Thorne · P. D. Jones · S. F. B. Tett · M. R. Allen
D. E. Parker · P. A. Stott · G. S. Jones · T. J. Osborn
T. D. Davies**

Probable causes of late twentieth century tropospheric temperature trends

Received: 17 December 2002 / Accepted: 14 August 2003 / Published online: 24 October 2003

Abstract We assess the most probable causes of late twentieth century (1960–1994) tropospheric temperature changes. Optimal detection techniques are used to compare observed spatio-temporal patterns of near-surface and tropospheric temperature change with results from experiments performed with two different versions of the Hadley Centre climate model. We detect anthropogenic forcings, particularly well-mixed greenhouse-gases, with a less certain sulfate aerosol cooling influence. More limited evidence exists for a detectable volcanic influence. Our principal results do not depend upon the choice of model. Both models, but particularly HadCM3, appear to overestimate the simulated climate response to greenhouse gases (especially at the surface) and volcanoes. This result may arise, at least in part, due to errors in the forcings (especially sulfate) and technical details of our approach, which differs from previous studies. We use corrected and uncorrected versions of the radiosonde record to assess sensitivity of our detection results to observational uncertainties. We find that previous corrections applied to the radiosonde temperature record are likely to have been sub-optimal in only taking into account temporal consistency. However, the choice of corrected or uncorrected version has no systematic effect upon our main conclusions. We show that both models are potentially internally consistent explanations of observed tropospheric temperatures.

P. W. Thorne (✉) · P. D. Jones · T. J. Osborn · T. D. Davies
Climatic Research Unit, School of Environmental Sciences,
University of East Anglia, Norwich, NR4 7TJ, UK

P. W. Thorne · S. F. B. Tett · D. E. Parker
P. A. Stott · G. S. Jones
Hadley Centre for Climate Prediction and Research,
Met Office, Fitz Roy Road, Exeter, EXI 3PB, UK
E-mail: Peter.Thorne@metoffice.com

M. R. Allen
Atmospheric, Oceanic and Planetary Physics,
University of Oxford, Clarendon Laboratory,
Oxford, OX1 3PU, UK

1 Introduction

A number of previous climate change studies have used some form of "optimal detection" approach to determine the most likely causes of recently observed climate. These have almost exclusively considered near-surface temperatures (Tett et al. 1999, 2002a, (hereafter referred to as T99 and T02)); Stott et al. 2001, (hereafter referred to as S01); Hegerl et al. 1996, 1997; Barnett et al. 1999), zonally averaged radiosonde-based upper-air temperatures (Thorne et al. 2002a; Hill et al. 2001; T02; Allen and Tett 1999, (hereafter referred to as AT99); Santer et al. 1996), or a combination of the two (Jones et al. 2003). These studies consistently detect the effect of anthropogenic and natural forcing on twentieth century climate. However, there are numerous other climate indicators and depending upon such a small subset is clearly a limitation. If the results can be repeated for a larger range of independent (or quasi-independent) variables, then our confidence in the reality of a pronounced human effect on climate will be enhanced. Recently investigators have begun to consider changes in ocean heat content (Barnett et al. 2001; Levitus et al. 2001) MSU (Microwave Sounding Unit, Christy et al. 2000; Mears et al. (submitted 2003)) satellite-based tropospheric and stratospheric temperatures (Tett et al. submitted 2002b; Santer et al. 2003a), tropopause height (Santer et al. 2003b), and surface pressure (Gillett et al. 2003).

In this study, we consider full spatial fields rather than just zonally averaged temperature fields within the free atmosphere (Thorne et al. 2002b provides a rationale). We use large area averages (LAAs) to represent these full spatial fields. We employ five-year means of annually averaged modelled and observed LAA data to ensure that we only consider the large spatio-temporal scales at which we have confidence in the models ability to simulate the climate system (Stott and Tett 1998). We consider a total of six tropospheric temperature diagnostics: three layer averages, and three "lapse rates" (the differences between layer averages). The derivation of our input fields is discussed in Sect. 3 along with a brief summary of the optimal detection algorithm.

We use observations of near-surface temperatures from the HadCRUTv record (Jones et al. 2001). We augment this with available tropospheric data from the HadRT2.1 radiosonde temperature record (Parker et al. 1997). For simplicity we define the troposphere as consisting of any values at or below the 300 hPa level, regardless of both season and latitude. We consider the period 1960–1994 both for consistency with previous zonal-mean detection studies (AT99; T02; Thorne et al. 2002a), and because this is the period when we have the highest confidence in the veracity of, and maximum coverage within, the HadRT dataset (Thorne 2001). Observations are compared to model output from two versions of the Hadley Centre's Ocean/Atmosphere General Circulation Model (AOGCM). These models have been run both for long control integrations with no changes in external forcings and for small (generally four-member) ensembles to assess the historical and likely future climate responses to time-varying external (natural and anthropogenic) climate forcing influences. The model and observational datasets are discussed in Sect. 2. At this point, however, we justify the selection of which observed atmospheric temperature dataset to use, and our decision to use tropospheric values only.

There exist at least three potential sources of observed tropospheric temperatures: the MSU satellite-based temperature record (Christy et al. 2000, 2003; Mears et al. 2003), reanalyses using operational numerical assimilation schemes (Kalnay et al. 1996; Gibson et al. 1997), and instruments carried on radiosondes (Parker et al. 1997). The MSU record is near-global in coverage and temporally continuous. However, it exists only for 1979 to date, and it has been shown on several occasions that a longer record increases the chances of successful detection (e.g. Santer et al. 1995; Barnett et al. 1998). Further, there remain potentially significant uncertainties due to orbital decay effects (Wentz and Schabel 1998), and platform changes (Hurrell and Trenberth 1997), amongst others, which may affect the long-term temporal homogeneity of the MSU record (NRC 2000; Mears et al. 2003). Reanalyses are likely to contain significant artificial discontinuities due to increasing data availability through time (Barnett et al. 1999; Santer et al. 1999; Pawson and Forino 1998), particularly the advent of satellite observations in the late 1970s. They are also model based data being output from numerical weather forecast models with time-invariant physics and resolution. They are therefore not truly valid for use in any model-observational intercomparisons. Numerous alternative radiosonde temperature databases exist (Angell 1988; Eskridge et al. 1995; Lanzante et al. 2003a, b for example), but the only currently available globally gridded (although incomplete) product, available in several different versions, is the HadRT record (Parker et al. 1997).

We only consider tropospheric temperatures because we see this region as the most relevant in terms of the climate change debate. We believe that on the large space and time scales considered in the present study the troposphere is well-mixed. Therefore our detection results should be consistent for the different temperature diagnostics if the models are an adequate explanation of the observed trends. We see no logical reason as to why stratospheric temperatures (or any other climate parameter) could not additionally be considered in future studies under our approach. However, the HadRT radiosonde record coverage degrades significantly above the tropopause, particularly early in the record, due to sonde burst. There is also good reason to expect errors within the radiosonde dataset to increase with altitude (Parker et al. 1997). Further, previous model validation studies suggest that both Hadley Centre models may not adequately resolve either natural internal variability or the dynamical responses to forcings within the stratosphere (Gillett et al. 2000; Collins et al. 2001). A consideration of full spatial field stratospheric radiosonde temperatures in detection studies is, therefore, premature at the present. Additionally, very different chemical and dynamical considerations may pertain within the stratosphere, at least potentially making any direct comparison of detection results between the two atmospheric regions non-trivial (see Subsect. 4.4) In discarding the stratosphere we are sacrificing degrees of statistical freedom, but previous zonal mean detection studies have been shown to be insensitive to the inclusion or otherwise of stratospheric temperatures (Thorne et al. 2002a).

In Sect. 4 we consider the results for each temperature diagnostic separately. We focus our analysis on the lower troposphere and the near-surface, as recent differences in trend between these regions have been a source of controversy (see NRC 2000 for a review). Section 5 considers whether any fundamental discrepancies exist between our results for all our input temperature diagnostics. Finally, we present our conclusions in Sect. 6.

2 Observed and model datasets

2.1 Observed datasets

The HadCRUTv dataset (Jones et al. 2001) is available as a monthly 5° longitude by 5° latitude gridded product of global near-surface temperature anomalies relative to 1961–90 from the late nineteenth century to present. Although data availability is discontinuous in both space and time through the record as a whole, it is fairly consistent over the period 1960–1994. HadCRUTv is a modified version of the HadCRUT dataset (Jones et al. 1999) employed in previous detection studies considering near-surface temperatures (Allen et al. 2002; T99; T02; S01; Barnett et al. 1999; Hegerl et al. 1996, 1997; Santer et al. 1995). Variance corrections have been applied to both land and ocean components in HadCRUTv to account for the effects of time-varying sampling (Jones et al. 2001). These corrections should yield more stable EOFs (empirical orthogonal functions) (Jones et al. 2001) an important consideration in optimal detection. Other gridded near-surface temperature products exist (e.g. Hansen et al. 1999, 2001), and it would be useful in future studies to use more than one observational dataset to assess sensitivity to this choice. However, this observational uncertainty is likely to be order of magnitude less than in the upper-air, so we do not consider it here.

The HadRT2.1 record (Parker et al. 1997) is a monthly 10° longitude by 5° latitude globally gridded radiosonde temperature product from 1958 to date as anomalies relative to 1971–1990. Data are available on 9 WMO (World Meteorological Organisation) standard reporting levels. We only consider data from the 850 hPa, 700 hPa, 500 hPa, and 300 hPa levels as higher levels are likely to sample stratospheric air for at least part of the year (outside the tropics). Simple near-neighbour spatio-temporal quality control checks have been performed on the dataset to identify obviously spurious values (Thorne 2001), resulting in the discarding of approximately 5% of the data. There are data gaps in space and time, with much larger areas of missing data than in HadCRUTv. To avoid any potential biases we therefore sub-sample and regrid the HadCRUTv record to match the sparser HadRT lower tropospheric field before undertaking our analyses. This ensures that any differences in our results between the near-surface and free troposphere could not arise due to spatio-temporal sampling considerations alone (see Santer et al. 2000 for a more detailed justification). The resulting observed datasets are highly biased in their coverage towards land and, particularly, Northern Hemisphere mid-latitude continental regions (Fig. 1).

To assess the sensitivity of our results to likely observational uncertainties in the free troposphere we consider two versions of the HadRT dataset: HadRT2.1 and HadRT2.1s. Both have been corrected globally for significant change-points in individual station series post-1979 at the time of known observational practice changes (Gaffen 1996), with reference to co-located MSUc data (Christy et al. 1998) as described in Parker et al. (1997). HadRT2.1 has had these corrections applied within the troposphere, whereas HadRT2.1s has not. A consideration of both versions should permit an assessment of the sensitivity of our results to the effects of the correction approach.

We caution that our datasets do not fully sample observational uncertainty. Numerous other both radiosonde and satellite-based upper-air temperature products exist. These differ both in terms of their raw input data and their treatment of suspected inhomogeneities. The resulting range of climate trajectories is larger than that spanned by our two versions of the HadRT dataset (Seidel et al. submitted 2003 provide an intercomparison of different datasets). It would be highly desirable to repeat our analyses using additional datasets to fully assess the sensitivity of our results to dataset treatment and choices.

2.2 Model datasets

Two versions of the Hadley Centre's AOGCM are considered: HadCM2 (Johns et al. 1997), and HadCM3 (Pope et al. 2000; Gordon et al. 2000). Although these are different versions of the same centre's GCM, sufficient differences exist between the two generations, and the forcing histories applied for equivalent scenarios for each model (S01; T02, respectively), to justify treating them as being at least quasi-independent. Both models have a resolution of 3.75° longitude by 2.5° latitude in the horizontal in their atmospheric components. HadCM3 has a finer resolution ocean component (1.25° longitude by 1.25° latitude). Both models consist of 19 layers within the atmosphere. We simply sub-sample to our four HadRT2.1 tropospheric levels and the near-surface, and bilinearly interpolate the model data on each of these levels to the coarser resolution observational grid. We then mask out values where HadRT or sub-sampled HadCRUTv data are missing.

