3. An intercomparison of modelled and observed fields

Before proceeding to a rigorous quantitative detection and attribution study using the optimal detection methodologies as described in chapter 4, it is pertinent to undertake a degree of model validation. This chapter summarises the results of such an exercise using HadCM2 and HadCM3 model data and the HadRT radiosonde record, and discusses the possible use of certain variables as input diagnostics in subsequent detection and attribution studies. Intercomparisons are made at various temporal and spatial scales. In section 3.1, the methodologies used in this chapter are described. Section 3.2 gives results from an intercomparison of the fields derived from the two versions of the Hadley Centre model and the observed HadRT fields. The diagnostics are considered in terms of increasing complexity, from global-mean temperatures to fully 4-dimensional fields (longitude, latitude, height, time). Section 3.3 discusses the potential for using the full global radiosonde temperature fields, as well as likely suitable input variables for the later, more formal, quantitative detection and attribution studies. In section 3.4 the principal results of this chapter are summarised.

3.1 Method

3.1.1 Model data treatment

HadCM2 and HadCM3 are both fully 4-dimensional (x, y, z, t) coupled GCMs (see section 1.3 for more details). The model temperatures on pressure levels fields are available as annual (December to November years) and, for HadCM2, seasonal values. Both models have a horizontal atmospheric resolution of 2.5° Latitude by 3.75° Longitude. Therefore, the first step in any intercomparison is to re-interpolate these data to the resolution of the HadRT observations (5° x 10°), implemented by bilinear interpolation (Press et al., 1992). A different interpolation technique could have been used, but this is very unlikely to have a first order effect upon subsequent results, especially at the large spatial scales considered in formal detection studies. It is best to reinterpolate the models to the observational grid both because the model fields are complete, unlike the observations, and the interpolation is to a coarser grid.

Santer et al. (1999) and Allen and Tett (1999), amongst others, state that for any meaningful comparisons to be undertaken between datasets, they must exhibit the same spatio-temporal sampling characteristics, otherwise they could yield spurious results. Therefore, the interpolated model fields are sub-sampled to the available HadRT data mask. Spatio-temporal masking also means that additional variance introduced due to missing data can be implicitly taken into account when estimating natural variability. Trivially, missing data will inflate the variance for any given point in the dataset and, therefore, sub-sampling must be important in estimating trend significance.

Both Hadley Centre models have been run a number of times for each perturbation experiment under identical forcing, but with different initial conditions (sampled from the control run), to form an ensemble of possible responses (Tett et al, 1999, 2001, Stott et al., 2001). In this chapter, and the remainder of this thesis, these ensemble members are averaged to form an ensemble mean response. This has the advantage of enhancing the signal-to-noise ratio (Allen and Tett, 1999). It is, therefore, expected that the observations will exhibit greater variability than the model fields with which they are being compared.

In both models there are five ensemble average fields that are broadly comparable in terms of the forcings applied. The following are the model equivalent forcings with HadCM2 quoted first; GSO≡ANTHRO, GS≡TROP-ANTHRO, G≡GHG, LBB≡SOLAR, and VOL≡VOLCANIC (Tett et al., 1999, 2001, Stott et al., 2001, see also chapter 1). In addition, HadCM2 contains a further solar ensemble (SOL), and HadCM3 contains a NATURAL ensemble that combines the effects of SOLAR and VOLCANIC. In most subsequent analysis and figures the anthropogenic ensembles are emphasised, although results from other forcings are also referred to. In all comparisons, V2 (see chapter 2) of both HadRT2.1 and 2.1s are used, although the statistics, and areas of significance in the Figures, quoted in Section 3.2 relate solely to the HadRT2.1s fields. The use of both HadRT versions in the qualitative comparisons provides a visual estimate as to the uncertainty associated with the observations and their treatment within the troposphere.

The HadRT radiosonde temperature record is available for the period 1958 to present as anomalies relative to 1971-90 climatological values. These climatological averages include periods of missing data and, therefore, the sub-sampled model fields are used to calculate normalised values for each individual ensemble member relative to its 1971-90 climatological temperatures. This enables direct comparisons to be made with the observations. The masks are not exactly coincident, as the HadRT data are available on a higher (monthly) time resolution, but they are felt to be sufficiently similar. All subsequent results and analyses relate to these normalised model data rather than the absolute temperature values, as a consideration of patterns of change is desired here rather than absolute values. Therefore, any bias in the model absolute values will not be accounted for in this analysis. Both HadCM2 and, to a lesser extent, HadCM3 exhibit a noticeable tropospheric cold bias (Pope et al., 2000).

3.1.2 Diagnostics used in the intercomparison

The simplest comparison possible is that between global-mean modelled and observed temperature series. For the purposes of such a comparison, the annual global-mean temperatures on pressure levels are averaged via a simple calculation. Weighting is given according to latitude to compensate for decreasing gridbox areal coverage with increasing latitude, and an annual global-mean temperature series for each pressure level derived. These series are then visually compared to derive qualitative similarities and differences in the trends. As stated previously (see chapter 1), the use of such univariate diagnostics is of debatable value in any quantitative detection study as unambiguous detection or attribution is not possible under such an approach. Therefore, no attempt is made to assess the significance of any similarities or differences seen in this analysis in a rigorous quantitative manner.

To make comparisons easier for more complex analyses, decadal mean values are calculated, thus reducing the noise inherent in the annual timeseries. In the construction of decadal mean series, criteria must be implemented for the inclusion or rejection of individual grid-box series. In this chapter, a decadal average value is calculated for any given grid-box if more than five years contain data in any given decade, and there is no more than two consecutive years break in data availability. These criteria are essentially arbitrary in nature, being a trade-off between data coverage and the higher variance due to low temporal sampling concentration. Furthermore, in this chapter the criteria of the number of months required for the calculation of seasonal and annual values in the HadRT dataset is not taken into account, only V2 being used (see chapter 2). Two decadal datasets are constructed, a four-decade case covering the period 1958-1997, and a three-decade case over a sliding window from 1958-1987 to 1967-1996. The four-decade case has advantages in terms of trend length, which has been found by various authors to be critical in detection studies (Barnett et al., 1998, amongst others). However, the start and end periods of the HadRT record have poorer spatial sampling, as illustrated by Figure 3.1, and particularly poor representation in certain regions, which could have an impact on the results. Using a sliding three-decade window would have the advantage that the effect of the choice of start year can be factored into any analysis.

A last decade minus first decade analysis is considered in the comparisons here. This has the drawback that the transient nature of any change over the period under consideration will not be considered. It also reduces the data coverage slightly, as individual grid boxes may not necessarily contain data in both the first and last decades of any period. However, there are advantages to using a difference analysis. Firstly, the period over which the series are normalised is contained within the comparison. If two divergent trends are normalised over a mutual period then, during the normalisation period, both series will tend around zero and, therefore, yield spuriously high correlations. Using a first decade minus last decade analysis ensures that the trend end points are considered, thus reducing any chances of spuriously claiming a correlation when none exists. Furthermore, expectations are that such an analysis will yield a stronger signal. For the three-decade case, the period 1963-1992 is used in this analysis, as it yields the greatest areal data coverage of all possible three-decade periods considered.

Increasing the dimensionality of the problem from a consideration of global-mean values, zonal-mean (latitude-height) diagnostics are considered. In this comparison the first decade minus last decade diagnostic is used. As was the case for global-mean temperatures, the comparison is limited to be solely qualitative in nature; no attempt is made to rigorously quantify the level of agreement between zonal-mean fields. A fixed-signal optimal detection exercise is carried out using zonal-mean temperature fields in chapter 5, complementing previous such studies (Allen and Tett, 1999, Tett et al., 2001 for example).

A further increase in the dimensionality leads to a consideration of temperatures on various pressure levels. Available HadRT pressure levels are the WMO standard reporting levels: 850hPa, 700hPa, 500hPa, 300hPa, 200hPa, 150hPa, 100hPa, 50hPa, and 30hPa. The intra-ensemble variance is used to construct a realisation of the noise due to internal climate variability for grid-box temperatures at each pressure level. Estimates of this variance are made for both HadCM2 and HadCM3, and are treated independently since the variance is likely to be model-dependent. The noise estimate is used to determine those areas of significant disagreement (on a grid-box basis) with the observed trends, using a high threshold of $\pm 3\sigma$ to guard against spurious rejection. The assumption under this approach is that the model perfectly captures the response to all forcings important in explaining the observations. Therefore, assuming a normal distribution due to natural internal climate variability, only one percent of the grid boxes should fail this test. This is unlikely to be the case even if the model were to perfectly capture all the important forcings. HadCM2 has been shown to underestimate near-surface temperature variability at scales below 2000 kilometres (Stott and Tett, 1998). Expectations are that this will also hold true for upper air temperatures, and that HadCM3 will exhibit similar characteristics. However, the rank of agreement should provide a robust indication as to similarities between modelled and observed fields.

Additionally, to provide an indication of overall field similarity, a root mean squared difference (RMSD) statistic is used to provide a quantitative realisation of the overall field similarity at each pressure level. Such a statistic is sub-optimal, in that it does not take into account spatial auto-correlation effects and, therefore, will likely give artificially significant results (Wigley et al., 2000). In the method used here, this is not expected to be important, as all the fields considered are likely to exhibit similar spatial auto-correlation structures, although this assumption is not tested in any explicit manner.

For both statistics a null hypothesis of "no change" field is constructed whereby all grid-box data points are set to zero. The same tests are applied to this field as they are to the modelled fields. If the tests provide better results for the modelled runs than for this null hypothesis field, then it is concluded that the model provides a degree of skill in predicting the observed changes relative to "no change". For RMSD statistics, this approach is likely to be highly conservative, as the "no change" null hypothesis field employed in the comparison is perfectly spatially auto-correlated (and therefore likely to yield a spuriously low RMSD value). This would be an unobtainable result in the real world, at least on annual to multi-decadal timescales.

Finally, in an attempt to discriminate in the height dimension, consideration is extended to changes in tropospheric lapse rates utilising the same two simple statistical measures as for temperatures on pressure levels. Three lapse rates are considered: an entire troposphere (300-850hPa); an upper troposphere (300-500hPa); and a lower troposphere (500-850hPa) lapse rate. Expectations are that considering lapse rates will reduce noise due to the effects of missing data, as the troposphere is well mixed on annual timescales and, therefore, temperatures on individual pressure levels within the troposphere will tend to co-vary for any given point in Latitude-Longitude space. In the current analysis, lapse rate calculations only consider information from the two pressure levels quoted, being a simple upper level minus lower level calculation, and do not consider any intervening levels. Lapse rates will only provide any useful information if the troposphere, or the portion thereof considered, exhibits differential rates of temperature change with altitude.

