6. The most likely causes of late 20th Century observed temperature changes at the near-surface and within the free troposphere

Changes in upper air and near-surface temperatures for the period 1960-1995 are considered in an optimal regression detection study under a common methodological framework to identify the most likely causes of recent climate change. Firstly, observed and modelled fields are pre-processed. Six input tropospheric temperature variables are defined: 3 layer average temperatures, and 3 lapse rate temperatures. Detection results using the OLS regression approach (AT99, chapter 4) are subsequently considered for all plausible model signals, identifying the most likely signal combinations which explain recently observed changes for each input temperature variable. Initial detection results from a TLS regression approach (AS01, chapter 4) are also considered to assess whether they are likely to be highly sensitive to the choice of regression algorithm. Detailed results from OLS regression for the most plausible combinations of forcings are then considered for each input temperature variable to enable meaningful statements to be made regarding the most likely causes of recently observed changes. Finally, the lines of evidence are then drawn together from the separate temperature variables to begin to assess whether HadCM2 and HadCM3 are plausibly adequate explanations of the observations throughout the troposphere and, if so, which forcings are most likely to be important in explaining the observations.

6.1 Observed and modelled field pre-processing

Globally gridded, although incomplete, observed near-surface (HadCRUTv, Jones et al., 2001) and upper-air (HadRT, Parker et al., 1997) temperature datasets exist for a common period from 1958 to present. In this study, a variance-corrected version (HadCRUTv) of the near-surface HadCRUT temperature series is considered. The non-variance corrected version was used in previous detection studies (Tett et al., 1999, 2001 for example). Two versions of the HadRT dataset are also considered: HadRT2.1 which has had corrections applied by Parker et al. (1997) for suspected inhomogeneities throughout its depth, with reference to MSUc series temperatures (Christy et al., 1998); and HadRT2.1s which has had these corrections applied only

within the stratosphere. Both the HadRT versions considered here have had obviously dubious values removed in well-sampled regions, following the analysis described in chapter 2.

The HadCRUTv and HadRT datasets exhibit time-varying coverage changes over the period under consideration. The HadRT temperature dataset has greatly reduced coverage compared to the HadCRUTv dataset, and is highly biased towards Northern Hemisphere mid-latitude continental regions (see chapters 2 and 3). To avoid systematic bias in any comparisons undertaken between near-surface and upper-air temperatures, the HadCRUTv temperature record is re-interpolated to the coarser resolution HadRT grid (from 5°×5° to 10°×5°). It is subsequently sub-sampled to the time-varying coverage of the lower tropospheric HadRT temperatures (Santer et al., 1999, provide a detailed justification for such an approach). Given that previous studies (Tett et al., 1999, 2001, Stott et al., 2001, 2001a) have considered the sensitivity of detection results based upon near-surface temperatures to a plethora of potential uncertainties, this solely provides a further sensitivity study, as to whether meaningful statements regarding signal detection and / or consistency can still be made given artificially degraded data coverage.

The HadRT record exists for nine WMO standard reporting pressure levels from the lower troposphere (850 hPa) up into the lower stratosphere (30 hPa). Tropospheric upper-air temperatures are further processed here to reduce the dimensionality of the problem being considered. Stratospheric levels are removed, as they are effectively de-coupled from the troposphere across the tropopause. Furthermore, the coverage, and the confidence in the accuracy of, the HadRT stratospheric records is reduced compared to the tropospheric records (chapter 2, Gaffen et al., 2000a, Parker and Cox, 1995). There are also large uncertainties regarding the abilities of both GCMs considered here within the stratosphere (Gillett et al., 2000, Collins et al., 2001). A mass weighting scheme, which was found in chapter 5 to yield the most consistent detection results, is employed to attain two composite tropospheric layer averages: an upper tropospheric (UT) (500 hPa and 300 hPa), and a lower tropospheric (LT) (850 hPa and 700 hPa) series. Near-surface temperatures (Surf) are treated as a further layer in this analysis to yield three separate layer average temperature series

estimates within the well-mixed troposphere. Differences between these tropospheric layers can be calculated to yield three crude tropospheric lapse rate series: UT-LT (free troposphere lapse rate), UT-Surf (entire troposphere lapse rate), and LT-Surf (lower troposphere lapse rate). In each case the lapse rate is simply calculated as the upper layer value minus the lower layer value.