Both models have been integrated in a long control run with no changes in external forcings. These exhibit essentially no long-term drift in global-mean quantities. However, flux adjustments are required for HadCM2 to avoid it drifting into an unrealistic state. In addition both models have been used to predict the time-varying response to a number of candidate external forcing mechanisms. In the present study we limit ourselves to the subset of forcing realisations which have direct equivalence between the models. However, we stress that differences exist between the models in the manner in which these forcings are applied (see S01 and T02 for more details). In particular, HadCM3 has more realistic greenhouse gas and sulfate aerosol forcings than HadCM2. The model experiments considered are summarised in Table 1 and are based on 4-member ensembles. From here onwards the HadCM3 scenarios are named using their HadCM2 equivalents to facilitate intercomparisons. The acronyms for the forcings considered are therefore: G (greenhouse gases), GS (as G, but additionally including Sulfate aerosol effects), GSO (as GS, but incorporating stratospheric ozone depletion effects post-1979), VOL (volcanic forcing), and LBB (solar forcing). These are different to the acronyms used in T02, but results are for the same forced model runs and can be compared directly.

3 Methodology

We employ the optimal detection technique of AT99 (see also Allen and Stott 2003; T02) to a series of 3-dimensional (latitude, longitude, time) input fields. The technique is an application of ordinary least squares (OLS) multiple regression to ascertain simulated model signal scaling estimates required to recreate a best fit to the observational dataset. The scaling estimates are the weightings required on the individual model signals, based upon ensemble means. Normally, as is the case here even for our spatially averaged input fields, the dimension of the input fields (70 for our preferred diagnostic) is greater than the number of independent noise samples used to perform optimisation. Therefore the optimisation and analysis is undertaken only in the phase-space of the leading EOFs

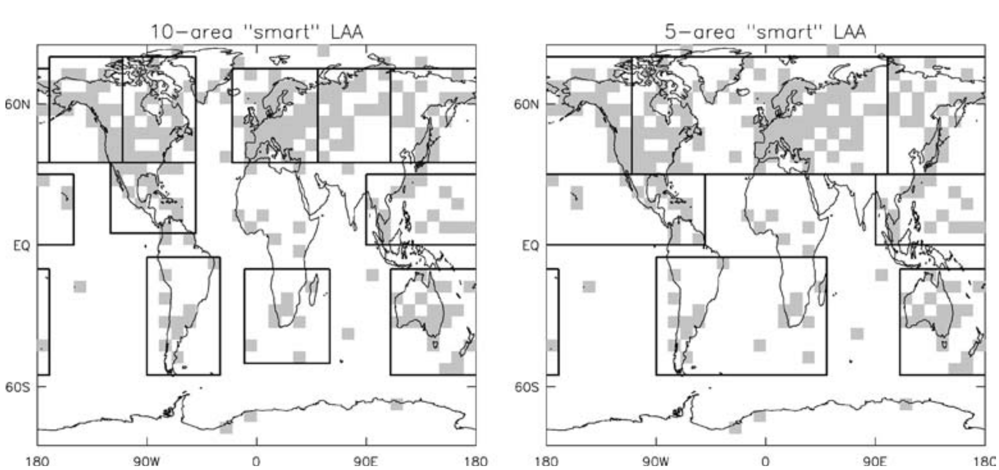


Fig. 1 Areas selected for *smart* large area average diagnostics. The *shaded areas* indicate grid boxes (5° latitude by 10° longitude) which contain values in the input fields for at least one of the seven 5-year periods considered

Table 1 Brief summary of the model signals used in the current study. The detection method can be used to gain, for example, an estimate of the S signal by recombining G and GS under the assumption of no non-linear actions between the forcings. See T02 for further details

Forcing	HadCM2	HadCM3
Well-mixed greenhouse gases	G 1860–2100 (All greenhouse gases treated as CO_2 equivalent concentrations)	GHG 1860–2100 (Individual constituent greenhouse gases considered.)
Greenhouse gases plus anthropogenic sulfate aerosols	GS 1860–2100 (As G, but additionally considering the direct effect of sulfate aerosols)	TROP-ANTHRO 1874–2000 (As GHG, plus includes effects of sulfate aerosols and the impact of tropospheric ozone changes)
Greenhouse Gases plus Anthropogenic sulfate aerosols plus changes in stratospheric ozone	GSO 1860–1995 (Identical to GS until late 1970s when the additional effects of stratospheric ozone depletion are included)	ANTHRO 1860–2100 (Identical to TROP-ANTHRO until late 1970s when stratospheric ozone depletion effects are incorporated)
Solar	LBB 1860–1996 (Solar forcing history based upon the reconstruction of Lean et al. 1995)	SOLAR 1860–1999 (Solar forcing history based upon the reconstruction of Lean et al. 1995)
Volcanic	VOL 1890–1997 (Volcanic forcing history based upon the dust veil index of Sato et al. 1993)	VOLCANIC 1860–1999 (Volcanic forcing history based upon the dust veil index of Sato et al. 1993)

of simulated climate variability as estimated from a segment of the model control run in this input field space. The number of EOFs considered is termed the truncation dimension. Optimisation is undertaken by rotating the fields such as to maximise their SNR (signal to noise ratio) in this EOF space by downweighting high variability modes.

OLS yields both a best-estimate of the individual model signal scalings and their associated uncertainties. From this it is possible to determine whether each model signal is detected in the observations, in which case the uncertainty range in the estimate will encompass only positive values. Detection is assessed at the (two-tailed) 90% level, such that there is a ∽5% chance of a false positive result. Further, we can begin to attribute the observed changes to a given set of forcings. All potentially important forcings have to be considered to be able to unambiguously attribute the observed changes to a set of forcings. We are confident that the ensembles we consider contain responses to what are currently considered the most plausible major climate forcing factors of the late twentieth century.

Two distinct approaches have arisen to attribution. The first specifies that the scaled model signal estimates should all be statistically consistent with the observations (e.g. Hegerl et al. 1997), such that the uncertainty in the scaling encompasses unity. This approach is difficult to quantify in a statistical sense as it is couched backwards. We are testing for the non-rejection of our original null hypothesis, used to gain our uncertainty limits, that the model signal strength is indistinguishable from zero. Rejection of the null at probability *P* does not imply acceptance of the alternative (that the model signal is consistent with the observations) at 1 – *P*, as the statistic is assymetric, and any such claims are likely to be over-optimistic (Levine and Berliner 1999). The second approach simply attributes the observed changes to the most parsimonious set of signals which are detected simultaneously (e.g. T02). There is no requirement for the model signal scaling estimates to be "consistent" with the observations under this approach. However, it is assumed that any model uncertainty arises primarily within the amplitude and not the pattern of the modelled response. Barnett et al. (1999) show a large degree of similarity in the leading EOFs of response to GHG between a number of models, though it is less certain that responses to sulfate aerosol or other forcings will exhibit similar behaviour. Therefore it is assumed that the model signal responses can simply be scaled to recreate the observations and, potentially, to constrain predictions of future climate changes (e.g. Allen et al. 2000, 2002; Stott and Kettleborough 2002).

In the present study, where we are comparing results from distinct tropospheric temperature variables, we feel that the former definition of attribution (Hegerl et al. 1997) is more useful. We use our "attribution" criteria of model signal "consistency" with the observations to flag where the models are likely to be significantly over- (scaling required significantly less than 1) or under- (scaling required significantly greater than 1) estimating the magnitude of the forcing response under the caveat that it is likely to be a weak test yielding too many consistent estimates (see previous discussion of attribution approaches). For any given signal an adequate model would not significantly over-estimate the response in some temperature diagnostics and under-estimate it in others. We are assuming that our detection analyses are free from (gross) systematic analytical biases. Such biases could arise due to methodological considerations (Allen and Stott 2002), sampling artefacts, imperfect model forcings, or significant model and observational error. Furthermore, in some cases a significant over- or under-estimate of the amplitude of the forcing response in a given temperature diagnostic by a model may arise due to chance alone. We stress that the choice of attribution approach does not significantly impact our principal results.

The AT99 detection approach includes a consistency check on the residuals of the regression against an independent realisation of natural climate variability gained from a separate section of model control run. The residuals of the regression should be statistically indistinguishable from this independent estimate of natural internal climate variability. If this test does not pass then we are likely to be incorporating EOFs of control space which are unrealistic (undersampled in the control or poorly captured in the model) and we flag the results as being likely to be dubious. In addition to a consistency check, the OLS approach also includes tests for signal degeneracy (T99) and low SNRs (T02), both of which can significantly bias the results. Inputting degenerate (essentially similar) signals leads to highly uncertain and poorly constrained estimators. For weak (noisy) signals there are known problems with the AT99 approach, whereby the estimates tend to be biased towards zero, affecting the chances of successful detection and attribution (Allen and Stott 2002). We explicitly calculate SNRs following the approach of T02 and flag those cases where they are low and, therefore, likely to lead to significantly biased estimators. Discussions of the various technical aspects of the AT99 optimal detection approach have been extensively addressed (AT99; T99; T02; S01; Allen and Stott 2002), and so are not covered in any more detail here. The reader is directed to Sect. 4 and Appendix C of T02 for a detailed description. Here we discuss further solely those input-field pre-processing and analysis aspects which are novel to the present study.

We consider six indicators of tropospheric temperatures: three "layer average" temperatures and three "lapse rate" temperatures. The layer averages are defined as the surface (SURF henceforth), lower troposphere (mass weighted average of 850 and 700 hPa data, LT henceforth), and upper troposphere (mass weighted average of 500 and 300 hPa data, UT henceforth). We caution that our LT diagnostic is not equivalent to the TLT temperature diagnostic in the MSU temperature series (Christy et al. 2003). The three lapse rate diagnostics are simply the differences between these individual layer series, defined as upper layer value minus lower layer value and only calculated for those grid boxes where data exist for both layers. The lapse rates are: free tropospheric lapse rate (UT-LT henceforth), entire tropospheric lapse rate (UT-Surf henceforth) and lower tropospheric lapse rate (LT-Surf henceforth).

Previous studies employing the approach of AT99 using longitude, latitude, time fields of near-surface temperature have considered a spherical harmonic representation of the decadally averaged gridded data to ensure that they only consider large-scale climate features (Allen et al. 2002; T99; T02; S01). Given the sparsity of observational dataset coverage in some regions of the globe (particularly oceanic and southern high latitudes) in our datasets (Fig. 1), it is unlikely that the spherical harmonic approach would be stable. However, all that is required is a representation of the observed and modelled climate variations at the large space and long time scales at which we have confidence in the ability of the models (Stott and Tett 1998). We therefore choose to use large area averages (LAAs henceforth) (Fig. 1) to represent the spatio-temporally varying observed and model temperature fields. In all cases we use five-year averaged segments of data (1960–1964, 1965–1969 etc.). To derive a 5 year mean value at least three years in each five must contain at least two months with data in three of the four seasons. We use five year periods rather than decadal periods to retain a sufficient time resolution for a full spatio-temporal analysis.