Significance of the grid-box similarity and RMSD statistics for temperature on pressure levels and lapse-rate diagnostics is estimated by creating synthetic estimates of physically plausible model and "no change" fields. For these purposes a 1200-year chunk of the HadCM3 control, run with no changes in prescribed external forcings, is employed. Independent segments of the control are extracted, yielding a total of 36 three-decade segments and 27 four-decade segments. Each control segment is converted to anomalies in the same manner as the other model fields. These segments are then added to the ensemble mean fields, scaling by one over the number of ensemble members to account for ensemble size effects, to create a series of synthetic realisations. These synthetic realisations are then treated in exactly the same way to provide a range of grid-box similarity and RMSD values consistent with HadCM3 predicted natural internal variability for each ensemble. Collins et al. (2001) note how HadCM3 may generally underestimate this natural variability when compared to the observations, at least in a zonal-mean sense, especially within the stratosphere and mid-latitude Northern Hemisphere troposphere. Hence, results quoted here might be systematically biased; being based upon the assumption that HadCM3 control is a demonstrably adequate representation of real-world variability.

A very simple (naïve) measure of model skill is subsequently realised. If an ensemble average realisation truly exhibits skill relative to a null hypothesis of "no change" in explaining the observations, then it should exhibit skill over a broad range of scales. Ensembles are treated on a level-by-level (or lapse-rate) basis and, if any level has a lower proportion of grid-box values in significant disagreement and a lower RMSD value than the "no change" null hypothesis field, then its score is incremented by one. This analysis is repeated for both the ensemble average (best guess value) and population of synthetic data for that ensemble average (to yield an estimate of uncertainty) for each ensemble, and the "no change" field.

3.2 A comparison of modelled and observed upper air temperature diagnostics

3.2.1 Global-mean temperature trends

Figure 3.2a shows the results of a comparison of HadRT2.1s with output from the HadCM2 forced runs. The most striking feature is that the observations exhibit far greater variation than the model ensemble averages, as expected. How much of the variation relates to residual errors in the observations is uncertain, but following the quality control procedure described in chapter 2, and previous efforts by Parker et al. (1997), it is considered likely to be relatively small. Comparing the model trends with the observations, two forcings immediately appear important. In terms of describing the overall trend in the observations the GSO (greenhouse gases, sulphate aerosols and stratospheric ozone) run is the most consistent explanation. Both of the other anthropogenic forcings (GS and GHG) underestimate the observed stratospheric cooling, implying that ozone changes are important in explaining recently observed stratospheric temperature trends. However, it is also apparent, following periods of explosive volcanic activity such as Pinatubo in 1991, that VOL is important in explaining the observations, although HadCM2 appears to over-estimate the lifetime of the effect by a few months to a year.

Results considering HadCM3 fields (Figure 3.2b) are broadly comparable to those for HadCM2. The best overall explanation arises when anthropogenic factors are taken into account (ANTHRO≡GSO in HadCM2), ozone changes again being required to explain the stratospheric trends. Volcanic influences are also needed to explain some of the short term stratospheric warmings and tropospheric coolings seen in the observations although, as for HadCM2, the lifetime of the effect would seem to be overestimated. The absolute magnitude of the model-predicted effects of volcanic incidents is also greater for HadCM3 than is the case for HadCM2, but there is insufficient evidence in the observations to suggest which, if either, is a more realistic estimate. Comparing Figures 3.2a and 3.2b, there are no other noticeable discrepancies between the models at the global-mean scale when considering equivalent forcing scenarios.

3.2.2 Zonal-mean temperature changes

Figure 3.3 shows differences in the zonal mean temperatures between 1963-1972 and 1983-1992. The top two panels provide a guide as to the likely effect of uncertainty in the observations when considering zonal mean diagnostics. HadRT2.1s and HadRT2.1 fields are generally of consistent sign, with differences in magnitude generally of less than half a degree Celsius. For anthropogenic fields, patterns tend to be consistent between the two models for comparable forcing scenarios. HadCM3 anthropogenic forcing runs tend to estimate stronger tropospheric and, particularly, tropical upper tropospheric warming than their HadCM2 equivalents. Both models estimate greater warming in this region than is evident in either set of observations. However, the observations are sparse at these elevations in this region and, therefore, any individual erroneous grid boxes will potentially have a large effect upon the calculated observed zonal-mean value. Confidence in the observed zonal-mean values is generally poorer outside the well-sampled Northern Hemisphere mid-latitudes. Qualitatively, the best explanation of the observations occurs when all anthropogenic forcings are considered, regardless of the model being used. Zonal-mean changes due to natural forcings according to HadCM2 and HadCM3 (not shown here) are small and of variable sign, implying that a large part of the natural forcing signals (on decadal timescales) is likely to arise from noise due to random variations, and temporal sampling effects. These results are, broadly speaking, consistent with previous quantitative zonal-mean detection study results considering these model fields (Tett et al., 1996, Allen and Tett, 1999, Tett et al., 2001).

3.2.3 Changes in temperatures on pressure levels

The observed and modelled changes between the periods 1963-1972 and 1983-1992 are compared graphically before advancing to considering the simple statistical measures outlined in section 3.1. Results for the differences between 1958-1967 and 1988-1997 are essentially similar to those shown here, but they contain substantially fewer data (> 20% at all levels, increasing with height). For the purposes of this comparison, two levels are concentrated upon: a lower stratospheric series (100hPa; above this altitude data availability degrades significantly, see Figure 3.1); and a mid-tropospheric series (500hPa). Subsequently, results from the simple statistical indicators (section 3.2.2) for both three-decade and four-decade cases are also discussed for all other levels.

Fields from the 100hPa (lower stratosphere) level are shown in Figure 3.4. Areas of model output highlighted by boxes indicate those regions which are outside ± 3σ of the HadRT2.1s observations, where σ is estimated from the respective model's intra-ensemble variability (shown in the bottom panels in grey scale). Both HadRT series are identical at this level, the difference between the two series only being in the tropospheric corrections applied. During the period there has been a general stratospheric cooling observed, although this is by no means uniform, and in some regions (Southern US, Central Pacific, Caucasus), there is warming. There appear to be some anomalous points remaining (e.g. Southern China, Iberia), but such points could appear anomalous solely due to temporal sampling effects in HadRT. Having already undertaken a degree of spatial quality control upon the HadRT dataset (chapter 2), it is supposed that this provides at least part of the explanation. Equally, there exists the potential for residual biases in the observed data due to, for example, sonde balloon burst effects, which could yield spurious cooling trends at this level.

Considering HadCM2 (HadCM3) anthropogenic fields at the 100 hPa level (Figure 3.4), only GSO (ANTHRO) captures the large-scale cooling seen in the observations. There are differences between the respective model fields for the two models however; most noticeably over Oceania where the model-predicted patterns are of opposite sign. Also, there are a large number of grid boxes where the model-predicted change is inconsistent with the observations at the 3σ level (boxed areas). For HadCM2, GS predicts a small warming with areas of very slight cooling, whilst HadCM3 TROP-ANTHRO predicts regions of both warming and cooling, both of greater magnitude, with a greater overall warming. Both models predict large scale warming under the G (GHG) scenario, the warming being greater for HadCM2 than HadCM3.

The qualitative results described above are confirmed in Tables 3.1 and 3.2. Table 3.1 presents the percentage of grid-box values that are greater than 3σ from the observations for all model ensembles and, for both three- and four-decade analyses. Considering the three-decade case at the 100hPa level for both models, the GSO/ANTHRO fields are consistent with the greatest proportion of the individual grid-box observations. This is a robust result for HadCM3, being insensitive as to whether a four-decade or three-decade trend length is considered. However for HadCM2, LBB provides the best explanation when considering a four-decade analysis. Although SOLAR also exhibits skill for the HadCM3 four-decade analysis, it is not as marked. Both of these ensembles utilise the Lean et al. (1995) solar forcing reconstruction. Confidence would be higher if there were a similar skill exhibited for SOL in HadCM2, an independent solar forcing ensemble based upon Hoyt and

Schatten (1993) and Willson (1997). However, this is not the case, implying either that the SOL forcing history is incorrect, or that the positive result for the Lean et al. (1995) forced runs in both models is a statistical fluke. Considering RMSD values (Table 3.2) for both three- and four-decade analyses, GSO/ANTHRO consistently provides the best explanation of the observed trends at 100 hPa. For neither model in the four-decade case does LBB / SOLAR exhibit skill and, therefore, confidence in the robust presence of a solar signal in the observations is further reduced.

The 500hPa HadRT temperature fields as shown in Figure 3.5 (top panels) are largely similar to each other, as corrections for the HadRT2.1 series have only been made at a few points (Parker et al., 1997), and are relatively small in many cases. HadRT2.1 exhibits accentuated high-latitude Northern Hemisphere cooling, and greatly reduced warming over Southern Australia. In both HadRT series, the large-scale overall trend is one of warming in the tropical and extra-tropical regions, with more chaotic patterns of change in high northern latitudes. In neither case are the patterns zonally homogeneous, implying that such information could be employed to gain extra power to discriminate between forcing mechanisms, compared to previous zonal-mean detection studies (Allen and Tett, 1999, Tett et al., 1996, 2001, Santer et al., 1996a).

Although the observations would tend to suggest a degree of zonal heterogeneity in temperature trends at 500hPa, model predicted trends are far more zonally homogeneous, especially in HadCM3. For both models, the anthropogenically forced runs show a warming, with only runs including stratospheric ozone depletion indicating any regions of cooling, and even then these are underestimated in extent compared to the observations. However, from a visual inspection it is difficult to conclude that any single anthropogenic forcing is a better explanation of recently observed changes. Considering Tables 3.1 and 3.2 for the three-decade case, a consideration of the effects of greenhouse gases and sulphate aerosols provides the best explanation of individual grid-box trends, although all model fields exhibit skill relative to “no change”. However, considering the four-decade case, solar forcings (SOL and SOLAR) also provide a similar degree of skill, on a grid-box basis, to anthropogenic forcings. Confidence in the result is reduced, as the HadCM2 solar realisation LBB (equivalent to SOLAR in HadCM3), does not show similar skill.