The resulting observed datasets contain large swathes of missing data, primarily over the oceans (see Figure 6.1). Hence it is highly unlikely to be suitable to undertake the spherical harmonic coefficients approach previously employed with near-surface temperature datasets in space-time optimal regression detection and attribution studies to perform spatial truncation (Allen et al., 2000, 2001, Tett et al., 1999, 2001, Stott et al., 2001, 2001a). However, it should be noted that all that is required in optimal regression detection studies is a representation of the spatio-temporally evolving patterns of any atmospheric parameter at the large spatial scales at which there is confidence in model skill (Stott and Tett, 1998, AT99). Therefore, it is proposed here that a number of areas be defined from which to calculate (latitudinally weighted) Large Area Averages (henceforth LAAs). The rationale behind such an approach is given in further detail in section 3.3.

Four sets of LAAs are defined for the purposes of the current study to assess the sensitivity of results to the choice of regions. Two "smart" LAA diagnostics, and two less smart or "naïve" LAA diagnostics are considered. "Smart" diagnostics take into account geographic coverage of the observations, in an attempt to derive LAA values based upon approximately equal numbers of contributing grid boxes, and for specific regions. Figure 6.1 shows the raw input lower tropospheric HadRT2.1s field, upon which are superimposed the area boxes used in the 10-area "smart" LAA diagnostic. Those datapoints not within area boxes are generally from data-sparse regions. Confidence in their upper air temperature records is reduced, as rigorous spatial quality control has not been possible on these data (chapter 2). Figure 6.2 shows how the raw observed field of Figure 6.1 translates into 10-area "smart" LAA values. This diagnostic is highly biased towards Northern Hemispheric continental mid-latitude regions, with no information arising from high latitudes of the Southern Hemisphere. A coarser resolution "smart" LAA approach is also employed using 5 areas, and this is shown in Figure 6.3. Both "naïve" LAAs assume no prior knowledge of observed

dataset coverage, except that it is insufficient poleward of 30°S. The remaining field is simply split into LAAs of equal area. This is achieved by splitting into three zonal bands, 90°N-30°N, 30°N-0°N, and 0°S to 30°S, as per G. S. Jones et al. (2001). These three bands are then further divided into two and four longitudinal bands of equal spacing to yield a 6-area and a 12-area "naïve" LAA diagnostic respectively. Expectations are that these diagnostics are likely to be highly sub-optimal, as both variance and sampling error are likely to vary widely between the component areas. Furthermore, for temperature variables including a free troposphere component, these "naïve" LAAs will incorporate HadRT data in which there is less confidence as near-neighbour consistency checks described in chapter 2 could not be applied in more data sparse regions. These "naïve" LAA diagnostics are, therefore, solely used in further consideration of those combinations of signals determined in detection analysis by "smart" LAAs to be the most likely explanation of recent trends in tropospheric temperature variables.

It should be noted that the choice of LAAs to be considered could only ever lead to a sub-optimal input to the detection algorithm and, therefore, to conservative results. In the current study LAAs are defined as rectangles to minimise the computational costs involved, although theoretically LAAs could be any shape, and even overlap. Considering overlapping regions may introduce complications into the analysis of any results, and so is avoided here. Scope remains for a more optimal approach. Care must be taken, however, in ensuring against introducing a common factor effect (see for example Santer et al., 1993) into any procedure that is used in optimising the choice of areas in any studies. This could systematically bias the results of any subsequent detection exercise, and lead to false positive detection results.