To assess the sensitivity of our results to the choice of LAAs representation we consider four different sets of LAAs, two *smart*, and two *na*. The *smart* LAAs try to take into account dataset coverage (hence the name *smart*) such that areas are chosen to represent approximately equal numbers of datapoints, and as far as possible, what we consider to be climatologically distinct regions. Therefore *smart* LAAs are heavily skewed towards Northern Hemisphere mid-latitude continental regions (Fig. 1). Furthermore, in undertaking such a procedure we omit a few of the tropical stations where recent differences between the surface and the troposphere have been shown to be most pronounced (Hegerl and Wallace 2002; Thorne et al. 2003). Our preferred *smart* set is defined by 10 regions, with a second choice limited to five much larger regions. The use of more areas in our preferred spatial preprocessing option should increase the power to discriminate between competing forcings, reducing the chances of signal degeneracy. In addition we consider two *na* diagnostics whereby we make no assumptions about the spatial coverage other than it being insufficient south of 30°S. We simply split the remaining areas of the globe equally into 6 and 12 areas (30°S–0°S, 0°N–30°N, and 30°N to 90°N, and then 2 (0°E–180°E and 180°E–360°E) and 4 (0°E–90°E, 90°E–180°E, 180°E–270°E and 270°E–360°E) equal longitude intervals respectively (see Jones et al. 2003)). Given that the number of contributing data points will vary widely between these areas, and that they are unlikely to describe distinct climatological units, we consider that such an approach is likely to be less optimal than our *smart* approach.

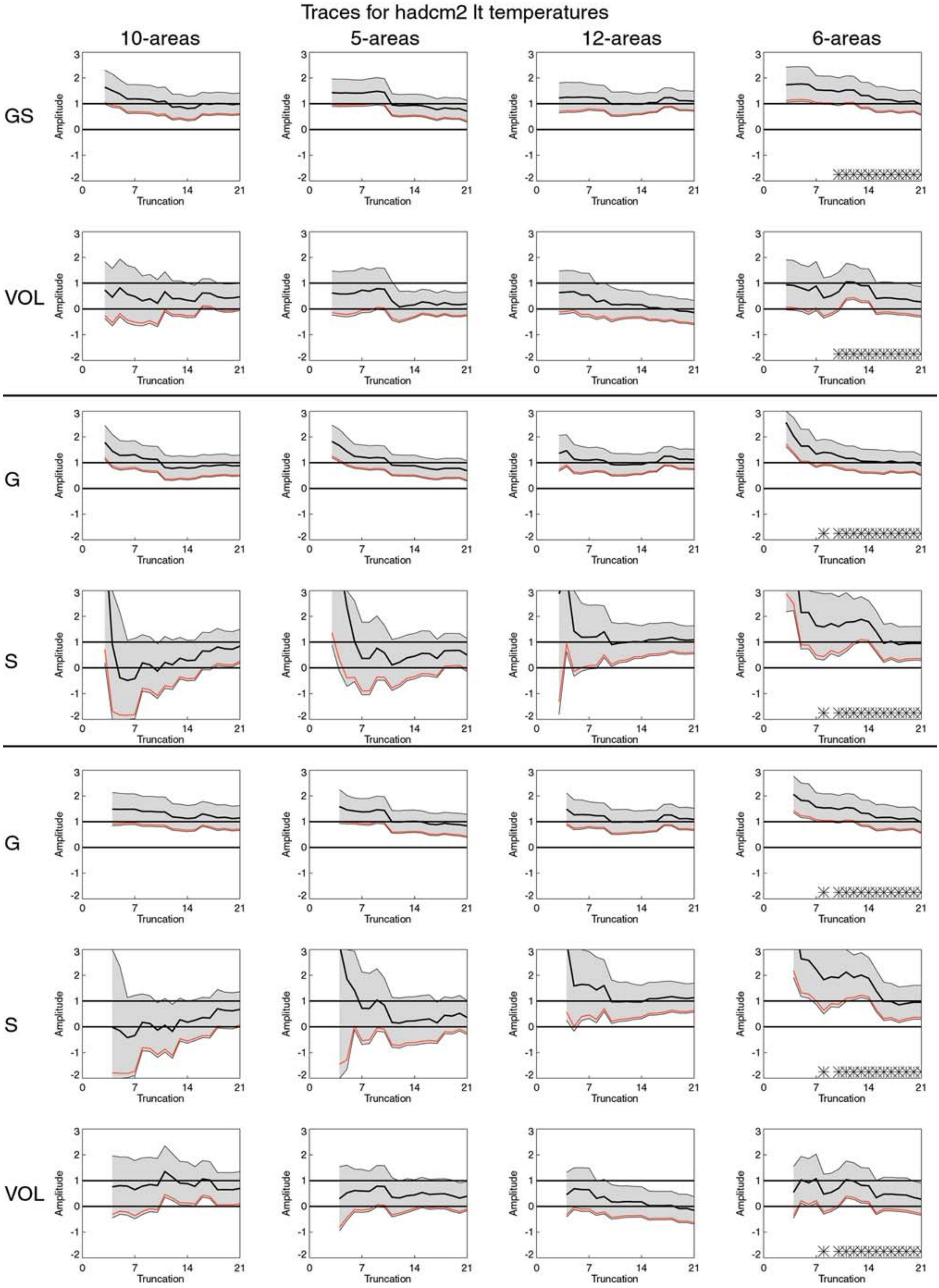
We use what we define as detection "traces" (see Fig. 2) to outline our principal results and assess their robustness to the choice of truncation. These "traces" show how our signal scaling estimates and their associated 90% uncertainty ranges change with increasing truncation from 4 to 21. The power of the detection approach increases with truncation (North et al. 1995; Hasselmann 1993) as more modes of variability are considered. The first few truncations tend to be dominated by global-scale variability, whereas at latter truncations we tend to incorporate more regional-scale modes. This may significantly impact at what truncations individual signals become detectable. For example, S has a more regional response pattern than G in both models considered here. Therefore we might expect G to be both better defined and more detectable at much earlier truncations than S. Conversely, G may become undetectable at high truncations as we concentrate more on regional modes where the model may fail to adequately capture the response. In some cases the consistency test on the residuals fails, and such cases are clearly marked. We consider all four LAA cases for both models. We claim that a given model signal is robustly detected if it is detected in all four LAA representations and at most truncations in our traces. We also assess model signal consistency with the observations (our attribution approach) in a similar manner.

Global-mean trend reconstructions are calculated based upon our preferred 10-area *smart* LAA diagnostics. The reconstructions are performed using the highest truncation (largest number of principal modes considered) at which the test on the residuals passes for each model. The optimised signals, weighted by the scaling factors, and observations are used to recreate an estimate of the true observed and model-predicted global-mean timeseries. Unless models are identical in every respect, the reconstructed observed trend will differ between models, as their projection onto their leading modes will not be identical. The global-means are also not true global-means, being limited to the available 10-area *smart* LAA data, which is spatially sparse.

4 A consideration of individual temperature diagnostics

We illustrate our results by concentrating on SURF and LT diagnostics in view of the recently observed trend discrepancies from 1979 to date between these layers, which cannot to date be unambiguously explained by model predictions (see Jones et al. 1997a; Santer et al. 2000, 2001; Thorne et al. 2003). However, we consider the longer period of 1960–94. This is important as there is evidence that pre-1979 the observed "discrepancy" was reversed (Jones et al. 1997a; Parker et al. 1997; Angell 2000; Brown et al. 2000; Gaffen et al. 2000). We also have relatively little weighting in our *smart* LAA diagnostics within the tropics, and most of the recently observed global mean trend discrepancy arises within this region (Hegerl and Wallace 2002; Thorne et al. 2003). Therefore, we might not expect to see a discrepancy between our detection results for these two layers, even if such a discrepancy exists on other space and time-scales in the real world.

In all cases results shown are those using HadRT2.1s tropospheric data. We find that for all upper air temperature diagnostics except UT-LT the use of HadRT2.1 as the observed radiosonde dataset results in an increased frequency of failure of the consistency test on the residuals, whilst having negligible impact upon our estimates of the signal scalings. This is most likely because the corrections applied to HadRT2.1 were undertaken with only a temporal rather than spatio-temporal consistency requirement (Parker et al. 1997). We cannot however rule out any errors in MSU data used as a reference in the corrections also affecting our results. Fields of radiosonde instrumental biases following corrections, which were applied in a patchy manner (only when metadata, which is incomplete for the vast majority of stations, existed and was coincident with a statistically significant difference), are unlikely to correspond to the leading modes of variability in either model (or the real world). Thus they will tend to artificially inflate the residuals term in our analyses whilst having negligible impact upon our scaling estimates (Thorne 2001). It would be useful to have a radiosonde dataset for which spatio-temporal consistency is maintained and robust error estimates are available.


Traces for hadcm2 lt temperatures
10-areas
5-areas
12-areas
6-areas
GS
VOL
G
S
G
S
VOL
Amplitude
Truncation
3
2
1
0
-1
-2
0
7
14
21

◀

Fig. 2 Changing amplitude scaling estimates with increasing truncation for HadCM2 lower tropospheric model signals. The OLS regression signal amplitude estimates are denoted by the *bold lines*, with 90% uncertainty ranges for assessing signal consistency denoted by *shading*. Detection confidence limits are denoted by a *red line*, detection occuring when above zero. Detection and consistency limits differ for methodological reasons (S02). *Asterisks* mark "inconsistent" residuals, indicating likely problems in our analysis. *Top two rows* relate to GS + VOL, the *middle two* G + S, and *lower three* G + S + VOL

4.1 Deciding which signals to be considered

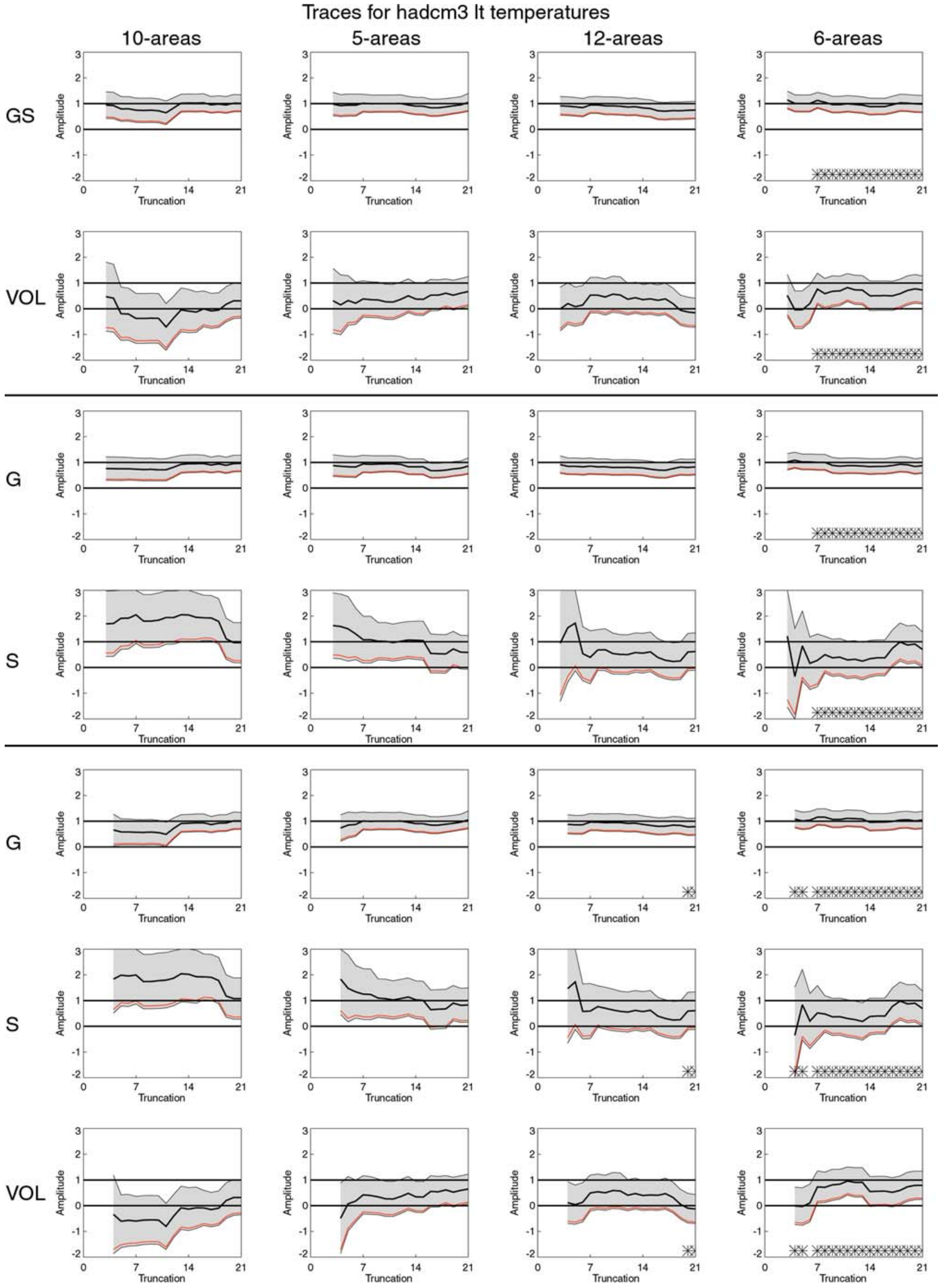
We begin our analysis by summarising detection results for all possible combinations of input signals for each tropospheric temperature diagnostic. Clearly with a very large matrix of potential one to five input signal combinations (there are 25 combinations in total) it is sensible to focus solely upon the most plausible explanations in our detailed analysis. Analysis is undertaken using either truncation 21 (the estimated maximum number of degrees of freedom of the control sections in the shortest model control used for optimisation (1.5 times the number of independent 35 year sections, AT99)), or the maximum truncation at which the test on residuals does not fail, if this is reached before 21. We limit ourselves to our two sets of *smart* LAA diagnostics in this initial analysis. In addition to using ordinary least squares (OLS) regression, we also use total least squares (TLS) (Allen and Stott 2002) to assess whether results are sensitive to implicitly accounting for the presence of noise in our small ensemble model signals.