These results for HadCM2 are consistent with RMSD values (Table 3.2), unlike the 100hPa case, giving increased confidence. For HadCM3 fields, the anthropogenic ensembles do not exhibit skill in the RMSD values for the four-decade analysis, whereas SOLAR does. A visual inspection of the anthropogenic HadCM3 fields in Figure 3.5 indicates that the most probable cause of the anthropogenic ensembles not exhibiting skill is that they overestimate the warming trend at the 500 hPa level. The SOLAR forcing (not shown) also contains a net overall warming, although this is greatly reduced compared to the anthropogenic fields.

Results for both grid-box consistency tests and RMSD statistics on all levels are shown in Tables 3.1 and 3.2 respectively. In general, there are more cases of fields exhibiting skill at the grid-box level, but not for RMSD values. This may relate to the conservative approach used in constructing the "no change" field, which is perfectly spatially auto-correlated. For both statistics, the models exhibit least skill at the 150hPa and 200hPa levels – the region of the tropopause. From a consideration of Figures 3.2a and 3.2b, the large-scale pattern of change is one of tropospheric warming and stratospheric cooling. Therefore, in this region the observed trends tend towards zero, and there is likely to be a reduction in the perceived model skill relative to the "no change" null hypothesis field, even if the model were to exhibit the same degree of skill as in other regions of the atmosphere. This may also impact upon the skill of quantitative detection studies in this region of the atmosphere, although there remain spatially distinct trends (not shown). For the four-decade analyses great care should be exercised in interpreting stratospheric values for both tables, as the sample size is very small.

Any model ensemble should exhibit skill in explaining observed changes at a range of scales relative to a null of "no change" if it includes some or all of the important forcing factors. The grid-box and RMSD values in Tables 3.1 and 3.2 are, therefore, used to ascertain estimates of overall model skill. For any level, a score of unity is given if both the RMSD and grid-box scale values exhibit skill relative to "no change" (values highlighted in yellow in Tables 3.1 and 3.2). The best guess scores summed over all levels are shown in Table 3.3. Also shown in brackets are the confidence limits resulting from the addition of independent HadCM3 control run segments as pseudo-noise estimates. There is no *a priori* reason why the uncertainty limits should be symmetric about the best guess score value for any ensemble. Interpreting the results is potentially very complex. Here, a very simplistic approach is pursued, whereby it is concluded that a model ensemble provides skill in explaining the observations if its score uncertainty does not encompass zero. Confidence in this statement is enhanced the greater the separation of the lower limit of the uncertainty range is from zero.

For both models and trend lengths, a consideration of all anthropogenic influences generally provides the best explanation, consistent with previous detection and attribution studies considering zonal-mean upper air temperature patterns (Santer et al. 1996a, Tett et al., 1996, Allen and Tett, 1999). However, in all cases at least some natural external forcings also exhibit a degree of skill, particularly SOLAR in HadCM3 and SOL for HadCM2 for four-decade analyses, implying that it is not necessarily only anthropogenic factors that are important in explaining recently observed tropospheric temperature trends. The uncertainty ranges in the best estimate scores vary widely between ensembles, being generally greater for the shorter three-decade analyses, and of smaller magnitude than the uncertainty range in the null hypothesis “no change” field. Adding small perturbations to the “no change” field will yield a non-zero score. Comparing the skill of the perturbed field to that of the original “no change” field in explaining the observations, expectations are that a number of levels will exhibit skill. Adding such random perturbations to other fields is only likely to have an effect on scores for those levels which exhibit similar skill in explaining the observations to the “no change” null hypothesis field.

3.2.4 Changes in tropospheric lapse rates

It can be seen from Figure 3.5 (top two panels), that there are distinct changes observed when considering an entire troposphere lapse rate diagnostic and, therefore, the use of lapse rates could prove useful in detection and attribution studies. Results using either an upper (300-500hPa) or a lower (500-850hPa) troposphere lapse rate yield broadly similar pattern observational fields.

HadRT2.1s and 2.1 exhibit similar changes in entire troposphere lapse rates between 1963-1972 and 1983-1992. In tropical regions the upper troposphere is generally warming relative to the lower troposphere. In these regions the Environmental Lapse Rate is approximately equivalent to the Saturated Adiabatic Lapse Rate, at least over ocean regions. Therefore, any surface warming will tend to be accentuated with height in tropical regions. Away from the tropics, the predominant change is one of relative lower tropospheric warming, although this is by no means uniform. There are regions of both distinct relative lower tropospheric cooling (over North East America, the Mediterranean, Central Asia, and parts of Australia (in 2.1s, but not 2.1)) and warming (most high Northern latitudes, and Southern Africa). Such a relatively complex observational field should provide power in discriminating between different potential forcing mechanisms. This makes the assumption that the changes are real, and not the manifestation of residual errors in the HadRT dataset. Confidence in this is increased as both models also exhibit zonally heterogeneous behaviour in their entire troposphere lapse rate diagnostics.

Considering the anthropogenic model runs, results for equivalent forcing scenarios show little overall field agreement between the models solely on the basis of visual inspection (Figure 3.5). The implication from this must be that changes in the model parameterisations (Pope et al., 2000) and forcing algorithms (Tett et al., 1999, 2001) have had important effects on lapse rate diagnostics. This illustrates the importance of not relying upon a single model in formal detection and attribution studies for such diagnostics, in agreement with previous detection studies considering near-surface temperatures (Hegerl et al., 2000, Barnett et al., 1999, 2000). As is the case for temperatures on pressure levels, more grid boxes fail to capture the observed changes than would be expected by chance (Figure 3.5 middle 3 panels; Table 3.1). It is difficult to determine which fields are providing the best overall explanation. Considering RMSD values (Table 3.2) for both models and trend lengths, a consideration of all anthropogenic forcings yields the best (or equal best) explanation. Ignoring the effects of stratospheric ozone depletion reduces the agreement, but it remains a better explanation than “no change”. The agreement is encouraging because the global average change is around zero (actually very slightly negative (Figure 3.5 top panels)), which means that the models are exhibiting a degree of skill in estimating the patterns of observed changes in entire troposphere lapse rates.

Results shown in Tables 3.1 and 3.2 for a lower troposphere and an upper troposphere lapse rate are less conclusive than those considering a whole troposphere lapse rate. This is likely due to the signal strength in most regions being additive, lower troposphere lapse rate trends being generally of the same sign as the upper troposphere lapse rate trends for any given grid box (not shown here). For the lower troposphere trends, it is possible that natural forcings could explain at least some of the observed trends, but this result appears to be both trend-length and forcing-history dependent, reducing overall confidence in the importance of natural external forcings.

Table 3.3 also presents best-guess scores and uncertainty ranges for lapse-rate diagnostics. The uncertainty estimates in lapse rates are proportionately larger than those for individual pressure levels, which implies that noise due to natural internal variability may be a limitation in quantitative detection studies that consider such diagnostics. There are very few model fields for which the uncertainty range in the score does not encompass zero and, therefore, for which it can be confidently concluded they exhibit true skill. For the three-decade case there are no model ensemble responses which exhibit skill using this diagnostic. In the four-decade case there is a possible SOLAR influence, although neither LBB nor SOL also exhibit skill, again reducing confidence in there being a robust solar influence. In the four-decade case for both models, for anthropogenic forcings to exhibit skill in explaining recently observed lapse rate changes, the effects of sulphate aerosol forcings (and stratospheric ozone depletion for HadCM3) must be considered.

3.2.5 Overall model skill

In Figures 3.7 and 3.8 the skill scoring system as described previously for temperatures on pressure levels and lapse-rate diagnostics is used to derive an overall ‘skill-score’ value based upon all variables considered (maximum possible value of 12) for the three-decade and four-decade analyses respectively. The assumption is that if the lower boundary of the uncertainty range is separated from zero, then the model ensemble exhibits skill in explaining the observations, confidence increasing with the degree of separation. Furthermore, if the best guess estimate is greater than the upper-bound of the uncertainty range in the null hypothesis “no change” field, then it is concluded that it should be detectable. Caveats apply to this assertion as the true numbers of degrees of freedom of the statistical approach are unknown, and so rigorous significance estimates cannot be attached to such statements.

In the three-decade analyses (Figure 3.7), ensembles for which it can be confidently asserted the models exhibit skill are those which consider anthropogenic forcings. At these timescales and using such 'skill-score' diagnostics, it is not possible to differentiate between the skill exhibited by the individual anthropogenic forcings. There are also potentially weak solar and volcanic signals evident, but skill is dependent on both model and forcing-scenario, which reduces confidence in these being truly detectable signals. None of the best guess estimates are outside the range of 'skill-score' values expected by chance, so no claims are made as to the likely detectability of signals on this timescale.

For the four-decade analyses (Figure 3.8), anthropogenic forcings exhibit much greater skill when the effects of sulphate aerosols are included for both models. Greenhouse gases on their own appear to be an inadequate representation of the observed trends. There is also a potentially much stronger solar signal than was the case in the three-decade analyses. However, LBB does not exhibit skill, although both SOLAR (a HadCM3 realisation of the same LBB forcing signal) and SOL do; therefore solar ensemble skill is found to be both model and forcing-history dependent. For HadCM2, the GSO ensemble has a best estimate ‘skill-score’ that is greater than the uncertainty range for the null hypothesis field and, therefore, it is likely to be detectable. There is no similar result for the HadCM3 ANTHRO field, however, reducing the confidence in there existing a demonstrable anthropogenic influence, at least considering the simple statistics employed here.

3.3 Suitability for use of the full field HadRT temperature record in formal detection and attribution studies

Results from section 3.2 point towards a discernible human influence upon recently observed free atmosphere temperatures. However, these results in themselves are insufficient to claim formal detection and attribution. For such purposes a more formal, quantitative, approach is required. Previous detection and attribution results have been introduced in chapter 1, and the optimal regression algorithm is explained in detail in chapter 4. Here, consideration is given as to the potential of extending current optimal regression detection and attribution studies to full field upper air temperatures.