A number of complex GCM ensemble predictions of the response of tropospheric and near-surface temperatures to a range of physically plausible external forcing factors, both natural and anthropogenic, are considered for the same period as the observations. Models are also used to derive estimates of natural internal climate variability. Two model versions of the Hadley Centre's GCM are employed in this study: HadCM2 (Mitchell et al., 1995, Johns et al., 1997) and HadCM3 (Pope et al., 2000, Gordon et al., 2000). These are summarised in chapter 1. HadCM2 and

HadCM3 data are bilinearly interpolated (Press et al., 1992) to the coarser resolution HadRT grid (from 3.75°×2.5° to 10°×5°). They are subsequently treated in exactly the same way as the observational data to allow direct comparisons to be made. In particular, the same annual missing-data mask is employed for the annual grid-box series in an attempt to incorporate the effects of incomplete temporal sampling in the observations. These masks are not quite exactly coincident, as the observations are available on a finer monthly resolution. This should not be a major constraint, as sampling error is likely to be much smaller in the model than the observations. Only the preferred version of V2 (chapter 2), which requires two months in each of three seasons in any year as a minimum condition for calculation of an annual mean value (see chapter 2), is used for each observed dataset. Results for a sub-set of cases (not shown here) were found to be equivalent to first order regardless of which of the three versions considered in chapter 2 is employed. Hence, it is not believed that principal results are overly sensitive to the temporal inclusion criteria used in calculating individual observed grid-box annual averages over a reasonable range.

Five-year LAA averages of annual mean grid-box data are calculated for all modelled and observed fields for the period 1960-1995, and stacked in chronological order to yield a space-time input vector to the regression algorithm. Figures 6.2 and 6.3 provide a spatial realisation of observed input fields using HadRT2.1s for the LT level for 10-area and 5-area "smart" LAA diagnostics respectively. The preferred "smart" input diagnostic in this study is the 10-area LAA diagnostic (Figure 6.2). Use of more data points as input to the regression should reduce degeneracy between input signals, as well as increasing the power of the algorithm to discriminate between competing hypotheses of the causes of recent climate change. This effect, albeit in time and not space, has previously been illustrated for near-surface temperatures by increasing the temporal resolution of the decadally averaged input diagnostics from annual to seasonal (Stott et al., 2001). If results can also be replicated in the 5-area "smart", and both "naïve" LAA diagnostics, then confidence will be increased, as they are not susceptible to spatial pre-processing sampling uncertainty.

Model signals considered in the present study are detailed in Table 6.1. There are some differences between equivalent model anthropogenic forcings (see also section 1.3). For the sake of clarity in the current analysis all HadCM3 model fields are aliased to their HadCM2 equivalents. In HadCM2 G considers well-mixed greenhouse-gases as a CO_2 equivalent concentration, whereas for HadCM3 the gases are treated independently using an interactive chemistry model. For HadCM2 GS, also considers the direct effect of sulphate aerosols. HadCM3, in its GS equivalent run additionally considers the direct and the primary indirect effect of sulphate aerosols, as well as changes in tropospheric ozone concentration. In both models GSO considers effects of stratospheric ozone depletion on top of those forcings considered in GS. In addition to anthropogenic ensembles, both models have signal response ensembles for both solar and volcanic natural external forcings. Here, consideration is solely given to the two forcings that are directly equivalent between the models being considered. These are a reconstruction of solar irradiance (Lean et al., 1995), labelled LBB, and a dust veil index reconstruction based upon volcanic activity (Sato et al., 1993), labelled VOL. Other natural forcing realisation ensembles exist for both models, and future studies might aim to consider these in addition to those analysed here, as well as responses to alternative reconstructions of natural forcings (Bard et al., 2000, Fligge and Solanki, 2000, Stothers, 1996 for example).

6.2 Upper tropospheric temperatures

It is sensible to first focus any detailed analysis for each input temperature variable upon those signal combinations that are the most likely causes of recently observed temperature changes. There exist over 30 possible input signal combinations for each model. This is reduced by considering only up to three-signal combinations, as four- and five-signal combinations are always found to be potentially degenerate, for all six input temperature variables being considered, for both models, as are three-signal combinations in some cases. Results of standard detection tests on upper tropospheric temperatures are given in Table 6.2 for all conceivable 1-, 2-, and 3-input signal combinations for each model, and for each version of the HadRT dataset. Two specific questions are addressed in this table: firstly whether the signal(s) is (are) significantly positive [detection], and if it is (they are), then secondly, whether

it is (they are) plausibly a consistent explanation of the observations [attribution]. In both cases assessments are made at the 90% confidence limit, in keeping with previous detection studies (AT99, Tett et al., 1999, 2001 for example). In previous studies (Tett et al., 1999, 2001 amongst others) the most parsimonious explanation that passes these two tests has been sought, and observed changes attributed to this combination of forcings. Results are shown for the OLS approach for both 10-area and 5-area "smart" LAA diagnostics, and for the TLS approach for 10-area "smart" LAA diagnostics. TLS provides an estimate as to the likely sensitivity of principal detection results to the most likely major sources of uncertainty. Detection results are quoted at a truncation of 21 (the maximum estimated degrees of freedom of the control segment used for optimisation in each model), or the maximum truncation which does not fail the consistency test on the residuals. Failure of the test on residuals implies that the regression is becoming unrealistic, and acts as a check on the reality of the results (AT99).

