Atmospheric Science Letters (2002)
doi:10.1006/asle.2001.0046

Ascribing potential causes of recent trends in free atmosphere temperatures

P. W. Thorne[1*], P. D. Jones[1], S. F. B. Tett[2], D. E. Parker[2], T. J. Osborn[1] and T. D. Davies[1]

[1]*Climatic Research Unit, University of East Anglia, Norwich, NR4 7TJ, U.K.*
[2]*Hadley Centre for Climate Prediction & Research, Met Office, London Rd, Bracknell, Berkshire, RG12 2SY, U.K.*

Abstract: We use globally gridded radiosonde temperature datasets in a simple climate change study. Two climate models, when run with historical and, particularly, anthropogenic forcings, exhibit a degree of agreement with radiosonde temperature observations for 1958–1998.

Keywords: Climate change, detection, radiosondes, climate models.

1. Introduction

Previous formal, quantitative, climate change detection and attribution studies have considered changes in near-surface temperatures (e.g. Santer et al., 1995, Tett et al., 1999, Stott et al., 2001), ocean heat content (e.g. Barnett et al., 2001), or zonally-averaged vertical temperature structure (e.g. Santer et al., 1996, Tett et al., 1996, Allen and Tett, 1999). Such studies suggest a detectable anthropogenic influence on the climate during the second half of the 20th century, the results being robust to changes in procedure and use of different climate models. However, this dependence of detection and attribution studies on such limited data is undesirable; if a signal can be detected in more parameters or fuller fields then our confidence will be increased. Here we use global three-dimensional radiosonde temperature datasets to begin to discriminate between potential causes of recent climate change. Statements of formal detection and attribution are addressed in subsequent papers (in preparation). In this paper, we employ very simple statistical measures and qualitative field comparisons, to illustrate that a degree of model skill is still evident without recourse to more complex detection and attribution methodologies (Allen and Tett, 1999, for example).

We use versions 2.1 and 2.1s of the gridded HadRT radiosonde temperature record (Parker et al., 1997). Both versions have been corrected for known heterogeneities post-1979. Corrections were applied with reference to collocated data from the satellite microwave sounding unit (MSUc) record of Christy et al. (1998), MSUd being unavailable at the time the corrections were applied. Corrections have been applied globally following Parker et al. (1997), although in HadRT2.1s they are applied solely in the stratosphere. Further editing of the datasets has been undertaken following near-neighbour consistency checks, resulting in the omission of approximately ten percent of

* Peter.Thorne@uea.ac.uk

1530-261X

the available grid boxes. We consider the MSUc record to be too short for use in detection and attribution studies at present. Even the HadRT record is barely adequate, but it is the longest upper air gridded dataset available. The longer NCEP/NCAR/ECMWF reanalyses have homogeneity problems, partly due to changing input data, which preclude them from signal detection studies (Barnett et al., 1999).

In this study we consider model output from two versions of the Hadley Centre's fully coupled ocean–atmosphere general circulation model (OAGCM), HadCM2 (Johns et al., 1997), and HadCM3 (Pope et al., 2000). Both models have been run in control mode with no changes in external forcings and are stable, although HadCM2 requires flux corrections to avoid drifting into an unrealistic state. In addition, both HadCM2 and HadCM3 have been used to simulate ensemble realisations (generally of 4 members) of climate change under a number of potential forcing mechanisms. In the current study we use results from these forced runs to perform a comparison with the HadRT data. We maintain the forcing acronyms used in previous studies (e.g. Tett et al., 1999, 2001, Stott et al., 2000, 2001) for consistency. Although the forcings applied to HadCM2 and HadCM3 were not identical, there are three anthropogenic ensembles which are broadly comparable between models. These are for HadCM2 (HadCM3) in increasing order of complexity: G (GHG) (increases in anthropogenic greenhouse gases only), GS (TROP-ANTHRO) (which additionally incorporates the effects of sulphate aerosols), and GSO (ANTHRO) (which also includes stratospheric ozone depletion). There are differences in both the greenhouse gas and sulphate aerosol forcings between the model generations. In HadCM2 the greenhouse gases are considered as CO_2 equivalent concentrations, whereas HadCM3 considers greenhouse gases individually and incorporates an interactive chemistry model. Further, HadCM2 considers solely the direct scattering effects of sulphate aerosols and neglects the indirect effect due to changes in cloud droplet concentration and size properties, while HadCM3 incorporates both effects. We also refer to integrations based upon the estimated history of natural forcings. For HadCM2 there are two solar based forcings, SOL (Hoyt and Schatten, 1993, Willson, 1997), and LBB (Lean et al., 1995), and a volcanic based forcing, VOL (Sato et al., 1993). HadCM3 has a forcing labelled NATURAL which combines the LBB and VOL forcings to estimate the sum effect of changes in natural external forcings. See Tett et al. (1999, 2001), and Stott et al. (2000, 2001), for more details of all the forcings, and the differences between the models.