Results indicated in Figures 3.4, 3.5, and 3.6, show that for both temperatures on pressure levels and lapse-rate diagnostics there is a degree of zonal heterogeneity in both the observations and the models. This means that in zonally averaging the data before proceeding to detection and attribution, as has been the case previously (Allen and Tett, 1999, and references therein), a proportion of information is being lost. Retaining such information should enable a better signal derivation and reduce problems of signal degeneracy (see Allen and Tett, 1999, and Tett et al., 1999, for a discussion on degeneracy) and, hence, improve upon previous detection and attribution studies.

The optimal detection methodology of Allen and Tett (1999) only requires a vector-input diagnostic from which the leading EOFs can be calculated. For the near-surface temperature record, spherical harmonics are employed (Tett et al., 1999, Stott et al., 2001) to consider the largest-scale properties which the model simulates best (Stott and Tett, 1998). The first five spherical harmonic wave numbers are used to derive an input vector of 25 values (Tett et al., 1999). The HadRT observations are far more data-sparse than the HadCRUTv near-surface observations, and contain little information over oceanic regions and in high Southern latitudes. Furthermore, confidence is lower in the quality of HadRT data from poorly-sampled regions, as near-neighbour consistency checks have not been possible in these regions (chapter 2). It is, therefore, inappropriate to use spherical harmonics in detection studies considering upper atmosphere temperature diagnostics using HadRT records, as they are unlikely to be stable, and could be greatly influenced by (potentially erroneous) values in data sparse regions.

For detection studies involving upper atmospheric temperatures a different input vector is therefore required. Any input vector must contain information on large-scale changes, and confidence must be high in the accuracy of the observations. Hence it is proposed that some form of large-area averages (LAA henceforth) be used in constructing the input vector in this thesis. To ensure data representativeness, a number of reporting grid boxes are desirable for any region to calculate a value. This further increases bias towards Northern Hemisphere mid-latitude continental regions. The number of contributing grid boxes required should be large enough to mitigate the effects of individual erroneous grid-box values. However, this needs weighing against the necessity for an input vector size sufficient to reduce signal degeneracy and increase confidence in the results of regression analysis, as well as requirements of spatial representation. A number of different input vectors derived from different combinations of LAAs should ideally be utilised to ensure against spurious results. In chapter 6 four choices of LAA inputs are used (see chapter 6 for details of their construction).

3.4 Conclusions

In this chapter it has been shown, consistently and across a range of indicators from univariate global-mean diagnostics to tropospheric lapse rates, that anthropogenic influences are likely to be required to explain recently observed HadRT radiosonde upper air temperature trends. These findings are consistent with those from previous quantitative detection studies considering zonally-averaged upper air temperatures (Santer et al., 1996a, Tett et al., 1996, 2001, Allen and Tett, 1999), and near-surface temperatures (Tett et al., 1999, 2001, Stott et al., 2001, 2001a) for the same models. However, the statistical indicators considered in this section are relatively simple in nature and cannot be directly used to robustly answer questions of detection and attribution. What they do provide, albeit rather weakly, is additional evidence for an anthropogenic influence on climate. This does not discount the possibility that natural external forcings could also be important in explaining, at least in part, recent trends. Nor does it entirely rule out natural variability as a (partial) cause of the recently observed changes.

Previous zonal-mean detection studies (Santer et al., 1996a, Allen and Tett, 1999, Tett et al., 1996, 2001) are likely to have been sub-optimal because both the observational and model datasets exhibit a degree of zonal heterogeneity. In zonally averaging these data, potentially useful information is, therefore, being lost, which could be used to discriminate between competing forcings.

Two GCM datasets have been considered, and results using simple statistical indicators have been shown to be model-dependent. There is no reason to believe that this will not also be the case for the more complex optimal detection and attribution statistical approaches described in chapter 4 which are used in subsequent studies in this thesis. Therefore, caution is advised when considering only a single model, as this could yield ambiguous results, consistent with findings for near-surface temperatures (Hegerl et al., 1999, Barnett et al., 1999, 2000). Furthermore, Santer et al. (1999) conclude that the use of a single observational dataset could also yield ambiguous results. No other gridded observed upper air temperature product of similar quality and length exists to date, both NCEP and MSU records containing known limitations. Therefore, both versions of HadRT should be used in subsequent analyses. This is far from perfect as the two datasets are not truly independent.

It has been argued here that the use of spherical harmonics to filter the data (Tett et al., 1999, Stott et al., 2001) is unlikely to be stable for detection diagnostics based upon HadRT data. This is due both to the heavily biased data coverage, and reduced confidence in data quality from data sparse regions (chapter 2). An alternative methodology is proposed whereby the input vector is derived from large-scale area averages for a number of distinct regions.

Percentage of grid boxes greater than 3 sigma different to the observations.

Pressure level	850 hpa	700 hpa	500 hpa	300 hpa	200 hpa	150 hpa	100 hpa	50 hpa	30 hpa	500-850	300-500	300-850
3 DECADES												
num. of grid boxes	182	189	192	191	187	185	186	104	81	181	191	181
HadCM2 gso	**40**	**37**	**42**	**39**	60	53	**37**	**49**	**57**	36	**36**	38
HadCM2 gs	**26**	**31**	**35**	**41**	**27**	**30**	45	**51**	**54**	**32**	38	35
HadCM2 ghg	**41**	**38**	**37**	**41**	36	33	70	**70**	**58**	36	45	41
HadCM2 sol	**38**	**35**	**37**	**39**	**30**	**32**	65	80	80	34	39	**33**
HadCM2 lbb	51	60	**54**	**47**	37	37	46	82	73	45	44	36
HadCM2 vol	**45**	**51**	**48**	**42**	36	35	59	79	83	42	**35**	35
no change	50	59	56	49	36	33	40	79	69	36	38	35
HadCM3 anthro	**27**	**29**	**43**	**51**	49	59	**52**	**56**	**57**	31	39	**31**
HadCM3 trop-anthro	**24**	**25**	**35**	**58**	**39**	**39**	70	**75**	**44**	**25**	40	**31**
HadCM3 ghg	**22**	**26**	**40**	**62**	**43**	44	74	77	**47**	37	47	43
HadCM3 natural	**41**	**50**	**59**	**59**	49	52	67	82	73	35	40	43
HadCM3 solar	**39**	**46**	**57**	**63**	**42**	44	**61**	80	74	34	35	43
HadCM3 volcanic	50	56	**60**	64	**44**	**39**	65	81	74	31	42	39
no change	45	52	61	64	47	41	63	76	67	30	38	38
4 DECADES												
num. of grid boxes	146	149	149	146	144	138	136	34	17	146	146	144
HadCM2 gso	**36**	**34**	**41**	**50**	64	**54**	**78**	**44**	**65**	60	**38**	**44**
HadCM2 gs	**26**	**24**	**34**	**53**	**41**	**50**	90	**82**	94	**48**	**43**	**45**
HadCM2 ghg	73	89	97	98	**47**	**47**	93	**88**	82	62	**41**	**51**
HadCM2 sol	47	**39**	**36**	**54**	**51**	64	94	**85**	82	63	**50**	64
HadCM2 lbb	57	54	60	**68**	65	62	**68**	**85**	88	57	62	65
HadCM2 vol	63	69	66	74	69	58	**85**	**88**	88	53	57	63
no change	42	45	48	70	59	58	87	91	76	53	53	61
HadCM3 anthro	38	**36**	**48**	**61**	**37**	**57**	**61**	**56**	**53**	47	**38**	**39**
HadCM3 trop-anthro	42	42	**53**	**68**	**51**	**48**	92	**91**	88	45	**43**	**42**
HadCM3 ghg	43	48	59	79	**58**	**53**	93	94	88	42	**40**	**45**
HadCM3 natural	41	48	57	77	74	72	**83**	**91**	76	42	49	55
HadCM3 solar	**31**	**36**	**40**	71	**62**	**58**	**82**	**91**	**71**	45	**38**	56
HadCM3 volcanic	55	56	60	79	67	68	**79**	**91**	76	47	53	**51**
no change	37	41	54	73	67	66	88	94	76	42	45	52

Table 3.1. Percentage of individual grid box values greater than 3σ from the observations for each temperature variable. Values highlighted in yellow indicate model skill compared to the null hypothesis field of "no change".

Root Mean Squared Difference values between modelled and observed fields.

Pressure level	850 hpa	700 hpa	500 hpa	300 hpa	200 hpa	150 hpa	100 hpa	50 hpa	30 hpa	500-850	300-500	300-850
3 DECADES												
num. of grid boxes	182	189	192	191	187	185	186	104	81	181	191	181
HadCM2 gso	0.04	0.034	0.034	0.033	0.05	0.057	0.05	0.059	0.059	0.028	0.025	0.043
HadCM2 gs	0.031	0.026	0.029	0.033	0.031	0.044	0.063	0.063	0.062	0.028	0.024	0.044
HadCM2 ghg	0.041	0.032	0.032	0.033	0.038	0.046	0.083	0.078	0.07	0.031	0.025	0.047
HadCM2 sol	0.037	0.031	0.031	0.031	0.032	0.044	0.075	0.084	0.097	0.028	0.023	0.043
HadCM2 lbb	0.05	0.044	0.042	0.037	0.038	0.049	0.065	0.092	0.09	0.031	0.028	0.052
HadCM2 vol	0.044	0.039	0.039	0.036	0.034	0.044	0.073	0.087	0.094	0.027	0.025	0.043
no change	0.047	0.042	0.042	0.04	0.035	0.044	0.058	0.086	0.085	0.028	0.024	0.045
HadCM3 anthro	0.036	0.03	0.032	0.031	0.039	0.053	0.05	0.062	0.069	0.028	0.023	0.04
HadCM3 trop-anthro	0.033	0.028	0.03	0.033	0.036	0.044	0.065	0.08	0.061	0.027	0.024	0.04
HadCM3 ghg	0.034	0.029	0.033	0.039	0.037	0.047	0.073	0.088	0.066	0.032	0.027	0.048
HadCM3 natural	0.048	0.042	0.042	0.038	0.037	0.049	0.062	0.091	0.09	0.031	0.027	0.049
HadCM3 solar	0.047	0.039	0.039	0.036	0.035	0.044	0.059	0.088	0.09	0.03	0.024	0.047
HadCM3 volcanic	0.051	0.046	0.045	0.04	0.035	0.045	0.061	0.089	0.088	0.029	0.026	0.047
no change	0.047	0.042	0.042	0.04	0.035	0.044	0.058	0.086	0.085	0.028	0.024	0.045
4 DECADES												
num. of grid boxes	146	149	149	146	144	138	136	34	17	146	146	144
HadCM2 gso	0.038	0.033	0.038	0.048	0.059	0.072	0.077	0.212	0.291	0.04	0.032	0.059
HadCM2 gs	0.038	0.038	0.047	0.062	0.06	0.081	0.126	0.234	0.29	0.04	0.035	0.062
HadCM2 ghg	0.041	0.043	0.054	0.067	0.056	0.079	0.15	0.258	0.315	0.043	0.033	0.062
HadCM2 sol	0.039	0.032	0.037	0.053	0.056	0.082	0.159	0.304	0.417	0.045	0.035	0.068
HadCM2 lbb	0.047	0.04	0.041	0.056	0.074	0.087	0.109	0.292	0.371	0.04	0.042	0.074
HadCM2 vol	0.052	0.045	0.048	0.058	0.066	0.084	0.109	0.29	0.341	0.038	0.039	0.068
no change	0.047	0.04	0.042	0.055	0.064	0.083	0.114	0.296	0.374	0.04	0.038	0.07
HadCM3 anthro	0.043	0.042	0.052	0.061	0.055	0.079	0.077	0.169	0.196	0.04	0.032	0.058
HadCM3 trop-anthro	0.048	0.047	0.057	0.069	0.062	0.075	0.126	0.27	0.246	0.04	0.035	0.062
HadCM3 ghg	0.051	0.052	0.064	0.078	0.068	0.084	0.139	0.294	0.269	0.042	0.036	0.066
HadCM3 natural	0.051	0.044	0.046	0.058	0.071	0.092	0.112	0.291	0.381	0.041	0.034	0.066
HadCM3 solar	0.04	0.035	0.039	0.056	0.064	0.084	0.114	0.286	0.377	0.039	0.034	0.066
HadCM3 volcanic	0.061	0.053	0.053	0.062	0.072	0.088	0.111	0.286	0.371	0.043	0.041	0.075
no change	0.047	0.04	0.042	0.055	0.064	0.083	0.114	0.296	0.374	0.04	0.038	0.07