Detection results for HadCM2 signals in Table 6.2 indicate strong evidence for a detectable anthropogenic influence upon recently observed upper tropospheric temperatures. This result is largely independent of which HadRT version, LAA diagnostic, or regression algorithm is being considered. Well-mixed greenhouse-gases are always detected for HadCM2 under the OLS regression approach, either individually or when input in a fixed-ratio combination with other anthropogenic forcings. Sulphate aerosol forcing influences are also occasionally detected on their own for HadCM2: more so when HadRT2.1s is the observed dataset than HadRT2.1. A stratospheric ozone depletion forcing is only ever detected in HadCM2 fields when combined with other anthropogenic forcings. Confidence in a detectable stratospheric ozone depletion influence on upper tropospheric temperatures is therefore low for HadCM2. Results using the TLS approach are slightly more ambiguous for HadCM2 than those under OLS, particularly so for three-signal input combinations, which standard tests (Tett et al., 1999, Mardia et al., 1979) suggest are potentially degenerate. However, the primary conclusion that remains is a demonstrable anthropogenic influence under TLS. There is, additionally, some evidence under a TLS approach for a detectable HadCM2-predicted volcanic influence, although this is not repeated under the OLS approach, reducing confidence in this result. For neither regression approach, is there evidence for a HadCM2-

derived solar influence upon upper tropospheric temperatures. For the 5-area "smart" LAA diagnostic and HadCM2 fields, OLS results considering HadRT2.1 systematically fail at low truncations, whereas they do not for HadRT2.1s. Importantly, there is no systematic bias in what is detected. This result is dependent upon the input pre-processing, as the 10-area "smart" LAA diagnostic results for HadCM2 considering HadRT2.1s and HadRT2.1 do not exhibit similar systematic behaviour, thus reducing confidence in a discernible systematic influence of observational uncertainty upon results.

Detection results for HadCM3 upper tropospheric temperature model fields in Table 6.2 fail at early truncations in most cases, especially under the OLS approach. For HadCM3, as for HadCM2, well-mixed greenhouse-gases are robustly detected, and sulphate aerosols slightly less so. In addition there is stronger evidence for a HadCM3 predicted volcanic influence in the observations, with detection for OLS as well as TLS regression approaches. There are a number of cases where a solar influence is detected for HadCM3, either on its own, or when considered in combination with volcanic influences. Further analysis shows that these detections occur at very small truncations, residuals being inconsistent at higher truncations and, therefore, confidence in a solar influence is greatly reduced. Also, solar influences are never detected for HadCM3 when considered in combination with anthropogenic forcings, further reducing confidence in a discernible solar influence.

Subsequent analysis concentrates upon the GS + VOL, G + S (whereby the component forcings are allowed to have a ratio other than that forced in GS), and G + S + VOL combinations, being the most likely explanations of the observations according to both models in the analysis of Table 6.2. Recourse is made solely to results from OLS regression for this analysis. Results for TLS regression in Table 6.2 indicate that fundamental discrepancies should not exist between the two regression approaches for either model. AS01 show how OLS estimates may be negatively biased in comparison to those for TLS, especially in their upper bounds and for weak or poorly defined signals. It is therefore important to additionally consider SNRs to ensure that OLS results are unlikely to be significantly negatively biased. In addition to detection results for those "smart" LAA diagnostics considered in Table 6.2, detection results from "naïve" LAA diagnostics are also considered. Global mean

temperature reconstructions are additionally analysed for the 10-area "smart" LAA diagnostic to assess whether the models truly capture the observed transient upper tropospheric temperature changes, at least on a global scale.