In section 2 we describe our methodology for creating simple statistical measures of similarity between fields of data, and outline the likely principal limitations of such methods. In section 3 we analyse changes in temperature on individual pressure levels, and changes in tropospheric lapse rates, to evaluate overall coupled-model skill. In section 4 we summarise our findings.

2. Methodology

Model fields are spatially complete and have a grid resolution of 2.5° latitude × 3.75° longitude, whereas the observations are incomplete and available on a 5° × 10° grid. So first we re-grid the model data onto the observational grid. Next we only use the HadRT data, which are monthly anomalies (from the 1971–1990 mean) for the period 1958–1998, when an annual observed value for a gridbox has data for two or more months, in three or more seasons. Model simulated annual-mean data are then discarded where there are no observed data: the remaining values being used to create a reference climate for 1971–1990 for each member of each ensemble. Subsequently, the model data are reduced to anomalies relative to their respective climatologies. By masking before creating anomalies from each model ensemble member, we account for the effects of temporal sampling variability in subsequent analysis. All members of each ensemble are then used to form an ensemble average response in an attempt to maximise the signal strength.

To reduce noise further we consider decadally-averaged data. For any grid box to have a decadal mean calculated, it must contain at least five years of data with no gap exceeding two years. We focus on fields of differences between the last decade and the first decade. This

further enhances the signal strength, although any information on the transient nature of any climate change will be lost. Also, it reduces the chances of finding similarities solely as an artefact of using anomalies from the common period 1971–1990 in the respective data series, which is included in both cases considered here. During this period values in all series will be approximately zero, so a spuriously high similarity could occur solely due to the normalisation procedure. The observational data exist for 40 years, but have poor coverage both early and late in the record. Therefore, the choice of decades to compare is a difficult one, being a trade-off between signal strength, which is greater for longer trends, and data coverage, which degrades for the longest trends. So, we limit our analysis to a four decade diagnostic, 1988–1997 minus 1958–1967, and the three decade diagnostic which exhibits maximum data coverage, 1983–1992 minus 1963–1972. We begin by making a very simple qualitative intercomparison of the observed and best guess anthropogenic (GSO/ANTHRO) fields at the 500 hPa level, and for a deep layer lapse rate (300 minus 850 hPa) diagnostic for the three decade case. This provides insights into the dataset coverage, and the spatial coherency of the observed and modelled fields.

To enable quantitative comparisons to be made between fields, the model intra-ensemble variability is used to estimate the standard deviation of the decadal means at each grid box. Any intra-ensemble variability is considered to be due to climate noise. This will be model dependent, so we derive separate estimates of the standard deviation for HadCM2 and HadCM3, yielding at least 24 independent realisations in each case (at least three per ensemble). We are assuming that the standard deviation does not change in this analysis under transient forcings. We take two approaches to quantifying the agreement between observed and model decadal difference fields. These approaches assume that both the pattern and the amplitude of the model fields are correct. Firstly, on a grid-box basis, each ensemble average is compared to the observations. If the ensemble average does not agree within three sigma with the observational field then it is flagged. Effectively we are proposing that the ensemble mean response is not significantly different from the observations at the 1 percent confidence limit at the grid-box level. Our expectations are that this will fail in more than 1 percent of all cases even if the model perfectly captures the climate change, because Stott and Tett (1998) have shown that HadCM2 underestimates variability at scales below 2000 km, and we expect HadCM3 to behave similarly. Using such an approach will still yield useful evidence as to modelled-observed agreements. Secondly, in an attempt to gain information on overall field similarity we use a Root Mean Squared Difference (RMSD) statistic. The use of a RMSD statistic does not account for spatial correlation within the decadal difference fields and will, therefore, yield artificially significant estimates (Wigley et al., 2000). Here, however, for a given model, we are comparing the performance of simulations with the same spatial correlation structure and, therefore, the bias is likely to be similar in all cases. If forcing were to generate a new correlation structure, then this would no longer hold true. For both approaches we also calculate the values for a ''null hypothesis of no change'' whereby we set all the ''ensemble mean'' grid-box anomalies to zero. This null hypothesis field is perfectly auto-correlated and, therefore, is a conservative approach for the RMSD statistic.