Results differ markedly in detail between temperature diagnostics (Thorne 2001, gives much greater detail than is possible here), with signals in general being less detectable for SURF and, particularly, lapse rate variables than for UT and LT. The TLS approach leads to a systematic decrease in signal detectability, although this arises almost entirely in those cases where input signals are identified as being potentially degenerate. Importantly use of OLS versus TLS does not influence the relative detectability of different signal types. So we later proceeded using solely OLS analyses as these are less likely to be unduly influenced by the likely presence of non-negligible observational errors in the HadRT data than TLS analyses (Allen and Stott 2002). We find that the most detectable signals according to both models and all temperature diagnostics are, in descending order; G (or GS/GSO), S, VOL, LBB, O (see Thorne 2001 for details of this analysis). The few detections of LBB and O are rare enough to have occured by chance alone. Therefore we discount these forcing mechanisms as being detectable explanations of the observed climate changes. In doing so, we are assuming that the models do not grossly fail to capture the surface and tropospheric temperature response to either forcing. Our choice of temporal averaging period may reduce our chances of successful detection of an LBB signal as it is approximately half the solar sunspot cycle length. We subsequently consider in greater detail, using all four LAA representations, detection results for the following three input signal combinations: G + S, GS + VOL, and G + S + VOL.

4.2 Lower tropospheric temperatures

Detection traces for LT temperatures are shown for HadCM2 in Fig. 2. For the GS + VOL signal combination, GS is both robustly detected and consistent with the observations, except with the 6-area *na* LAA averages combined with high truncation. The VOL signal is only very occasionally detected, and the response is significantly overestimated within the model. Analysis of SNR values (not shown here), indicates that this is highly unlikely to be due solely to a weak model volcanic signal leading to a biased estimator. However, in the real-world El Chichon and Pinatubo were coincident with ENSO events (although weak at the time of Pinatubo) which might have mitigated the observed response leading to an apparent bias towards both models overestimating the VOL response in all tropospheric temperature diagnostics we consider (Brown et al. 2000; Santer et al. 2001). G is robustly detected in the G + S input signal combination and in all four LAA diagnostics. In the LT observations, at almost all truncations, G has an amplitude consistent with unity. The S signal is less robustly detected, although it too is consistent with unity in all LAA diagnostics, and at the majority of truncations. The S signal is more robustly detected in the *na* diagnostics as these better sample hemispheric assymetry which is an important component of the S signal. Individual signal components of the three-signal regression generally track their estimates from the two-signal regressions discussed. The overestimation problem for VOL within the model is partly reduced when G and S are allowed to vary rather than being a fixed ratio (GS), leading to a slightly increased frequency of detection of a VOL signal. This result is seen in cases where the scaling estimate on the S signal is close to zero in the 10-area *smart* LAA diagnostic (although not for other input LAA diagnostics). Therefore, there may be a degree of co-linearity between the responses to S and VOL leading to degenerate solutions (confirmed by standard degeneracy tests, T99).

GS + VOL signal combination detection traces for HadCM3 LT temperature fields, shown in Fig. 3, indicate robust detection of a GS signal,which is consistent with unity. Results for VOL are more uncertain: the VOL signal is rarely detected at truncations that pass the consistency test for residuals. HadCM3 also tends to overestimate the amplitude of the VOL response. In our preferred *smart* 10-area LAA diagnostic, the best-guess VOL signal scaling estimate is negative over a range of truncations. The inverse of the signal, a warming of the lower troposphere with volcanic events, is required to best explain the observations at these truncations which, on purely physical grounds, is unlikely to be correct.


Traces for hadcm3 lt temperatures
10-areas
5-areas
12-areas
6-areas
GS
VOL
G
S
G
S
VOL
Amplitude
Truncation
3
2
1
0
-1
-2
0
7
14
21

◀

Fig. 3 As Fig. 2 except for HadCM3 signals

This is unlikely to be an artefact of a weak signal, as SNR analysis shows that the signal is distinguishable from natural variability. However, our confidence limits still encompass positive values, so we do not discount a VOL influence. Furthermore, this result is not repeated for our other LAA representations. Nevertheless, our confidence in a VOL influence on temperatures within the LT is reduced. For HadCM3, in the G + S signal combination, G is both robustly detected and consistent with the observations. S is also robustly detected in *smart* LAA diagnostics but not *na* LAA diagnostics, and in all cases it is consistent with the observations. The S signal estimates are very poorly constrained at low truncations, improving marginally with increasing truncation. Estimates of the S signal amplitude are also seen to vary widely between LAA diagnostics, with the 10-area *smart* LAA diagnostic being an outlier yielding consistently larger scaling estimates. Reasons for this behaviour are unknown, but it is unlikely to be due to signal covariance leading to overestimated compensating G and S amplitudes as there is no similar shift in the G scaling estimates. This underscores the importance of exploring the sensitivity of results to a broad range of pre-processing choices. Three-way regression results for HadCM3 mirror those for the individual components discussed.

Global-mean LT temperature reconstructions are shown in Fig. 4 for both models. Reconstructions based upon HadCM2 input fields are unable to resolve the observed minimum in the mid-1970s, instead producing a more monotonic increase. This may be due to known inaccuracies in the sulfate aerosol forcing history in HadCM2 or its crude parameterisations of their effects. For HadCM2, both sulfate and volcanic forcings help to explain the observed maximum in global-mean LT temperatures in the late 1980s, with Pinatubo and sulfate effects causing a slight cooling in the early 1990s. Without accounting for volcanic influences, the increase in reconstructed global-mean LT temperatures is more monotonic than observed, implying that volcanic influences may be an important component of recent LT temperature trends (a finding in agreement with Santer et al. 2001). All HadCM3 reconstructions fall within the uncertainty range of the observations at all times. The volcanic influence is almost zero, except for Pinatubo in the early 1990s when there is a cooling of a few hundredths of a degree in the five year average. At the truncation being considered the weighting on the VOL signal is close to zero in HadCM3 so we are down-weighting the VOL component in the HadCM3 reconstructions vis-à-vis those for HadCM2. Most of the warming trend is derived from greenhouse-gases, although this is moderated by the cooling effects of sulfate aerosols, which are particularly important in explaining trends early in the period. HadCM3 better captures the temporal response possibly due to a more detailed S forcing history than HadCM2 (T02, S01). These global mean plots confirm the principal "detection" findings in that greenhouse gases are necessary for both models to adequately explain global-mean observed warming trends. Sulfate aerosols and volcanic influences are less necessary, as they both tend to cool the LT over this period and explain local and more high-frequency detail.

4.3 Near-surface temperatures

It is instructive to compare the current analysis for SURF with previous work comparing signals from the same models with the non-variance corrected HadCRUT series (T99, T02; S01). These studies have consistently yielded detectable well-mixed greenhouse gas (G) and, more tentatively, sulfate aerosol (S) influences for the latter half of the twentieth century. There is also evidence for solar and volcanic influences on SURF, although detection of these is shown to be sensitive to methodological choices, at least for HadCM2 (S01). These studies considered the full available observed field rather than our more data-sparse sub-sampled representation. We also use: a representation of spatial patterns other than spherical harmonic coefficients; temporal sampling other than decadal resolution; and a consideration of shorter than 50-year trend lengths. Effects of these differences in approach cannot be completely separated in the present analysis (see Gillett et al. 2002 for an example as to how to quantify such effects).

Detection traces for HadCM2 SURF temperatures are given in Fig. 5. For GS + VOL, GS is both occasionally detected and marginally consistent with the observations for some, but by no means all, truncations for each LAA diagnostic. Best-guess scaling estimates for GS in HadCM2 are consistent between LAA diagnostics in suggesting that the simulated signal must be approximately halved to provide a best fit to the observations. This is at odds with previous analyses (see Fig. 2 of T99) which indicated that the GS SURF signal was consistent with the observations. It is also at variance with LT results, which showed that the HadCM2 GS signal scaling was consistent with observations (Fig. 2). There is no evidence that GS SNRs for SURF are systematically decreased compared to GS SNRs for troposphere temperatures, which could lead to negatively biased estimators. This might be expected given that SURF temperatures are likely to contain many more degrees of freedom (Jones et al. 1997a, b) than LT. The HadCM2 VOL signal scaling in SURF is zero or close to zero in all four LAA diagnostics, and includes negative values, implying a non-significant volcanic influence. Considering traces for the G + S signal combination, both signals tend to be overestimated in amplitude in HadCM2 in all LAA diagnostics at all truncations, although S is not significantly overestimated (the scaling estimate is not significantly less than unity). G is occasionally detected in all four input diagnostics, whereas S is detected only once (in the