Detection traces for HadCM2 upper tropospheric temperature fields considering HadRT2.1s observations are shown in Figure 6.4 for all three input signal combinations identified previously, and all four LAA diagnostics. Detection traces illustrate the changing results with increasing truncation (number of principal modes being considered in the regression). It is important to consider traces rather than results at single truncations to ensure against making ambiguous statements as to the causes of recent climate change. Uncertainty ranges are inflated to account for finite ensemble size when considering consistency, but not for detection (Stott et al., 2001). A signal is detected at a given confidence interval if its lower detection limit is positive, and consistent if the inflated uncertainty limits encompass unity at any given truncation. Confidence is increased if results are insensitive to choice of both truncation and input LAA diagnostic. More weighting is given to "smart" LAA diagnostics in this and subsequent sections as expectations are that these are less likely to be biased than "naïve" LAA diagnostics for reasons discussed in section 6.1. If the uncertainty limits of the amplitude estimate distribution are significantly greater than unity, then the model signal must be multiplied by a scaling factor greater than 1 and, therefore, the model is significantly underestimating the response. The opposite is true for those cases where the limits are less than unity. Confidence is low in results for those truncations where the test on the residuals fails, as residuals of the regression do not resemble an estimate of internal climate variability from an independent section of model control.

In the GS + VOL signal combination for HadCM2 in Figure 6.4, GS is both robustly detected and a consistent explanation of the observations, at all truncations and for all LAA diagnostics, even in those cases where residuals from the OLS regression are inconsistent. Detection of a volcanic signal in this HadCM2 signal input combination is far less certain. In fact detection of VOL is only achieved in the 5-area "smart" LAA diagnostic, and then only marginally so at a few truncations. However, OLS best-guess amplitude estimates are overwhelmingly positive, implying that there exists some predictive skill in the volcanic signal estimate in

HadCM2. Under an OLS approach, the uncertainty range is conditioned independently of the observations, but the best-guess estimate is that resulting from the regression, which depends upon the observations. If for any signal at any truncation the best-guess estimate is negative, this implies that the inverse of the signal is required in the regression to recreate the observations. Physically this cannot be correct, so cases where the best-guess amplitude estimators are negative greatly reduce confidence in the presence of the model-derived signal even where the uncertainty range in the statistic still includes positive values. In the majority of cases HadCM2 significantly overestimates the true volcanic forcing upper tropospheric temperature response in the observations. SNR values from Tables 6.8 and 6.9 show that the VOL signal is well defined, and so this discrepancy is likely to be real.

The GS and VOL amplitude estimators for HadCM2 shown in Figure 6.4 tend to exhibit synchronicity in trends with changing truncation for all four LAA diagnostics. This is confirmed in Figure 6.5, which shows the solution ellipses for the three layer average temperature diagnostics at a truncation of 21. The upper tropospheric analysis yields an ellipse whose principal axis is close to the diagonal, implying a high degree of covariance in the solutions (AT99). The GS signal is one of general global-scale warming throughout, whilst VOL is one of shorter-term global-scale cooling events, for HadCM2 within the upper troposphere (see Figure 3.1a). To reproduce the observations they will co-vary – if scaling on one increases, scaling on the other will also be likely to increase to maintain the same overall global-mean temperature change. This is an over-simplistic view, as consideration is actually being given to the full spatio-temporal (rather than global-mean) evolution of the climate system in the OLS algorithm. It is, however, likely to explain a large part of the observed synchronicity between OLS regression amplitude estimators with changing truncation.

Splitting the HadCM2 GS signal into its constituent parts, G + S, G is both robustly detected and consistent for all four input LAAs considered (Figure 6.4), in common with GS in the GS + VOL combination. Results for S are more equivocal, with detection occurring solely at higher truncations for all four LAA input diagnostics. In all LAA diagnostics, and at all truncations, the HadCM2 S signal is a consistent explanation of the observations. S yields more variable amplitude estimates with

changing truncation, than any of the other signals considered, particularly at low truncations. This latter behaviour is most likely due to the low SNR of the HadCM2 S signal, particularly at lower truncations and for "smart" LAAs (Tables 6.8 and 6.9). TLS solutions should ideally be sought at these truncations to avoid biased solutions (Simon Tett, personal communication, 2001). The S signal response is likely to be more spatially heterogeneous, projecting proportionately more onto the higher (generally regional rather than global) modes of atmospheric variability due to the relatively short atmospheric residence time of sulphate aerosols. There is some evidence for this in the higher SNR values at truncation 21 (Table 6.9) compared to truncation 11 (Table 6.8). In common with the GS + VOL signal analysis, there is a degree of synchronicity in the estimators but with S changing by a consistently greater proportion than G, possibly as a result of its relatively weak signal strength.