We consider all nine standard levels (30 hPa, 50 hPa, 100 hPa, 150 hPa, 200 hPa, 300 hPa, 500 hPa, 700 hPa and 850 hPa) and lower (500 minus 850 hPa), and upper (300 minus 500 hPa), as well as deep tropospheric lapse rates. If a model exhibits skill in predicting the observed changes, then it should exhibit skill over a range of spatial scales. Therefore, if any model ensemble average exhibits both a lower percentage of individually anomalous grid boxes, and a smaller RMSD, than the ''null hypothesis of no change'' field, then we infer that it has a degree of skill relative to the null hypothesis in explaining the observations for that field. The ensemble score is incremented by 1 for every level or lapse rate which meets these criteria, yielding a maximum potential score of 12. There are obvious potential pitfalls in such a binary approach to a quantitative assessment of model skill, in that it does not provide any quantification as to how much better or worse the model field is than our ''no change'' field at any level. However, we feel both that it is complimentary to results from more formal detection approaches, and that the use of a model control segment to assess the range of plausible scores goes some way to communicating

the level of agreement of each ensemble mean realisation relative to our null hypothesis "no change" field. We employ a 1200-year section of HadCM3 control run data to generate 36 three decade and 27 four decade non-overlapping realisations of the interdecadal climate noise. We then add these realisations to each ensemble interdecadal difference field for the appropriate interval, scaling to account for ensemble size effects, and calculate the score value for each of these noise-polluted realisations, yielding a range of possible scores. If this range does not encompass zero then we conclude that the signal is likely to have a demonstrable manifestation in at least some of the observations, confidence increasing with the degree of separation. We also assess the likely range of scores due to chance alone, by adding the same synthetic realisations of noise to the "no change" field and assessing scores in an identical manner. If the best guess score for any signal is separated from this "no change" uncertainty range, then our confidence in its likely detectability is enhanced. We do not attach rigorous estimates of statistical significance to the results as the true degrees of freedom of the control segment are unknown, and our approach is deliberately simplistic, so no claims are made of unambiguous detection or attribution. Results using this approach are also going to be critically dependent upon the adequacy of the HadCM3 control section, and there is some evidence, at least within the stratosphere, that the variability may be significantly underestimated (Collins et al., 2001).

2.1 Comparison of modelled and observed fields

Figure 1 (top panels) shows the difference in observed decadally-averaged temperatures between 1963–1972 and 1983–1992 at the 500 hPa level. Data availability is heavily skewed towards Northern Hemisphere continental regions. In both versions of HadRT there is general warming, but with coherent regions of cooling over northeastern North America and northeastern Asia. The heterogeneous observations in some regions are probably due to temporal sparsity of data, along with likely residual systematic errors in some of the basic radiosonde observations (e.g. Parker et al., 1997). Figure 1 (lower panels) shows best guess model realisations with all anthropogenic forcings. The variability in the model fields is lower, most likely at least in part because they are ensemble averages, analysis not shown here yielding higher variability in individual ensemble members. Although both of the models capture a large-scale warming they fail to predict the regions of large-scale cooling seen in the observations, at least in the ensemble average, and HadCM3 ANTHRO overestimates the global average warming.

Lapse rates have been recommended as a detection diagnostic (Santer et al., 1996). They are less affected than individual level temperatures by the noise of inter-annual variability because, on an annual basis at least, regional temperatures at different tropospheric levels tend to co-vary. The HadRT deep troposphere (300 minus 850 hPa) lapse rates (Figure 2) exhibit geographically diverse behaviour, with areas of both relative upper tropospheric warming and cooling. HadCM2 and HadCM3 fields also exhibit heterogeneous behaviour, although less so than the observations, analysis again suggesting that this is most likely, at least in part, due to being ensemble averages. However, the standard deviation fields (not shown here) are of a similar magnitude to those for single level temperatures, whilst the absolute changes are smaller in magnitude, so noise may be a limitation in future quantitative detection studies considering lapse rates.

2.2 Assessing overall model skill

In the three decade case (Figure 3), no best guess ensemble scores are outside the range of that expected by chance due to natural variability. However, anthropogenic fields for both models show the greatest degree of skill relative to "no change". For the four decades