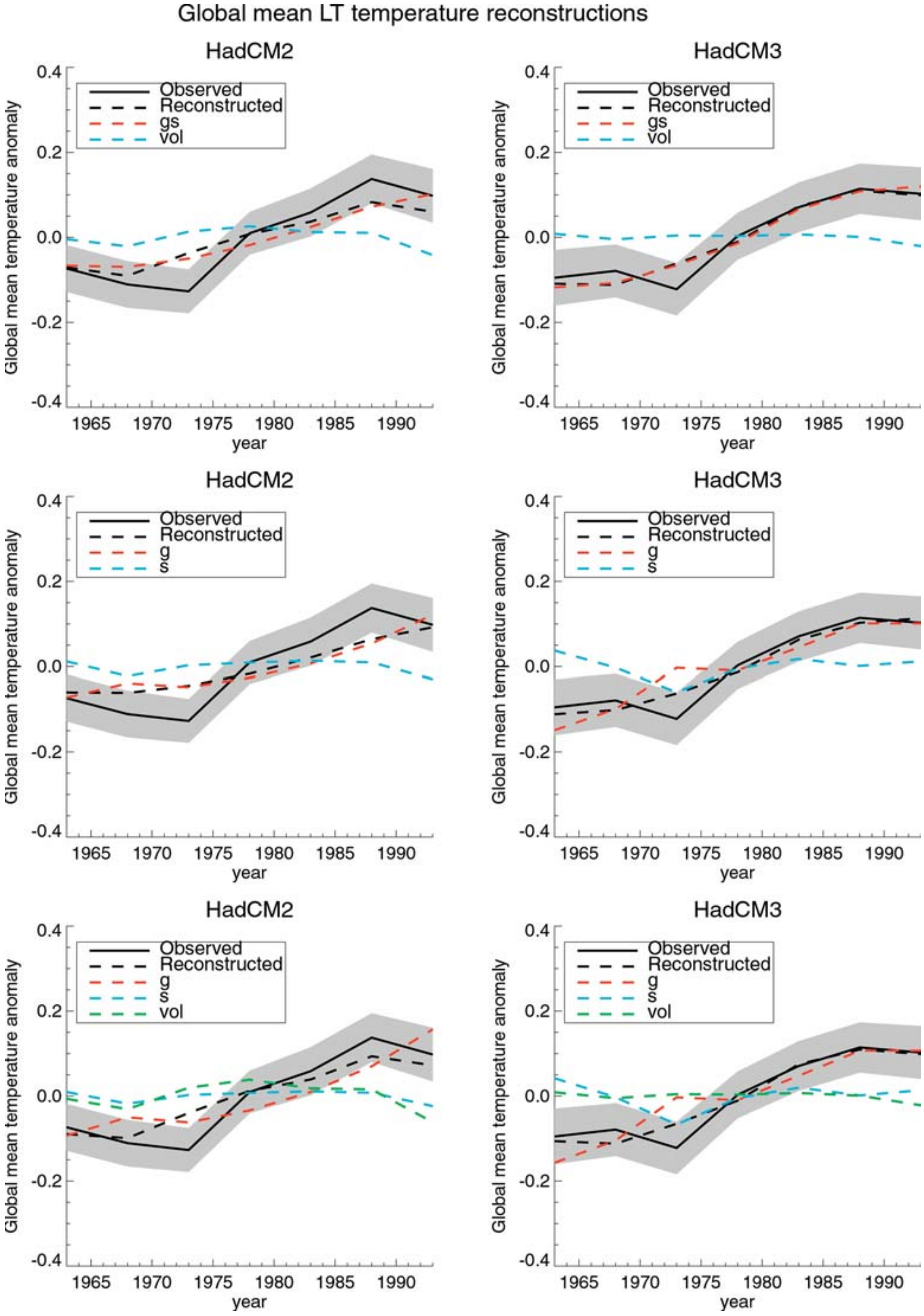


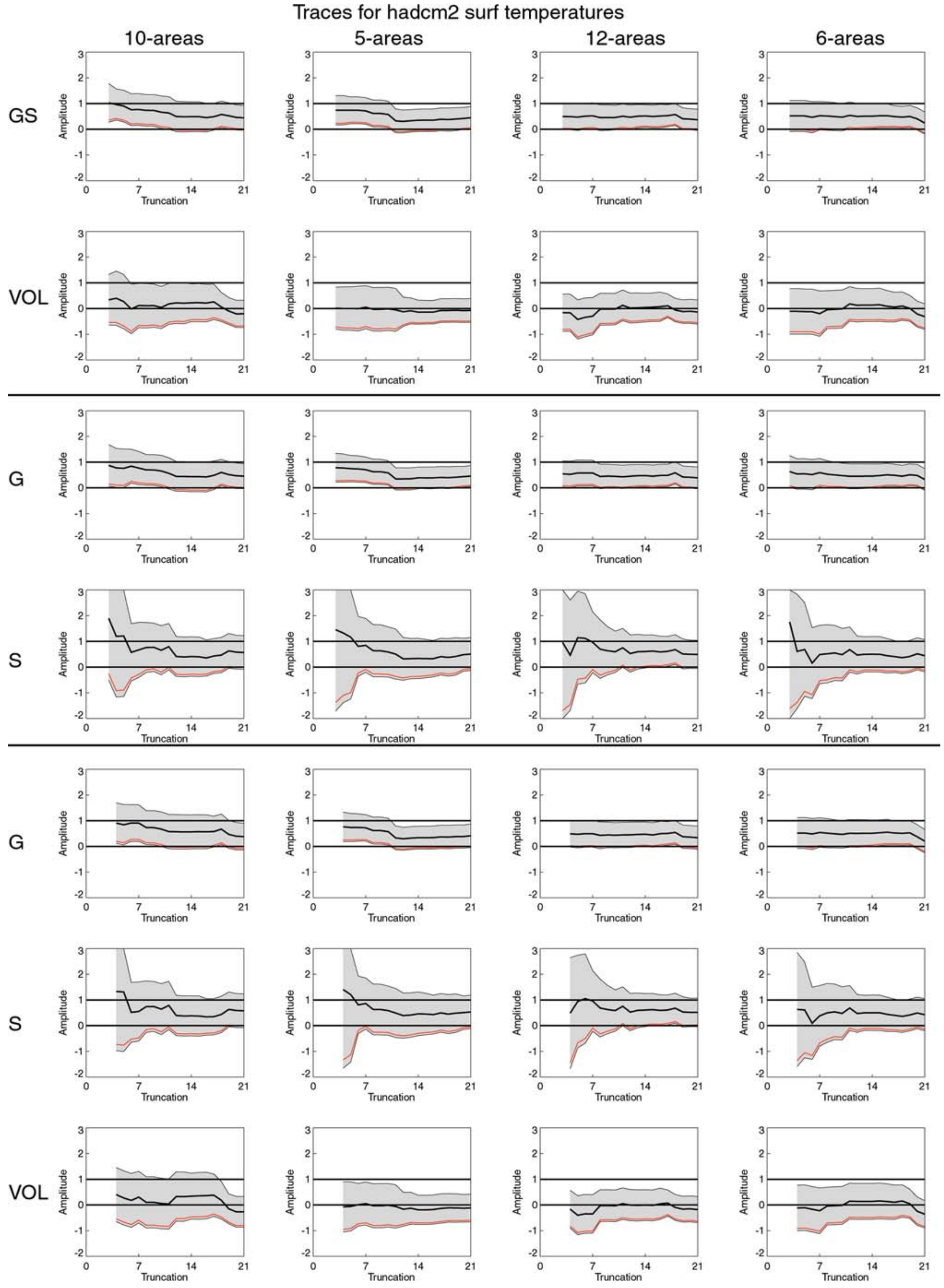
Fig. 4 Observed and reconstructed global-mean LT temperature series based upon HadCM2 (*left panels*) and HadCM3 (*right panels*). The "observations" are projections of observations onto the leading modes of model simulated internal variability, and therefore differ between models. The reconstructions are based upon the sum of the signals in the phase space of the leading modes multiplied by their amplitude estimates arising from the regression

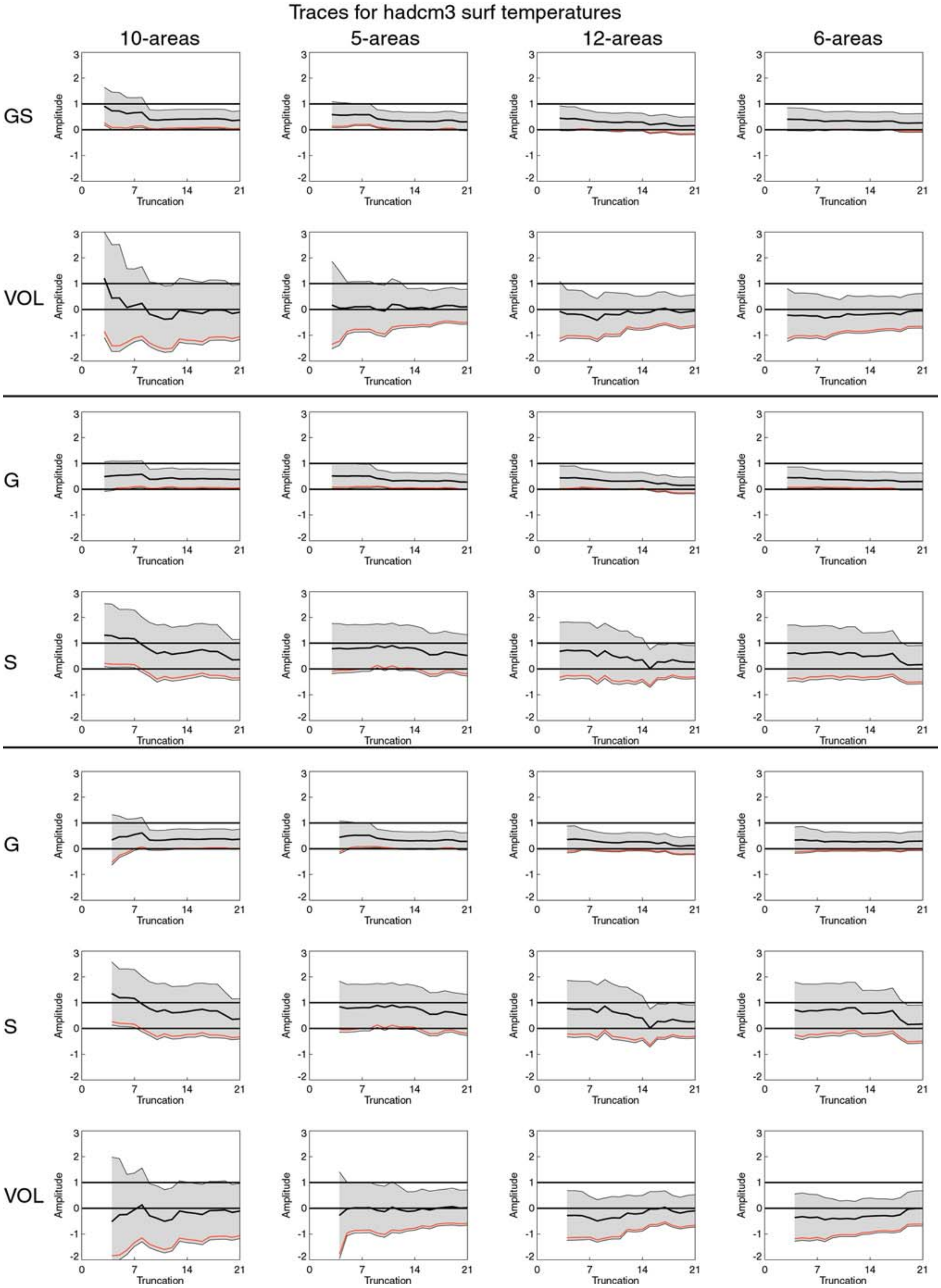
12-area *na* LAA diagnostic). The S signal is very noisy (SNR of about 2), particularly at low truncations, for all LAA input representations and, therefore, our estimates may be negatively biased (T02). However, the GS signal result in the GS + VOL combination (which is dominated by G) is similar to G in G + S, so a negatively biased S estimator, (which could cause a negatively biased G estimator) is unlikely to explain the observed overestimation of both G and S signals. It is also possible that the S signal pattern is significantly in error given the uncertainties in this forcing. Individual components of the three-way regression for SURF track their components in the two-way regressions described.