Table 3.2. RMSD values between modelled and observed fields for each temperature variable. Values highlighted in yellow indicate model skill compared to the null hypothesis of “no change”.

Table of best guess skill scores and their uncertainty ranges for HadCM2 and HadCM3 fields for three decade and four decade diagnostics.

Ensemble	Three decades levels	three decades lapse	Four decades levels	Four decades lapse
HadCM2 GSO	4 (3-5)	0 (0-2)	5 (3-7)	2 (1-2)
HadCM2 GS	5 (3-6)	0 (0-2)	4 (3-6)	2 (2-3)
HadCM2 GHG	4 (3-5)	1 (0-1)	1 (1-2)	0 (0-1)
HadCM2 SOL	1 (1-6)	2 (0-3)	4 (2-6)	0 (0-1)
HadCM2 LBB	1 (0-4)	0 (0-1)	1 (0-2)	0 (0-1)
HadCM2 VOL	2 (1-5)	1 (0-2)	0 (0)	0 (0-2)
HadCM3 ANTHRO	4 (3-6)	0 (0-2)	2 (2-4)	2 (1-3)
HadCM3 TROP-ANTHRO	5 (3-7)	1 (0-3)	2 (1-3)	1 (0-2)
HadCM3 GHG	4 (2-6)	0 (0-1)	0 (0-2)	1 (0-2)
HadCM3 SOLAR	2 (1-4)	0 (0-1)	4 (1-6)	2 (1-3)
HadCM3 VOLCANIC	1 (0-4)	0 (0-2)	2 (1-3)	0 (0-1)
HadCM3 NATURAL	1 (0-3)	0 (0-1)	2 (0-2)	1 (0-2)
No change	0 (0-5)	0 (0-2)	0 (0-4)	0 (0-3)

Table 3.3 Overall skill scores for the model fields when compared to a null hypothesis of zero change. For any level or lapse rate to gain a score it must exhibit skill both at the grid-box level and in the RMSD statistic. The maximum possible score is, therefore, 12. The fields are split so as to consider skill scores and their uncertainty ranges for temperatures on pressure levels and lapse rate diagnostics separately. Model fields considering anthropogenic influences consistently provide the best explanation, although there are also likely solar, and possibly volcanic influences, and "no change" cannot be ruled out. The choice of decadally averaged data is probably sub-optimal when searching for natural external forcing signals.

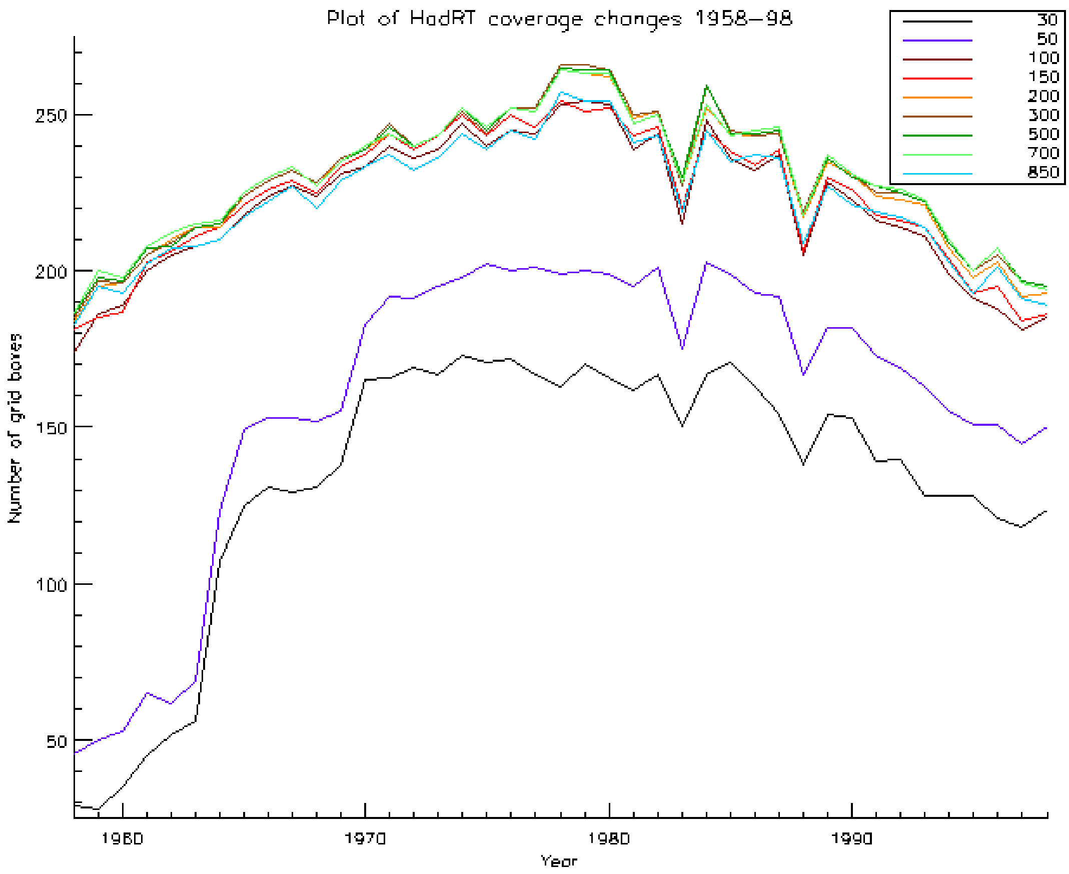

Figure 3.1. Changes in areal grid box coverage of the HadRT dataset over the period 1958-1998. The key in the top right hand corner denotes the pressure levels. Note how, below an altitude of 50hPa, the total number of observations is generally consistent between the levels. Decreases in data availability occur both early in the record due to a lack of observations, and late in the record due to delays in release of the data.

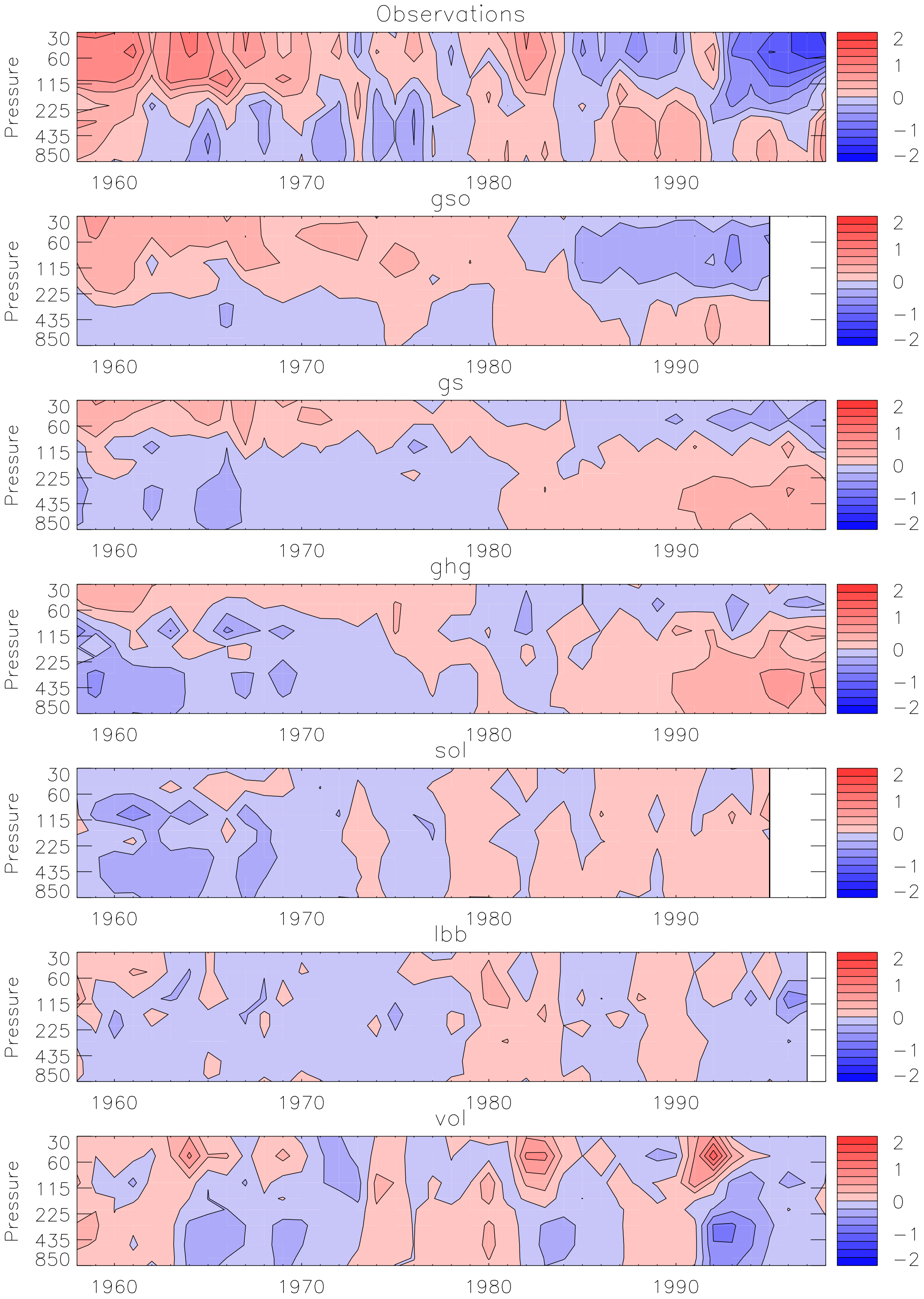

Figure 3.2a. Global averages for the HadRT2.1s temperatures and HadCM2 ensemble mean fields. Note that prior expectations are that model fields will exhibit less variability.