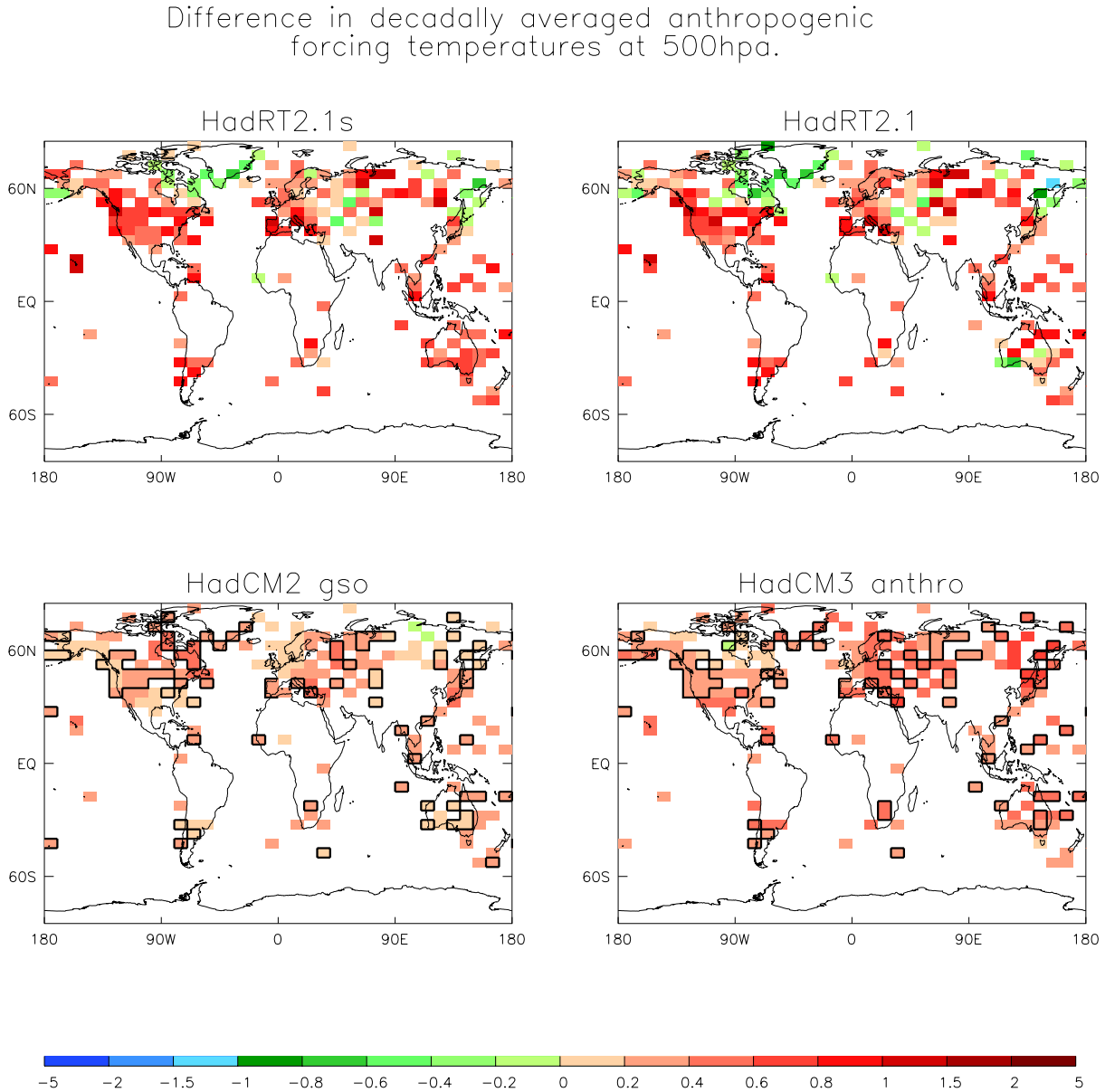

Figure 1. Observed and modelled fields of 500 hpa temperature change between 1963–1972 and 1983–1992. For the model fields boxed areas indicate regions in which the model disagrees with the observations at greater than 3 sigma. The scale is not linear.