HadCM3 detection traces for SURF are given in Fig. 6. As for HadCM2, GS, when considered in combination with VOL, is only occasionally detected and it is significantly overestimated within the model. For

▶

Fig. 5 As Fig. 2 except for SURF input temperature diagnostics


Traces for hadcm2 surf temperatures
10-areas
5-areas
12-areas
6-areas
GS
VOL
G
S
G
S
VOL
Amplitude
Truncation
3
2
1
0
-1
-2
0
7
14
21


Traces for hadcm3 surf temperatures
10-areas
5-areas
12-areas
6-areas
GS
VOL
G
S
G
S
VOL
Amplitude
Truncation
3
2
1
0
-1
-2
0
7
14
21

◀

Fig. 6 As Fig. 3 except for SURF input temperature diagnostics

HadCM3, the damping required on the GS signal estimate to fit the observations is even greater than for HadCM2. This is in keeping with previous analyses of near-surface temperature, which show that HadCM3 is more likely to overestimate the magnitude of the late 20th Century anthropogenic response (see Figs. 8 and 11 of T02, and T99). However, the results detailed here exhibit a greater overestimation bias than in T02. The systematic effect on results for both models indicates that it is most likely to relate to changes in input fields and their pre-processing between our study and previous studies (see Sect. 4.4 and Gillett et al. 2002). For HadCM3, in agreement with HadCM2, a detectable VOL signal is discounted, being at or around zero scaling in nearly all cases. G is only rarely detected in combination with S for HadCM3, and the simuluted signal almost always needs to be significantly reduced. The simulated S signal is generally consistent with observed near-surface temperatures, but only ever marginally detected in *smart* LAA diagnostics, and never detected in *na* LAA diagnostics. The greater G signal strength overestimation within HadCM3 according to our analysis leads to increased ambiguity regarding anthropogenic signal detection compared to previous analyses (T02). Results for the individual HadCM3 signals discussed above are generally insensitive to the consideration of a three-way regression.

Global mean SURF temperature reconstructions are shown in Fig. 7. For both models those reconstructions including VOL are physically implausible, since amplitude scalings applied to this forcing are negative at the truncation considered (but we retain them for completeness). We therefore concentrate on the the HadCM2 and HadCM3 G + S reconstructions. Both are within the 2σ uncertainty bounds of observed SURF temperatures, with G providing the major component of the warming trend, in agreement with results for LT. However, the absolute magnitude of the trend over the entire period is not well captured, being underestimated by both model reconstructions. This is likely due to the large down-weighting required on the G signal in both models according to our analysis. Previous analyses which yielded scaling estimates closer to unity better recreate the magnitude of the observed trends (T02; S01).

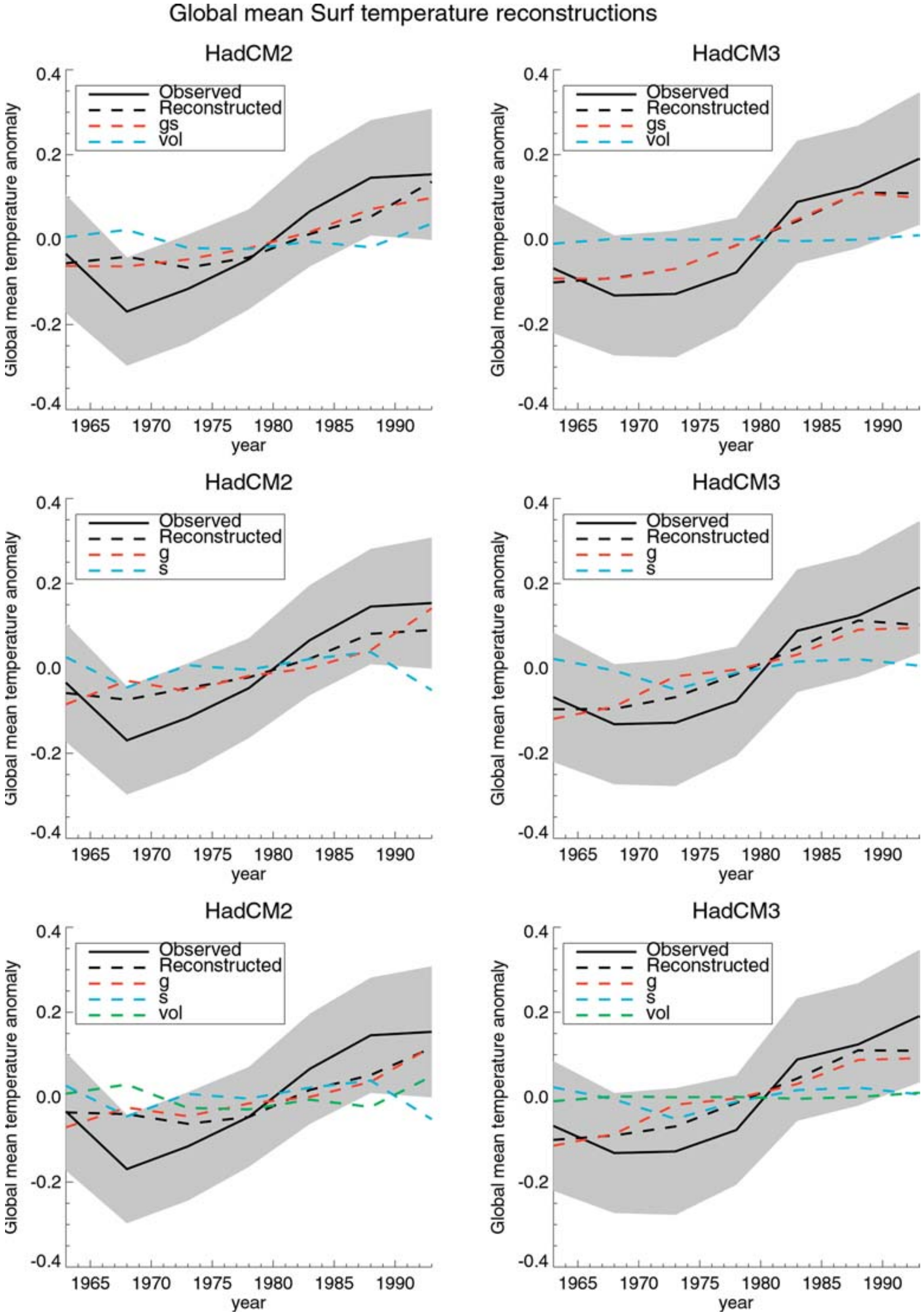
4.4 Results for all temperature diagnostics

In this section we summarise results for all six input tropospheric temperature diagnostics. The summary is based upon the same analysis as that performed in the previous two subsections. This detailed analysis is available online in Thorne (2001) for the remaining four temperature diagnostics (UT and the three "lapse-rate" diagnostics). We distill the large volume of results from each temperature diagnostic to provide a few meaningful indicators. Inevitably a degree of subjectivity is involved in this process, even though the raw results upon which these summaries are based are quantitative and consider uncertainty due to pre-processing choices. We conclude that we are very confident if results agree across all LAA representations and at almost all truncations in supporting a particular conclusion. We report no or very little confidence if the results overwhelmingly either do not support or contradict the conclusion. In between these extremes our results provide some degree of support but a degree of uncertainty remains based upon the choices of truncation and LAA representation. This degree of uncertainty is reflected in the confidence interval we assign. We stress that this subjectivity will only impact the absolute confidence values assigned to any statement rather than their relative values. We choose to concentrate on what we consider to be the four most useful indicators: whether a signal is detected, whether it is a consistent explanation of the observations (our attribution approach), whether the consistency test on the residuals is passed, and how well global mean trends are reproduced. Results for HadCM2 are shown in Table 2, whilst those for HadCM3 are shown in Table 3.

Both models agree that successful signal detection is more likely for both UT and LT than SURF and, in particular, than lapse rate diagnostics. Lapse rate diagnostics systematically have lower SNRs as much of the signal in the models is common to adjacent layers and hence removed in our lapse rate calculation. Therefore they may have a tendency to yield significantly negatively biased scalings (hence less detections). However our analysis (Thorne 2001, see also Sect. 5 and Fig. 8) shows that this systematic reduction in signal detectability for lapse rates arises predominantly from a systematic increase in the uncertainty range in our scaling estimates rather than negative biases in the scaling estimates themselves. This most likely relates to the optimisation tending to be less efficient for the "noisier" lapse rate diagnostics leading to larger uncertainties. Our analysis for all diagnostics also confirms our preliminary detection analyses (see Sect. 4.1) in ranking the signal detectability as, in descending order; G (or GS), S, VOL for both models.

For HadCM2 we find with high confidence that anthropogenic forcing factors (GS, G + S) are consistent with the observations for all diagnostics except SURF and LT-SURF, for which the model significantly overestimates the amplitude of the response to GS(G). For HadCM3 the overestimation of the response to anthropogenic forcings, in particular GS(G), is more widespread than for HadCM2 (consistent with T99 and T02). However, in agreement with the HadCM2 analyses this model overestimation of the GS(G) response is greatest for SURF and LT-SURF. There is also a tendency for both models to overestimate the amplitude of the volcanic forcing response. This is contrary to recent findings of Jones et al. (2003) who, by using a 4D input field including stratospheric values and making na assumptions about observational coverage, yield a VOL

Fig. 7 As Fig. 4 except for SURF temperature diagnostics



signal consistent with observations. We do not include stratospheric values and we downweight tropical regions, the two regions where a VOL response is likely to be strongest, which taken together likely explain this difference.

The systematic difference between SURF and LT-SURF and the remaining input diagnostics for the GS and G scaling estimates suggests that either the models or observations (or both) could be in error in the surface regions considered. Our results imply that the surface should have been warming faster than observed over 1960–1994 whereas the troposphere has been warming at the predicted rate if the models are correct. This is the opposite of the recently observed and much discussed global-mean trend discrepancy between observed and modelled lower tropospheric and near-surface temperatures (NRC 2000). We repeat our previous caveat that we are considering both a highly sub-sampled space and a longer time period than considered in the NRC report. We also note that our analysis points towards potential

Table 2 Summary of principal results for HadCM2

1 No to very low confidence
2 Low confidence
3 Medium confidence
4 High confidence
5 Very high confidence
> Model signal response significantly overestimated
eq. Model signal response consistent with the observations
< Model signal response significantly underestimated

	Detection				Consistent explanation of the observations												Consistent residuals	Reproduces Global mean trends
	GS	G	S	VOL	GS			G			S			VOL				
					>	eq.	<	>	eq.	<	>	eq.	<	>	eq.	<		
UT	5	5	3	2	1	5	1	1	5	1	1	4	1	4	1	1	4	5
LT	5	5	2	2	1	5	1	1	5	1	1	4	1	3	2	1	5	3
SURF	4	4	2	1	3	3	1	3	3	1	3	3	1	1	1	1	5	3
UT-LT	3	3	2	3	1	5	1	2	4	1	1	4	1	4	2	1	3	4
UT-SURF	3	2	2	1	2	4	1	2	4	1	2	4	1	1	2	1	4	3
LT-SURF	2	2	2	1	2	1	1	2	1	1	1	2	1	1	1	1	4	4

Table 3 Summary of principal results for HadCM3

1 No to very low confidence
2 Low confidence
3 Medium confidence
4 High confidence
5 Very high confidence
> Model signal response significantly overestimated
eq. Model signal response consistent with the observations
< Model signal response significantly underestimated

	Detection				Consistent explanation of the observations												Consistent residuals	Reproduces Global mean trends
	GS	G	S	VOL	GS			G			S			VOL				
					>	eq.	<	>	eq.	<	>	eq.	<	>	eq.	<		
UT	5	5	3	3	3	3	1	4	2	1	1	4	1	3	3	1	2	4
LT	5	5	3	1	1	5	1	2	4	1	1	4	1	4	2	1	4	5
SURF	4	4	2	1	4	1	1	5	1	1	2	4	1	1	1	1	5	3
UT-LT	2	2	1	1	3	2	1	3	2	1	1	1	1	1	1	1	2	1
UT-SURF	1	1	1	1	3	2	1	3	2	1	1	1	1	1	1	1	5	3
LT-SURF	2	2	2	2	3	1	1	3	1	1	2	2	1	2	2	1	5	4

biases in our SURF results when compared both to previous studies (T99, T02; S01) as well as to our results for remaining diagnostics. These earlier studies better reconstructed both the magnitude and evolution of the global-mean observed trends and yielded signal estimates consistent with unity. Our analysis does not permit definitive attribution of the causes of the differences in our results compared to previous studies. This would require a much more detailed experiment to assess our sensitivity to a range of plausible pre-processing choices, our current small sample space being grossly insufficient. Analysis in Sect. 5 suggests that the differences in results solely within our analysis between SURF and tropospheric temperature diagnostics may not be statistically significant.