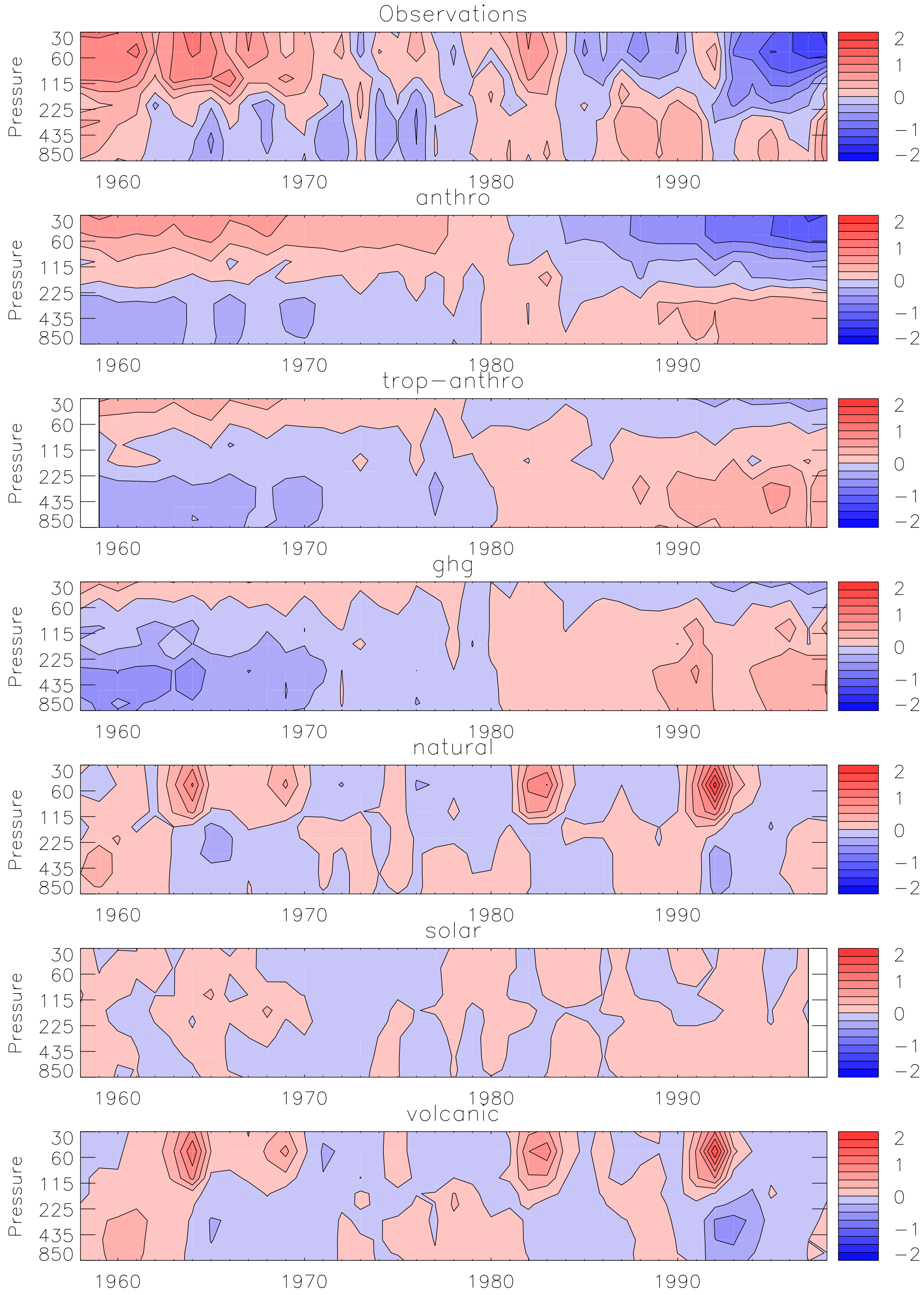

Figure 3.2b. Global averages for the HadRT2.1s temperatures and HadCM3 ensemble mean fields. Note that prior expectations are that model fields will exhibit less variability.

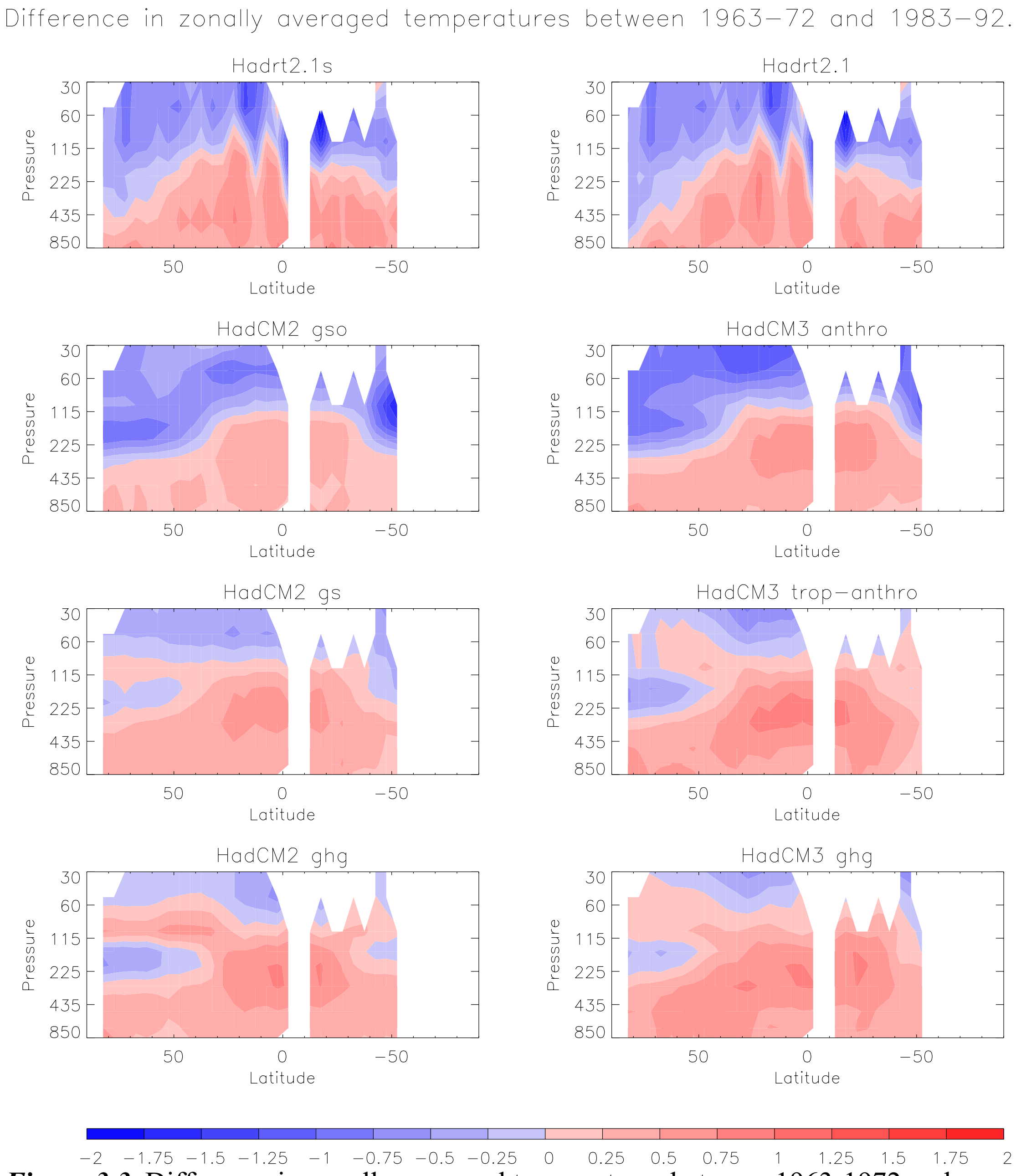

Figure 3.3. Difference in zonally averaged temperatures between 1963-1972 and 1983-1992. The top two panels show HadRT2.1s and HadRT2.1 fields, the differences between these datasets being relatively small. Remaining panels indicate ensemble mean model responses to various anthropogenic forcing scenarios for HadCM2 (left hand panels) and HadCM3 (right hand panels).