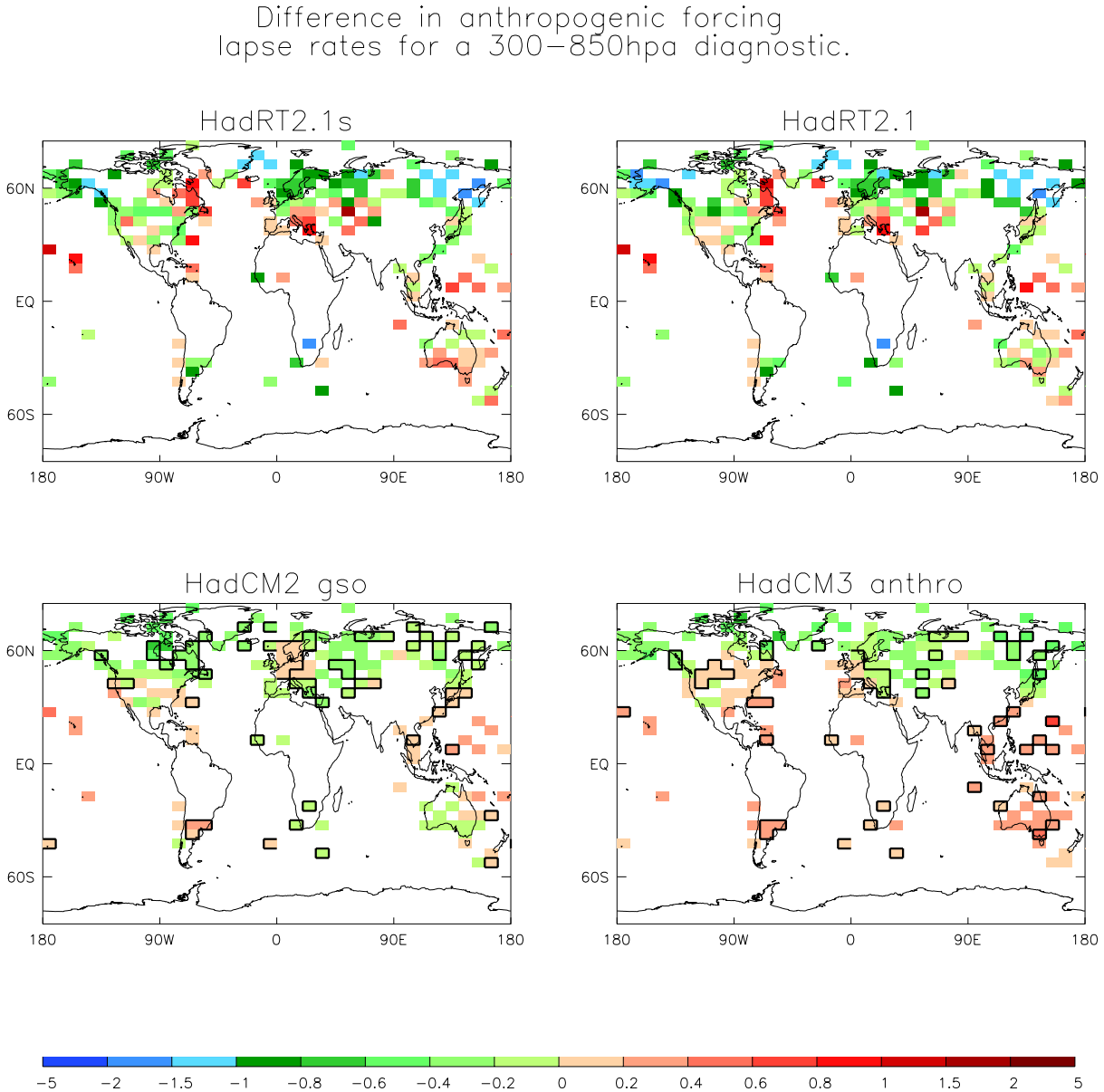

Figure 2. As Figure 1, except considering a tropospheric lapse rate (300–850 hPa).