For HadCM2, the consistency test on the residuals passes in nearly all cases for all input temperature diagnostics. The exception is for UT-LT lapse rates and this is mitigated if HadRT2.1 is used instead of HadRT2.1s, and hence is likely to be a result of observational rather than model error. For HadCM3 the consistency test also passes in most cases, the exceptions being UT and UT-LT. The latter can again be explained by observational error, whereas the former cannot. Given that HadCM2 does not experience similar behaviour for UT, the increased rate of consistency test failure is likely to relate at least in part to model error within HadCM3. HadCM3 probably fails to adequately capture at least some of the leading modes of internal climate variability in UT.

Finally we consider how well our estimators reconstruct observed global-mean trends. HadCM2 tends to underestimate the magnitude of the observed global-mean temperature trends in most cases using the weightings derived from our analyses. The reconstructions also occasionally fall outside the 2σ bounds of the observations, although this could happen due to chance alone. HadCM3 also tends to underestimate the magnitude of the observed trends, but generally stays within 2σ of the observed reconstruction. The exception is for UT, where at the truncations considered at least one of the VOL or S forcings involved always has a negative (unrealistic) weighting applied and therefore we cannot make any meaningful inferences. Much of the reason for the systematic tendency for both models to under-estimate the magnitude of the observed global-mean changes in at least those temperature diagnostics including SURF most likely relates to biases introduced by the reduced spatial sampling, leading to differences from previous studies, as discussed previously.

5 Are inter-level results consistent?

Analysis in Sect. 4 considered results for individual tropospheric temperature diagnostics as independent pieces of information. In both the modelled and real worlds we know that this is not true, and that on the large space and long time scales which we are considering the results should be consistent. As a first step towards an assessment of consistency we undertake a very simple analysis. All we wish to do is be able to identify whether evidence exists for fundamental discrepancies (arising from the modelled or the observed

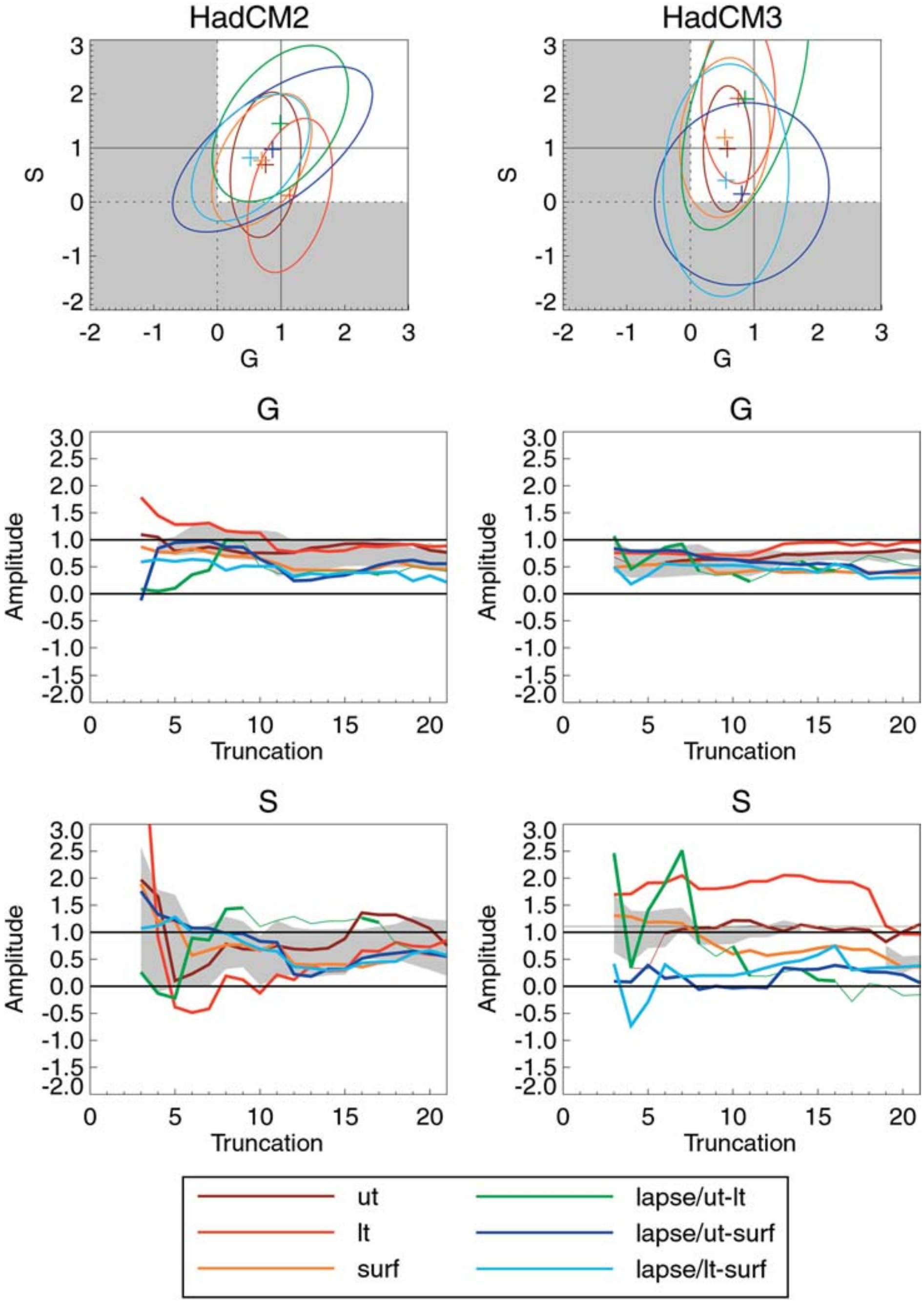


Fig. 8 The range of solutions gained for a G + S signal combination for all six input temperature fields and both HadCM2 (*left panels*) and HadCM3 (*right panels*). The *top images* give ellipses containing 90% of the joint probability distribution with the amplitude scaling estimate resulting from the regression denoted by a *cross* at truncation 9 (HadCM2) and 6 (HadCM3). The *lower panels* give traces over the entire range of truncations for both G and S. Areas denoted by *shading* in these *lower plots* relate to regions where the univariate uncertainty limits (at the 90% level) overlap entirely. The individual limits are not shown here for the sake of clarity. Where the test on the residuals fails the traces are shown as *thin lines*

datasets) in the results for either model. We proceed under the assumption that such gross errors would cause a lack of overlap in the uncertainty ranges of our individual signal scaling estimators for our six tropospheric temperature diagnostics. We make no effort to ascertain quantitative measures of the degree of agreement, although this would be highly desirable in future. We also limit our analysis to the G + S input signal combination.

Results for both HadCM2 and HadCM3 are shown in Fig. 8 for the 10-area *smart* LAA diagnostic. It can be seen from the ellipses in the top panels of this figure that for both models there exists a degree of overlap in the scaling estimates at the truncation being considered (9 and 6 for HadCM2 and HadCM3 respectively, the maximum truncation for which all diagnostics pass the consistency test on the residuals). In the bottom panels are univariate detection traces. These show that this overlap is independent of truncation over a reasonable range for both G and S signal estimators in HadCM2. At the majority of truncations both these signals are consistent with the observations, but they have a tendency to overestimate the signal strength (have a scaling estimate < 1) for most of the temperature diagnostics considered. For HadCM3, although the G signal estimates consistently overlap, the S signal estimates are more uncertain and whether there is overlap is critically dependent upon the chosen truncation. HadCM3 consistently overestimates the amplitude of the observed response to greenhouse gases, whereas the overlap, when it exists, is generally consistent with the observations for the S signal response. Individual S amplitude scaling estimates for HadCM3 are highly uncertain however, reducing our confidence in this result. The fact that the overlap is relatively robust to changing truncation for both models, at least for G, which our analyses have shown to be the most important explanation of the observed changes, does provide some confidence in their adequacy.

6 Conclusions

We have shown that applying an optimal detection algorithm to full spatio-temporal tropospheric temperature fields potentially adds power to zonal mean detection studies. Available near-surface and radiosonde observations were compared to climate change experiments performed with two versions of the Hadley Centre model: HadCM2 and HadCM3. Our analysis serves to confirm results of previous studies of tropospheric temperature changes over the latter twentieth century (e.g. AT99; T99, T02; S01) in that we detect with high confidence an anthropogenic signal for our analysis period of 1960–1994. We also sometimes detect volcanic influences, although we made no attempt to remove ENSO effects from the observed fields which might bias this result towards non-detection. We find that HadCM2 and particularly HadCM3 based simulations have a tendency to significantly overestimate the amplitude of the tropospheric temperature response to anthropogenic greenhouse gases. We conclude that this effect is likely to be real, although comparisons with previous studies suggest that artificially reduced dataset coverage and a shift in emphasis towards extra-tropical land regions to match available radiosonde coverage, amongst other differences, may be playing a significant role at the surface where the discrepancy is most pronounced. Therefore, at this stage we cannot unambiguously attribute recently observed changes in tropospheric temperatures to any combination of external forcing influences, although by far the most plausible causes are anthropogenic.

We have considered the sensitivity of results to several potential sources of uncertainty and find our principal conclusions to be robust, although further sources of uncertainty not considered within our analysis certainly remain. In particular it would be desirable to repeat these analyses using different modelled and observed datasets and to consider the reasons for our discrepancies with previous surface analyses in a robust and systematic manner. We aim to achieve this latter point through an analysis of the sensitivity of detection results to input field pre-processing choices. A simple comparison of this study and previous studies is insufficient to make conclusive statements as to the causes of the differences in results. Our analysis showed that corrections previously applied to the HadRT radiosonde temperature record might be sub-optimal in considering solely spatial rather than spatio-temporal consistency aspects. Work in progress aims to rectify this. We caution that any other radiosonde record corrections based on solely single station records in isolation may suffer from similar problems. We plan to repeat our detection analyses once improved radiosonde data are available using updated HadCM3 model fields for the longer period of 1958 to date to assess sensitivity of our results to radiosonde data uncertainties.

We extend traditional detection and attribution studies by stipulating that for multiple climate variables our detection results should overlap for any given model. We conclude under this approach that there is no evidence for a fundamental discrepancy between modelled and observed temperatures within the troposphere, at least under anthropogenic forcings. In the longer term we would envisage a more rigorous, quantitative, check being possible (see Ch. 7 of Thorne 2001 for a detailed discussion).

Acknowledgements This work was carried out while PWT was in receipt of NERC studentship GT04/98/76/AS. PDJ acknowledges support of the Office of Science (BER), U.S. Dept. of Energy, Grant No. DE-FG02-98ER62601. The Met Office authors are supported by the U.K. Department of the Environment, Food and the Rural Affairs under contract PECD7/12/37 and the Government Meteteorological Research program under Contract MSG-2/00. Through their contribution, the paper is Crown Copyright. HadRT, HadCM2 and HadCM3 data provided by the Hadley Centre. HadCM2 and HadCM3 data are available through the Climate Impacts LINK Project. HadCRUT2v data (an updated version of HadCRUTv) are available from the Climatic Research Unit website http://www.cru.uea.ac.uk. We thank Ben Santer and Gabi Hegerl for very useful reviews, which substantially improved the quality of this study.