Considering the three-way regression for HadCM2 of G + S + VOL, results are very similar to those discussed above for these respective signals in the two-way regressions (see Figure 6.4). This is encouraging, as if there were large differences then confidence in the OLS estimator would be greatly reduced, unless these differences could be explained in terms of signal degeneracy. Therefore, in this three-way case for HadCM2, G is robustly detected and consistent, S is sometimes detected but always consistent, and VOL is hardly ever detected and likely to be significantly overestimated in amplitude within HadCM2. All three estimators exhibit synchronicity in their estimates, in agreement with previous analyses of results for two-signal combinations.

It is noted that results for "naïve" LAA diagnostics (the final two columns of Figure 6.4) in this analysis of HadCM2, yield a far greater number of inconsistent residuals for all input signal combinations, providing some justification for the selection of "smart" LAA input diagnostics. In choosing "smart" diagnostics, those areas in which there is confidence in observational data consistency (chapter 2) were concentrated upon, effectively being given higher weighting in the analysis. Conversely for the "naïve" approach all points are given equal weighting, and non-negligible observational upper tropospheric temperature errors from data-sparse regions may have been incorporated. Additionally, sampling error should be approximately equal between areas for "smart" diagnostics, whilst being much more

variable in "naïve" diagnostics. This behaviour of results from the consistency test on the residuals is generally independent of both the model and input layer average temperature variable being considered (see Figures in this and subsequent sections).

Detection traces are first-order equivalent for HadCM2 upper tropospheric temperatures if HadRT2.1 is used in place of HadRT2.1s as the observational dataset (results not shown here for brevity). There is, however, an increased occurrence of consistency test failure for HadCM2 when HadRT2.1 is considered instead of HadRT2.1s for all LAA diagnostics. Corrections between these HadRT versions have been made solely in time on a grid-box basis, and not in space (Parker et al., 1997), with reference to MSUc series temperatures. The applied corrections consist of a shift in the means of individual grid-box series to create a "homogeneous" series. This may well have increased the temporal homogeneity of individual records, whilst degrading the spatio-temporally evolving structure of the observed field as a whole. As consideration is being given here to spatio-temporal patterns of change, the most likely explanation for the observed residuals' test behaviour is that at least some of the corrections introduced are spatio-temporally unlike anything seen in the leading modes of HadCM2. Corrections would therefore tend to inflate the residuals, leading to more frequent consistency test failure. This result is repeated for HadCM3 upper tropospheric temperature fields, so is not model-dependent, giving increased confidence in the hypothesis that at least some of the corrections applied to HadRT2.1 are degrading the spatio-temporal structure of the upper tropospheric temperature observations.

Traces for HadCM3 signal combinations considering HadRT2.1s observations are shown in Figure 6.6. The most immediately obvious difference when compared to the previous analysis for HadCM2 (Figure 6.4), is that there are many fewer cases for HadCM3 where residuals of the regression are consistent, confirming the findings in Table 6.2. Previous studies that have considered radiosonde temperature errors (Gaffen et al., 2000a, Parker and Cox, 1995 amongst others) conclude that observational errors are likely to increase with altitude. The increase in residual test failure between the models is, however, highly unlikely to be entirely a product of gross observational errors, unless such errors project strongly solely onto the principal modes of variability of HadCM3 and not those of HadCM2, which seems

highly improbable. Therefore gross problems may exist in the ability of HadCM3 to accurately predict upper tropospheric temperature changes. In particular, HadCM3 may be underestimating the magnitude of internal climate variability in this portion of the troposphere; leading more frequent failure of the residuals test, as the residuals appear too large in comparison (Collins et al., 2001, see eq. 15 in chapter 4 and accompanying discussion). This increased failure of the consistency test is independent of LAA diagnostic being considered, and hence is highly unlikely to be an artefact of these choices. Increased failure reduces confidence in the results for HadCM3 compared to those for HadCM2 within the upper troposphere.