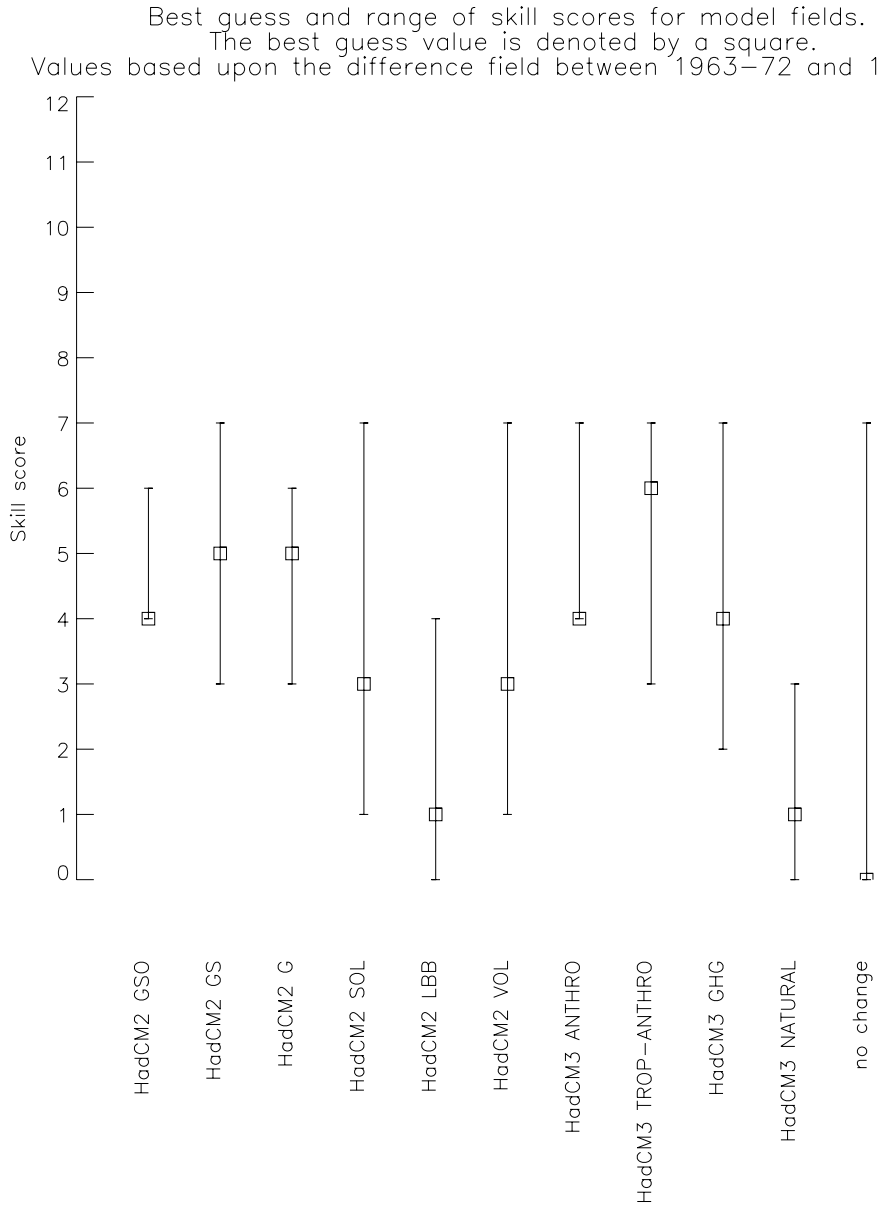

Figure 3. Plot of best guess and uncertainty ranges of scores for model ensembles using differences between 1963–1972 and 1983–1992. Best guesses are denoted by squares and the range of synthetic realisations is given by the bars (note: these do not denote strict statistical confidence limits).

case (Figure 4), the best guess score for HadCM2 GSO fields is greater than the range expected if "no change" were true, although the uncertainty ranges overlap. Further, GS exhibits a similar degree and range of skill, whereas G does not, indicating that the inclusion of anthropogenic aerosol effects is likely to be important in explaining the observations, at least for HadCM2. HadCM3 anthropogenic fields are broadly consistent with this interpretation, although they all exhibit a much lower degree of skill. GS and TROP-ANTHRO fields do not overlap, implying that there may exist significant differences in the response to the same forcing combination between the models: visual inspections show that HadCM3 tends to overestimate the tropospheric warming in all the anthropogenic simulations. We conclude that this is the most likely cause of the reduction in skill for HadCM3 as measured by our simple indicators, which assume both model pattern and amplitude are correct.

For natural forcings, HadCM2 SOL exhibits a degree of skill using our simple indicators, although LBB does not. Further the LBB and SOL best guesses are inconsistent with each other in the four decades case, which reduces our confidence in there being a discernible solar influence. If a solar influence were to be the sole reason for recent climate change, then HadCM2 is grossly underestimating the signal strength. There is no skill exhibited by the VOL forcing, most likely because of the choice of timescale and period used in the present study (Stott et al., 2001). Volcanic influences are only likely to be detectable on shorter time periods and during times of considerable explosive volcanic activity. For HadCM3, the NATURAL run exhibits no skill: whether this would continue to be true if the natural forcings were seperated to yield independent solar and volcanic signals is not known. The solar and volcanic signals could be cancelling to yield a very small and noisy signal.

3. Discussion

We have found a degree of geographical heterogeneity in upper air temperature and lapse rate changes, both in the observations and, to a lesser extent, in ensemble mean model fields. So previous detection studies using zonal mean temperature changes (e.g. Santer et al., 1996, Tett et al., 1996) may have been sub-optimal. In zonally-averaging upper air temperature series, we contend that potentially useful information is being discarded which could, in theory, be employed to discriminate between competing forcing mechanisms. We aim to address this in a more definitive manner in future work.