References

Allen MR, Tett SFB (1999) Checking for model consistency in optimal fingerprinting. Clim Dyn 15: 419–434

Allen MR, Stott PA, Mitchell JFB, Schnur R, Delworth TL (2000) Quantifying the uncertainty in forecasts of anthropogenic climate change. Nature 407: 617–620

Allen MR, Stott PA (2003) Estimating signal amplitudes in optimal fingerprinting, part I: theory. DOI 10.1007/s00382-003-0319-9

Allen MR, Gillett NP, Kettleborough JA, Hegerl G, Schnur R, Stott PA, Boer G, Covey C, Delworth TL, Jones GS, Mitchell JFB, Barnett TP (2002) Quantifying anthropogenic influence on recent near-surface temperature change. Accepted by Surveys in Geophysics

Angell JK (1988) Variations and trends in tropospheric and stratospheric global temperatures, 1958–87. J Clim 1: 1296–1313

Angell JK (2000) Difference in radiosonde temperature trend for the period 1979–1998 of MSU data and the period 1959–1998 twice as long. Geophys Res Lett 27: 2177–2180

Barnett TP, Hegerl GC, Santer BD, Taylor K (1998) The potential effect of GCM uncertainties and internal atmospheric variability on anthropogenic signal detection. J Clim 11: 659–675

Barnett TP, Hasselmann K, Chelliah M, Delworth T, Hegerl G, Jones P, Rasmusson E, Roeckner E, Ropelewski C, Santer B, Tett S (1999) Detection and attribution of recent climate change: a status report. Bull Am Meteorol Soc 80: 2631–2659

Barnett TP, Pierce DW, Schnur R (2001) Detection of anthropogenic climate change in the world's oceans. Science 292: 270–274

Brown SJ, Parker DE, Folland CK, Macadam I (2000) Decadal variability in the lower-tropospheric lapse rate. Geophys Res Lett 27: 997–1000

Christy JR, Spencer RW, Lobl ES (1998) Analysis of the merging procedure for the MSU daily temperature time series. J Clim 11: 2016–2041

Christy JR, Spencer RW, Braswell WD (2000) MSU tropospheric temperatures: dataset construction and radiosonde comparisons. J Atmos Oceanic Tech 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. J Atmos Oceanic Tech 20: 613–629

Collins M, Tett SFB, Cooper C (2001) The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 17: 61–81

Eskridge RE, Alduchov OA, Chernykh IV, Panmao Z, Polansky AC, Doty SR (1995) A comprehensive aerological reference data set (CARDS) – rough and systematic errors. Bull Am Metearol Soc 76: 1759–1775

Gaffen DJ (1996) A digitized metadata set of global upper-air station histories. NOAA Techn Memorand ERL ARL-211

Gaffen DJ, Santer BD, Boyle JS, Christy JR, Graham NE, Ross RJ (2000) Multidecadal changes in the vertical temperature structure of the tropical troposphere. Science 287: 1242–1245

Gibson JK, Kållberg P, Uppala S, Hernandez A, Nomura A, Serrano E (1997) ERA description. ECMWF Report ERA PRS1

Gillett NP, Allen MR, Tett SFB (2000) Modelled and observed variability in atmospheric vertical temperature structure. Clim Dyn 16: 49–61

Gillett NP, Hegerl GC, Allen MR, Stott PA, Schnur R (2002) Reconciling two approaches to the detection of anthropogenic influence on climate. J Clim 15: 326–329

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

Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, Johns TC, Mitchell JFB, Wood RA (2000) The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16: 147–168

Hansen J, Ruedy R, Glascoe J, Sato M (1999) GISS analysis of surface temperature change. J Geophys Res 104: 30,997–31,022

Hansen J, Ruedy R, Sato M, Imhoff M, Lawrence D, Easterling D, Peterson T, Karl T (2001) A closer look at United States and global surface temperature change. J Geophys Res 106: 23,947–23,963

Hasselmann K (1993) Optimal fingerprints for the detection of time-dependent climate change. J Clim 6: 1957–1971

Hegerl GC, Wallace JM (2002) Influence of patterns of climate variability on the difference between satellite and surface temperature trends. J Clim 15: 2412–2428

Hegerl GC, Von Storch H, Hasselmann K, Santer BD, Cubasch U, Jones PD (1996) Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. J Clim 9: 2281–2310

Hegerl GC, Hasselmann K, Cubasch U, Mitchell JFB, Roeckner E, Voss R, Waskewitz J (1997) Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change. Clim Dyn 13: 613–634

Hill DC, Allen MR, Stott PA (2001) Allowing for solar forcing in the detection of human influence on tropospheric temperatures. Geophys Res Lett 28: 1555–1558

Hurrell JW, Trenberth KE (1998) Difficulties in obtaining reliable temperature trends: reconciling the surface and satellite microwave sounding unit records. J Clim 11: 945–967

Johns TC, Carnell RE, Crossley JF, Gregory JM, Mitchell JFB, Senior CA, Tett SFB, Wood RA (1997) The second Hadley Centre coupled ocean atmosphere GCM: model description, spinup and validation. Clim Dyn 13: 103–134

Jones GS, Tett SFB, Stott PA (2003) Causes of atmospheric temperature change 1960–2000: A combined attribution analysis. GRL 30: Art. No. 1228

Jones PD, Osborn TJ, Wigley TML, Kelly PM, Santer BD (1997a) Comparisons between the microwave sounding unit temperature record and the surface temperature record from 1979 to 1996: real differences or potential discontinuities? J Geophys Res 102: 30,135–30,145

Jones PD, Osborn TJ, Briffa KR (1997b) Estimating sampling errors in large-scale temperature averages. J Clim 10: 2548–2568

Jones PD, New M, Parker DE, Martin S, Rigor IG (1999) Surface air temperature and its changes over the past 150 years. Rev Geophys 37: 173–199

Jones PD, Osborn TJ, Briffa KR, Folland CK, Horton EB, Alexander LV, Parker DE, Rayner NA (2001) Adjusting for sampling density in grid box land and ocean surface temperature time series. J Geophys Res 106: 3371–3380

Kalnay E et al (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77: 437–471

Lanzante JR, Klein SA, Seidel DJ (2003a) Temporal homogenization of monthly radiosonde temperature data: Part 1: methodology. J Clim 16: 224–240

Lanzante JR, Klein SA, Seidel DJ (2003b) Temporal homogenization of monthly radiosonde temperature data: Part 2: trends, sensitivities and MSU comparison. J Clim 16: 241–262

Lean J, Beer J, Bradley R (1995) Reconstruction of solar irradiance since 1610 – Implications for climate-change. Geophys Res Lett 22: 3195–3198

Levine RA, Berliner LM (1999) Statistical principles for climate change studies. J Clim 12: 564–574

Levitus S, Antonov JI, Wang J, Delworth TL, Dixon KW, Broccoli AJ (2001) Anthropogenic warming of earth's climate system. Science 292: 267–270

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic Press

NRC (National Research Council) (2000) Reconciling observations of global temperature change. National Academy Press, Washington, D.C., pp 85

North GR, Kim KY, Shen SSP, Hardin JW (1995) Detection of forced climate signals. 1. Filter theory. J Clim 8: 401–408

Parker DE, Gordon M, Cullum DPN, Sexton DMH, Folland CK, Rayner N (1997) A new global gridded radiosonde temperature data base and recent temperature trends. Geophys Res Lett 24: 1499–1502

Pawson S, Forino M (1998) A comparison of reanalyses in the tropical stratosphere, 1. Thermal structure and the annual cycle. Clim Dyn 14: 631–644

Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3. Clim Dyn 16: 123–146

Santer BD, Taylor KE, Wigley TML, Penner JE Jones PD, Cubasch U (1995) Towards the detection and attribution of an anthropogenic effect on climate. Clim Dyn 12: 77–100

Santer BD, Taylor KE, Wigley TML, Johns TC, Jones PD, Karoly DJ, Mitchell JFB, Oort AH, Penner JE, Ramaswamy V, Schwarzkopf MD, Stouffer RJ, Tett S (1996) A search for human influences on the thermal structure of the atmosphere. Nature 382: 39–46

Santer BD, Hnilo JJ, Boyle JS, Doutriaux C, Fiorino M, Parker DE, Taylor KE, Wigley TML (1999) Uncertainties in observationally-based estimates of temperature change in the free atmosphere. J Geophys Res 104: 6305–6333

Santer BD, Wigley TML, Gaffen DG, Doutraiaux C, Boyle JS, Esch M, Hnilo JJ, Jones PD, Meehl GA, Roeckner E, Taylor KE, Wehner M (2000) Interpreting differential temperature trends at the surface and in the lower troposphere. Science 287: 1227–1232

Santer BD, Wigley TML, Doutriaux C, Boyle JS, Hansen JE, Jones PD, Meehl GA, Roeckner E, Sengupta S, Taylor KE (2001) Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends. J Goephys Res 106: 28,033–28,059

Santer BD, Sausen R, Wigley TML, Boyle JS, AchutaRao K, Meehl GA, Roeckner E, Taylor KE (2003a) Comparison of the signal to noise properties of changes in tropopause height and atmospheric temperature. J Geophys Res 108: art no 4002

Santer BD, Wigley TML, Meehl GA, Wehner MF, Mears C, Schabel M, Wentz FJ, Ammann C, Arblaster J, Bettge T, Washington WM, Taylor KE, Boyle JS, Bruggemann W, Doutriaux C (2003b) Influence of satellite data uncertainties on the detection of externally forced climate change. Science doi:10.1126/science.1082393

Sato M, Hansen JE, McCormick MP, Pollack JB (1993) Stratospheric aerosol optical depths (1850–1990). J Geophys Res 98: 22,987–22,994

Stott PA, Tett SFB (1998) Scale-dependent detection of climate change. J Clim 11: 3282–3294

Stott PA, Kettleborough JA (2002) Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. Nature 416: 723–726

Stott PA, Tett SFB, Jones GS, Allen MR, Ingram WJ, Mitchell JFB (2001) Attribution of twentieth century temperature change to natural and anthropogenic causes. Clim Dyn 17: 1–17

Tett SFB, Stott PA, Allen MR, Ingram WJ, Mitchell JFB (1999) Causes of twentieth-century change near the Earth's surface. Nature 399: 569–572

Tett SFB, Jones GS, Stott PA, Hill DC, Mitchell JFB, Allen MR, Ingram WJ, Johns TC, Johnson CE, Jones A, Roberts DL, Sexton DMH, Woodage MJ (2002a) Estimation of natural and anthropogenic contributions to 20th Century temperature change. J Geophys Res 107: doi:10.1029/2000JD000028

Thorne PW (2001) Advancing climate change detection and attribution studies in the free atmosphere. PhD thesis, University of East Anglia, Norwich, NR4 7TJ, UK, pp 288 http://www.cru.uea.ac.uk/cru/pubs/thesis/PeterThorne2001/

Thorne PW, Jones PD, Osborn TJ, Davies TD, Tett SFB, Parker DE, Stott PA, Jones GS, Allen MR (2002a) Assessing the robustness of zonal mean climate change detection studies. Geophys Res Lett 29 art. no. 1920

Thorne PW, Jones PD, Tett SFB, Parker DE, Osborn TJ, Davies TD (2002b) Ascribing potential causes of recent trends in free atmosphere temperatures. ASL, doi:10.1006/asle.2001.0046

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