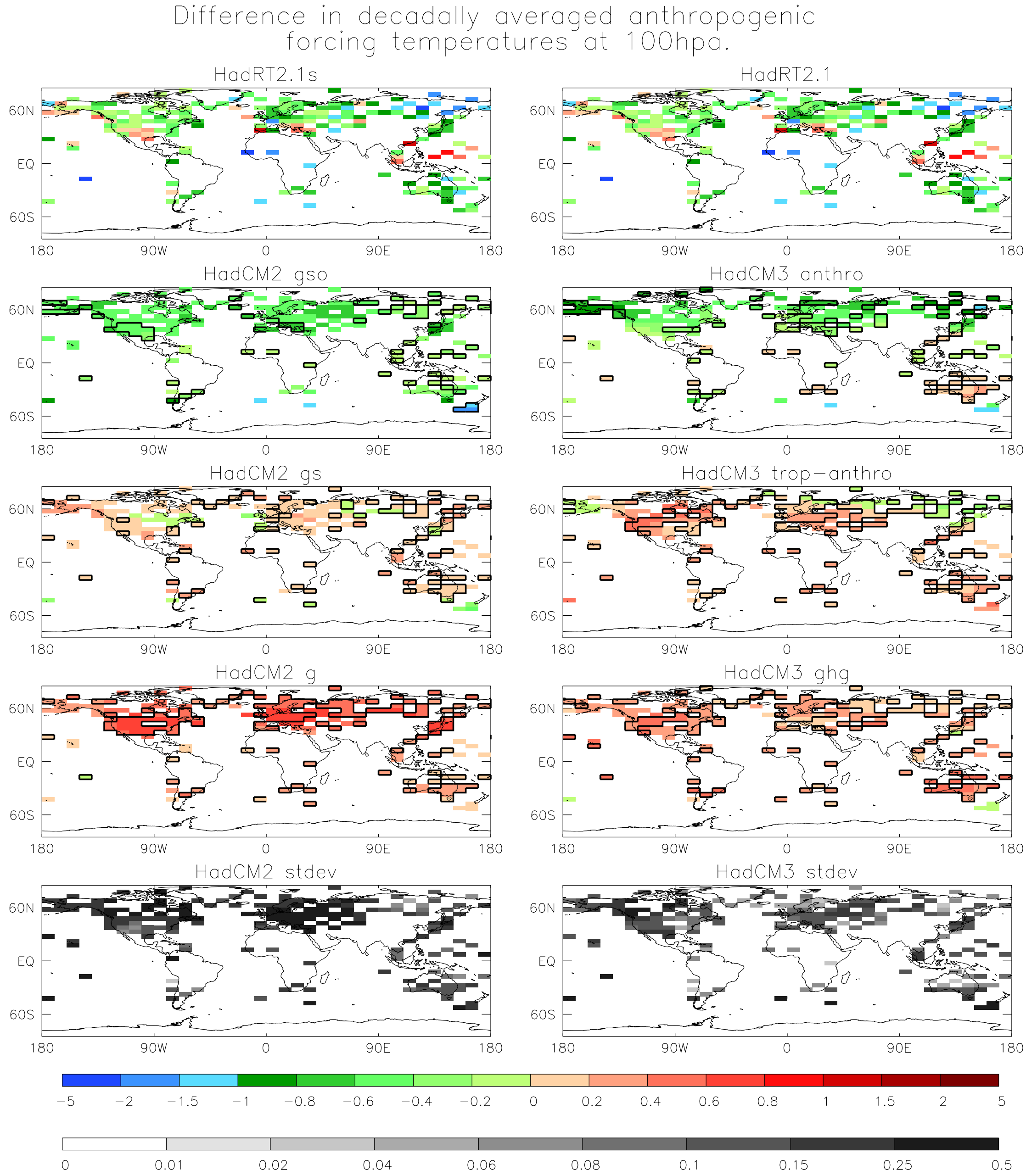

Figure 3.4. Difference in temperatures at the 100hPa level between 1963-1972 and 1983-1992. The top two panels show HadRT2.1s and HadRT2.1 fields, the differences between these datasets being relatively small. Remaining panels indicate ensemble mean model responses to various anthropogenic forcing scenarios, and the respective model σ fields for HadCM2 (left hand panels) and HadCM3 (right hand panels). In these model fields those points outside 3σ from the observations are boxed.

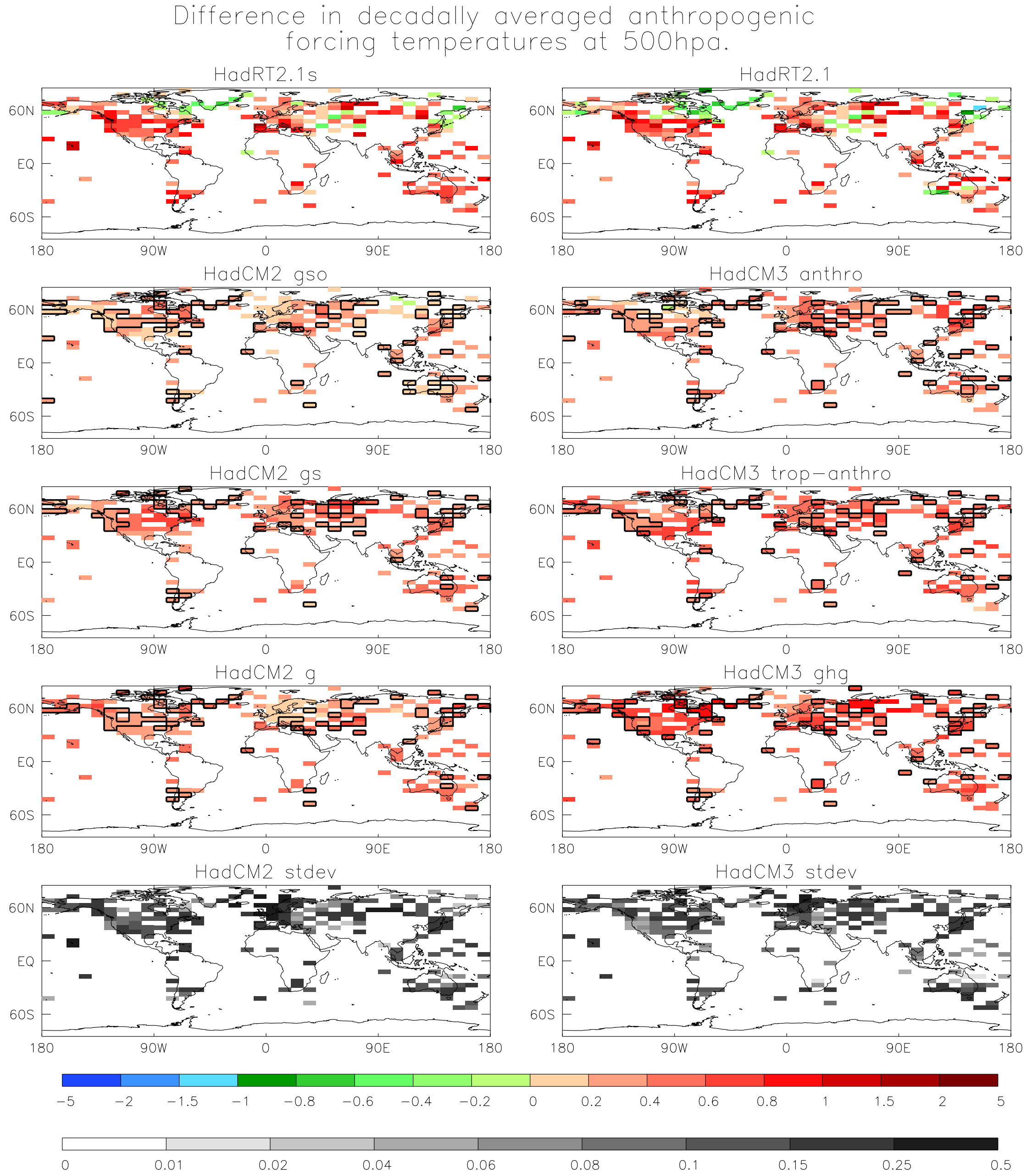

Figure 3.5. Difference in temperatures at the 500hPa level between 1963-1972 and 1983-1992. The top two panels show HadRT2.1s and HadRT2.1 fields, the differences between these datasets being relatively small. Remaining panels indicate ensemble mean model responses to various anthropogenic forcing scenarios, and the respective model σ fields for HadCM2 (left hand panels) and HadCM3 (right hand panels). In these model fields those points outside 3σ from the observations are boxed.