We have evaluated two potential parameters suitable for use in subsequent formal, quantitative, detection exercises; pressure level temperatures, and lapse rates. Two statistical tests have been applied to model fields of these parameters to objectively assess skill relative to a null hypothesis of "no change" in a simple quantitative manner assuming that model fields are correct in both pattern and amplitude. Anthropogenic forcing ensembles which incorporate the effects of sulphate aerosols provide the best overall explanation of the observed trends, although we cannot discount the possibility of a solar forcing influence. In no case is the model consistent in both pattern and amplitude with the observations throughout the troposphere and lower stratosphere. HadCM3 tends to overestimate the warming in its anthropogenic runs. In both models natural external influences are found to be too weak to account for the general warming observed. We conclude that observed trends are unlikely to have arisen by chance alone due to natural internal climate variability. These results are consistent with previous quantitative detection studies (Allen and Tett, 1999, Tett et al., 1999, 2001 for example). However, numerous caveats apply. The limited power of the quantitative approach employed here precludes us from making definitive statements of detection and attribution. We find our results are potentially sensitive to trend length and start year. Sufficient differences exist between HadCM2 and HadCM3 to conclude that a number of models may be required in detection studies using these data to avoid making ambiguous

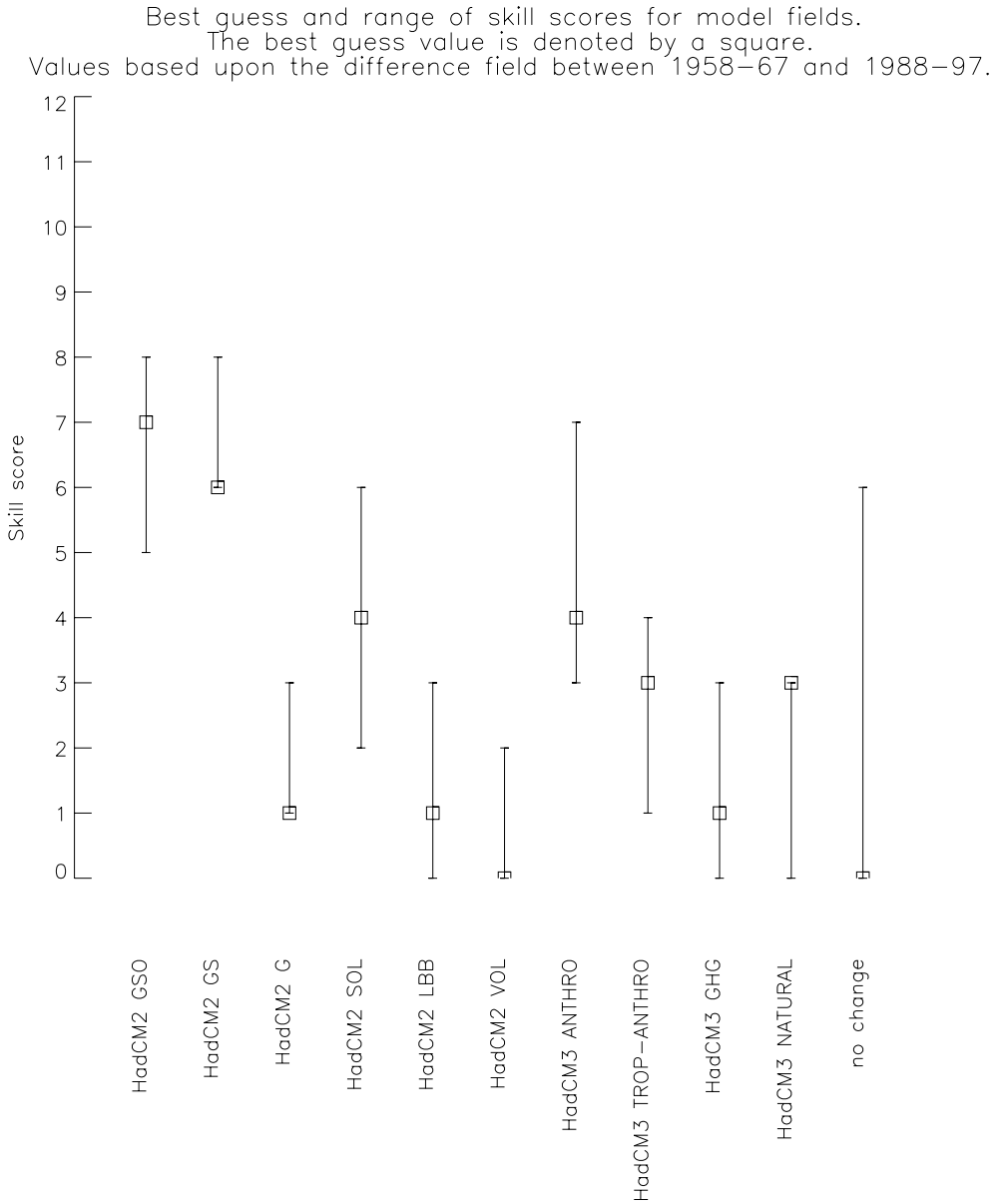

Figure 4. As Figure 3, except considering differences between 1958–1967 and 1988–1997.

conclusions. Results also differ when two different histories of the same forcing mechanism are applied. Therefore an accurate realisation of the forcing history is important to ensure unambiguous detection.