Considering traces for the GS + VOL signal combination for HadCM3 (Figure 6.6), GS is robustly detected in all cases, and marginally consistent in most cases where residuals pass the consistency test. HadCM3 tends to overestimate the magnitude of the GS response, sometimes significantly. A HadCM3 VOL signal is both detected, and a consistent explanation of the observations at higher truncations for "smart" LAAs, and detected although generally only when residuals are inconsistent for "naïve" LAAs. For all LAA diagnostics, signal amplitude estimators for a HadCM3 VOL signal are negatively biased at low truncations. For neither GS nor VOL is it likely that the overestimation is an artefact of poorly constrained signals, as the SNR values are large (Tables 6.8 and 6.9). Unlike the analysis for HadCM2, in HadCM3 the VOL signal is generally seen not to be significantly overestimating the amplitude of the volcanic forcing upper tropospheric temperature response.

Splitting the GS signal into G + S for HadCM3, the first observation is that residuals are hardly ever consistent for any LAA diagnostics. Where they are, G is robustly detected, although it is often significantly overestimated in amplitude within HadCM3 for all four LAA diagnostics. Detection of the S signal is less certain, but it is almost always a consistent explanation of the observations. The exception is in "naïve" LAA diagnostics, where at low truncations the best-guess S signal amplitude estimate is negative for HadCM3. Hence, confidence in the presence of a demonstrable S signal in upper tropospheric temperatures is reduced, as the best-guess estimators should be positive even if they are not detected when the model signals are not unphysical representations of the observations. Negative values may partly be a result of including observed (and modelled) data from poorly-sampled

regions, as this negative amplitude estimate result is not repeated for the "smart" LAA diagnostics, although the overall trend is. In agreement with analyses for HadCM2, the HadCM3 S signal suffers from a low SNR, although for HadCM3 the SNR does not generally increase with truncation (Tables 6.8 and 6.9). In some cases the HadCM3 S signal is likely to be significantly noise contaminated according to SNR analysis, and therefore the OLS results may be significantly negatively biased.

In the three-way HadCM3 regression of G + S + VOL, individual estimators are consistent with those in the two-way regressions, as was the case for HadCM2. Therefore, G is robustly detected, but tends to be overestimated in amplitude in the model, although not significantly so, in contrast to the two-way regression. Detection of S and VOL are less certain, although both are generally consistent explanations of the observations. In all LAA diagnostics, residuals are only consistent over a small sub-set of the entire range of truncations being considered. That G becomes a more consistent explanation of the observations, when S and VOL HadCM3 fields are also considered, can be understood from a consideration of the sign of the trends in the input fields, with G being positive and S and VOL negative, at least in a global-mean sense. A consideration of both S and VOL signals requires a greater proportion of warming from G to fit the observations and, therefore, G is less overestimated in amplitude.

In common with HadCM2 results, individual signal amplitude estimators for HadCM3 are seen to vary in a synchronous manner for all three input signal combinations. Analysis of results as per Figure 6.5 shows that there exists a large covariance between the solutions (not shown here). This is likely to be for much the same reasons as detailed previously for HadCM2 model signals.

Finally in this section, recourse is made to global-mean temperature plots, to assess how well both models capture the transient nature of observed upper tropospheric temperature changes at the largest scale. If either model can capture the spatio-temporal evolution of upper tropospheric temperatures, then confidence in their ability will be increased. For each model, the observations are projected onto the leading modes of variability considered in the regression. Trends are not "true" observed global mean trends, but rather a truncated representation of these trends.