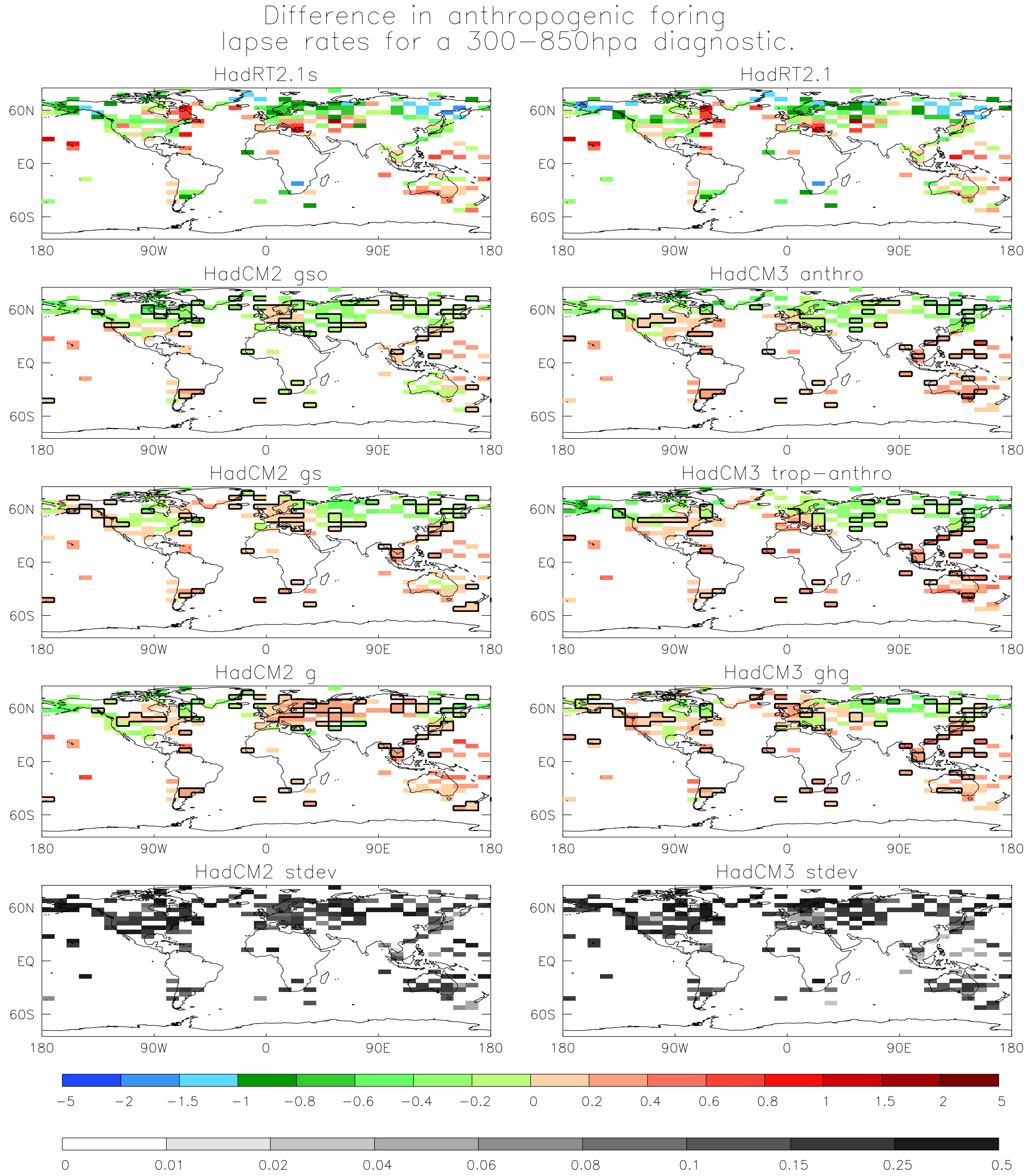

Figure 3.6. Difference in an entire troposphere (300-850hPa) lapse rate diagnostic between 1963-1972 and 1983-1992. The top two panels show HadRT2.1s and HadRT2.1 fields. Remaining panels indicate ensemble mean model responses to various anthropogenic forcing scenarios, and the respective model σ fields for HadCM2 (left hand panels) and HadCM3 (right hand panels). In these model fields those points outside 3σ from the observations are boxed.

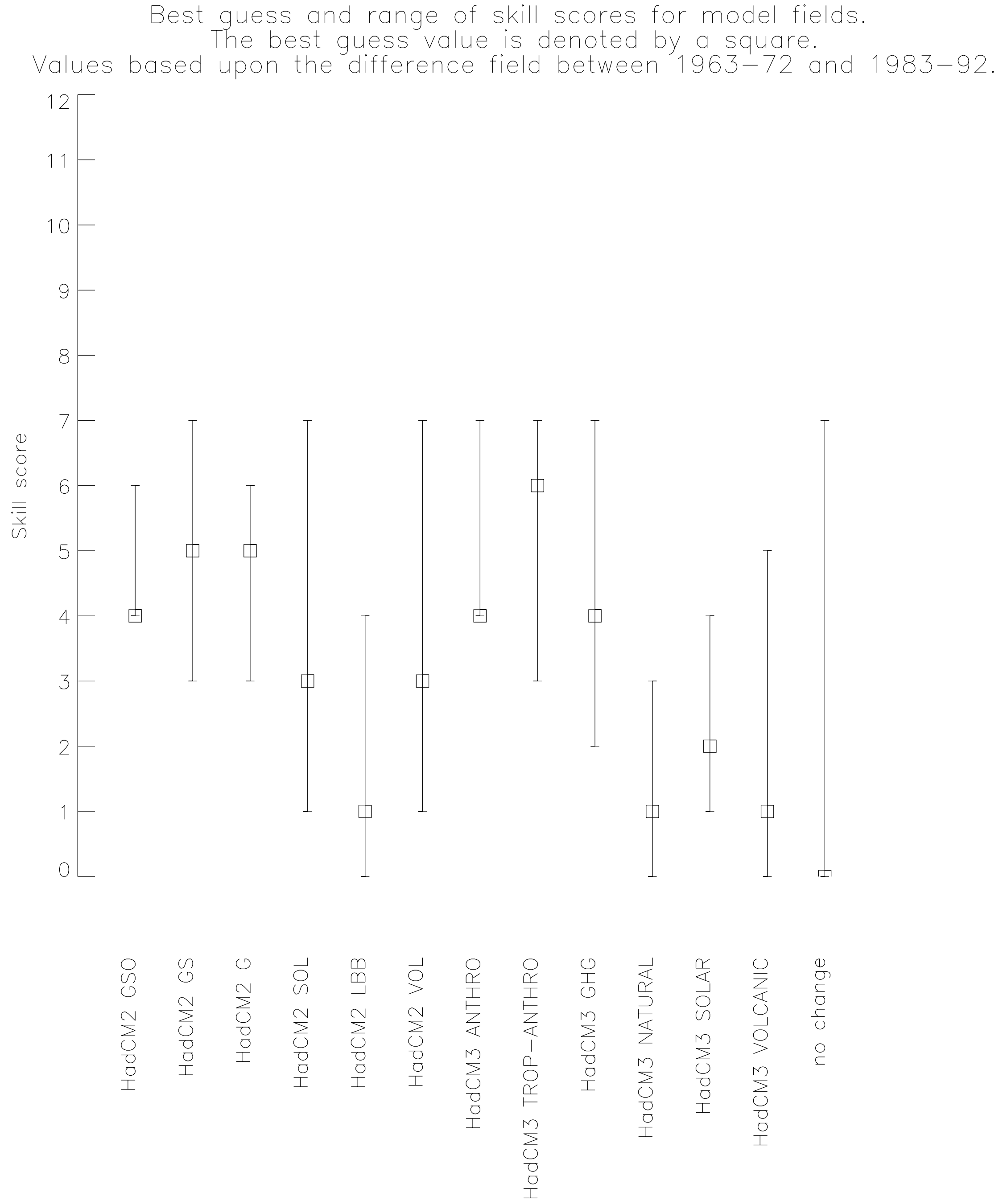

Figure 3.7. Best estimate and range of skill scores for a three-decade diagnostic. The best estimate is denoted by a square. There is no reason to expect the uncertainty range to be symmetric about this estimate. An ensemble is concluded to exhibit skill if its uncertainty range does not encompass zero. Confidence is increased with greater separation. In this approach a signal would be detectable if its best estimate was outside the uncertainty range of the null hypothesis no change field.

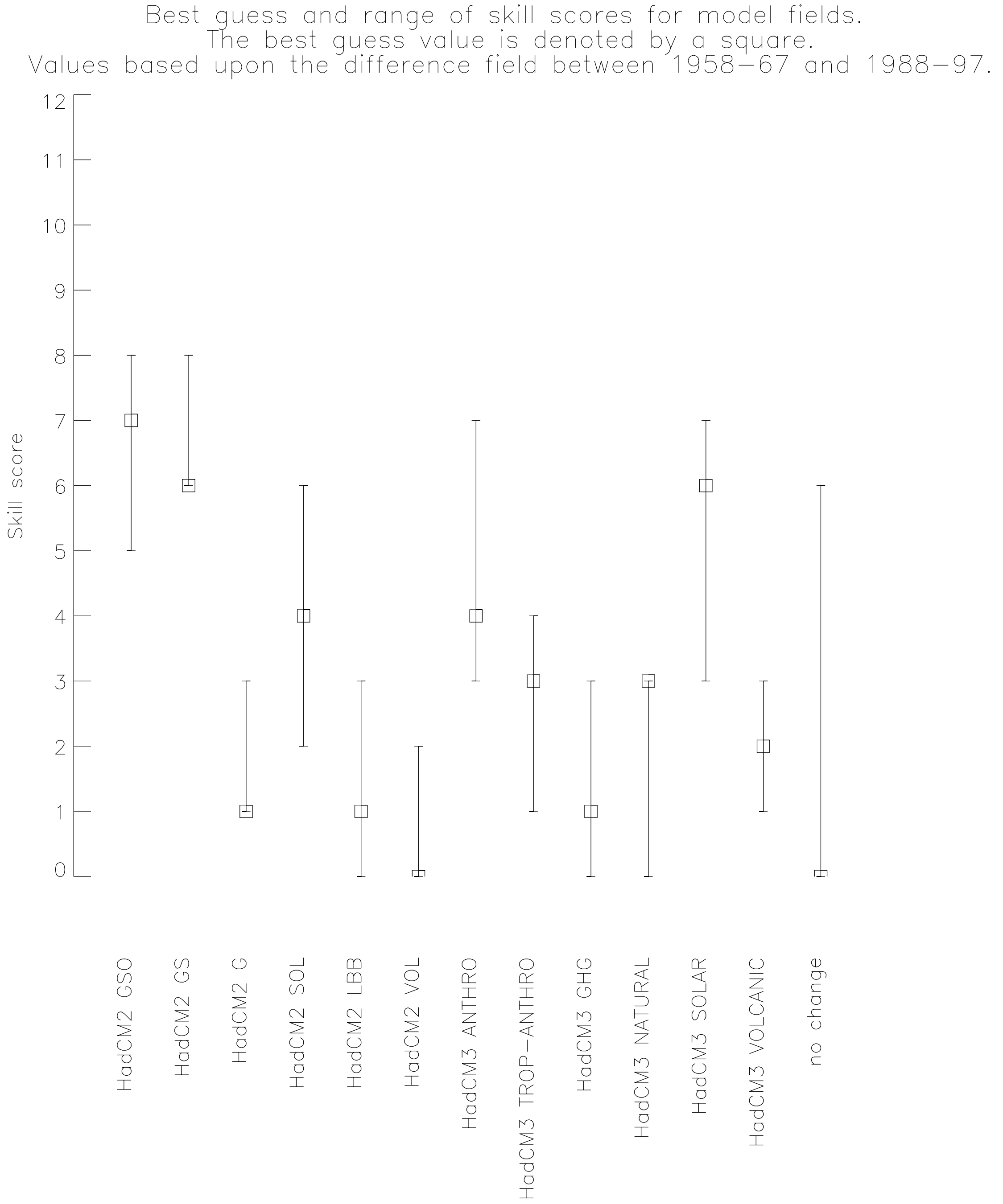

Figure 3.8. Best estimate and range of skill scores for a four-decade diagnostic. The best estimate is denoted by a square. There is no reason to expect the uncertainty range to be symmetric about this estimate. An ensemble is concluded to exhibit skill if its uncertainty range does not encompass zero. Confidence is increased with greater separation. In this approach a signal would be detectable if its best estimate was outside the uncertainty range of the null hypothesis no change field.