Acknowledgements

This work was carried out while PWT was in receipt of NERC studentship GT04/98/76/AS. The Met Office authors are supported by the U.K. Department of the Environment, Transport and the Regions under contract PECD7/12/37 and the Public Met. Service Research and Development 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*.

References

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

Barnett, T. P., Hasselmann, K., Chelliah, M., Delworth, T., Hegerl, G., Jones, P., Rasmusson, E., Roeckner, E., Ropelewski, C., Santer, B. and Tett, S., 1999. Detection and Attribution of recent climate change: A status report. *Bulletin of the American Meteorological Society*, **80**, 2631–2659.

Barnett, T. P., Pierce, D. W. and Schnur, R., 2001. Detection of anthropogenic climate change in the world's oceans. *Science*, **292**, 270–274.

Christy, J. R., Spencer, R. W. and Lobl, E. S., 1998. Analysis of the merging procedure for the MSU daily temperature series. *Journal of Climate*, **11**, 2016–2041.

Collins, M., Tett, S. F. B. and Cooper, C., 2001. The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments. *Climate Dynamics*, **17**, 61–81.

Hoyt, D. V. and Schatten, K. H., 1993. A discussion of plausible solar irradiance variations, 1700–1992. *Journal of Geophysical Research*, **98**(A11): 18895–18906.

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

Lean, J., Beer, J. and Bradley, R., 1995. Reconstruction of solar irradiance since 1610—implications for climate change. *Geophysical Research Letters*, **22**, 3195–3198.

Parker, D. E., Gordon, M., Cullum, D. P. N., Sexton, D. M. H., Folland, C. K. and Rayner, N., 1997. A new global gridded radiosonde temperature data base and recent temperature trends. *Geophysical Research Letters*, **24**, 1499–1502.

Pope, V. D., Gallani, M. L., Rowntree, P. R. and Stratton, R. A., 2000. The impact of new physical parameterizations in the Hadley Centre climate model—HadCM3. *Climate Dynamics*, **16**, 123–146.

Santer, B. D., Taylor, K. E., Wigley, T. M. L., Penner, J. E., Jones, P. D. and Cubasch, U., 1995. Towards the detection and attribution of an anthropogenic effect on climate. *Climate Dynamics*, **12**, 79–100.

Santer, B. D., Taylor, K. E., Wigley, T. M. L., Johns, T. C., Jones, P. D., Karoly, D. J., Mitchell, J. F. B., Oort, A. H., Penner, J. E., Ramaswamy, V., Schwarzkopf, M. D., Stouffer, R. J. and Tett, S. F. B., 1996. A search for human influences on the thermal structure of the atmosphere. *Nature*, **382**, 39–46.

Sato, M., Hansen, J. E., McCormick, M. P. and Pollack, J. B., 1993. Stratospheric aerosol optical depths (1850–1992). *Journal of Geophysical Research*, **98**(D12): 22987–22994.

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

Stott, P. A., Tett, S. F. B., Jones, G. S., Allen, M. R., Mitchell, J. F. B. and Jenkins, G. J., 2000. External control of twentieth century temperature by natural and anthropogenic forcings. *Science*, **290**, 2133–2137.

Stott, P. A., Tett, S. F. B., Jones, G. S., Allen, M. R., Ingram, W. J. and Mitchell, J. F. B., 2001. Attribution of twentieth century temperature change to natural and anthropogenic causes. *Climate Dynamics*, **17**, 1–21.

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

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

Tett, S. F. B., Jones, G. S., Stott, P. A., Hill, D. C., Mitchell, J. F. B., Allen, M. R., Ingram, W. J., Johns, T. C., Johnson, C. E., Jones, A., Roberts, D. L., Sexton, D. M. H. and Woodage, M. J., 2001. Estimation of natural and anthropogenic contributions to 20th Century temperature change. Submitted to Journal of Geophysical Research.

Wigley, T. M. L., Santer, B. D. and Taylor, K. E., 2000. Correlation approaches to detection. *Geophysical Research Letters*, **27**, 2973–2976.

Willson, R. C., 1997. Total solar irradiance trend during solar cycles 21 and 22. *Science*, **277**, 1963–1965.