The calculation is performed at the maximum truncation for which all three input signal combinations considered pass the test on the residuals for the 10-area "smart" LAA diagnostics. Unless the models are identical in every detail, and the same truncation considered, expectations are that projected observed trends will differ between models, as the projection matrix, which transforms from raw observations to truncated representation, will not be identical. If the best-guess reconstructed observations are adequate they should fall within the model estimated uncertainty due to natural variability at the 90% C.I. This makes the implicit assumption that the model internal variability estimate is correct in both pattern and magnitude. The "observations" themselves are expected to vary more than the reconstruction as they are only based upon a single observational dataset, whereas the reconstruction is based upon scaling factors on ensemble average signal predictions, which will have lower variance. It should be noted that the component forcing responses which make up the best-guess reconstruction of the observations are scaled by the best-guess amplitude estimators from the 10-area "smart" LAA OLS results at the truncation being considered. A different reconstruction would result from considering the base input (unscaled) forcing responses, or results based upon alternative LAA input diagnostics, or at different truncations. Only the 10-area "smart" LAA diagnostic results are considered in this analysis, being the preferred input diagnostic. Expectations are that overall trend agreement will be good, as the regression model attempts to scale the input signal fields to recreate the observations. A stricter test is whether the nuances of the time-varying response are adequately captured.

Figure 6.7 illustrates reconstructions of upper tropospheric globally-averaged temperatures for all three forcing combinations considered in Figures 6.4 and 6.6. As expected, the "observations" differ between models, as projection matrices are not identical. For HadCM2, the reconstructed upper tropospheric global mean temperature series estimate is within 2σ of the observed value throughout the 35-year period for all three input signal combinations. The majority of the observed positive trend in temperatures is due to well-mixed greenhouse-gases, with slight sulphate induced cooling throughout, and volcanic cooling early and late in the period. Purely visually, incorporating HadCM2-predicted volcanic effects improves the level of agreement between observations and reconstruction, even though this signal is not

detected in HadCM2 (see Figure 6.4 and accompanying discussion). For HadCM3 the reconstruction always significantly overestimates global-mean temperatures in the mid-1970's, and when volcanic forcings are included underestimates them in the late 1960's. The overall trend is again dominated by greenhouse-gas induced warming for HadCM3. Unlike for HadCM2, the VOL signal does not appear to increase the level of agreement of the reconstruction with the observations, if anything degrading it instead. This does not directly imply that VOL in HadCM3 is unimportant in explaining observed trends, as it may be important in explaining smaller spatial scale features. Global-mean upper tropospheric temperature plots help to confirm principal detection results, in that for both models the effects of well-mixed greenhouse-gases are required to explain the large-scale observed upper tropospheric warming, whilst the effects of sulphate aerosols and volcanic eruptions are more marginal in explaining the trends. In no case do the models grossly fail to reproduce the observed temporal evolution of global mean upper tropospheric temperature trends, increasing confidence in the ability of the models.

6.3 Lower tropospheric temperatures

Table 6.3 details detection results for the lower tropospheric temperatures (c.f. Table 6.2). In agreement with upper tropospheric analyses (section 6.2), results for both models for lower tropospheric temperatures imply the robust detection of well-mixed greenhouse-gases, either individually or in combination with other forcings. There is also some evidence for a demonstrable sulphate aerosol influence on lower tropospheric temperatures in both models. For HadCM2 there is stronger evidence for a detectable volcanic influence than was the case for upper tropospheric temperatures, whilst HadCM3 results still point towards a detectable volcanic influence. For each model there is also weak indication of a detectable solar influence on lower tropospheric temperatures, although in the majority of cases the detection fails at low truncations, reducing confidence in this result. It is further bought into question by the finding that a solar forcing influence is never detected in any input signal combination for either model which additionally considers the response to anthropogenic forcings. Given results detailed in Table 6.3, the same three input signal combinations are considered in further detail for lower

tropospheric as for upper tropospheric temperatures (GS + VOL, G + S, and G + S + VOL). A case could also be made for considering solar influences, but this is less than compelling given the restricted circumstances in which a solar influence is detected for either model.

For both models, but especially for HadCM3, detection results detailed in Table 6.3 indicate that there are far fewer cases where the residuals fail at low truncations for lower tropospheric than for upper tropospheric temperatures, at least when any signals are detected. This is confirmed in detection traces shown in Figures 6.8 and 6.9 (c.f. Figures 6.4 and 6.6), which illustrate that residual test failure at any truncation is much less likely for lower tropospheric temperatures than for upper tropospheric temperatures in both models, at least for the three specific input signal combinations being considered in further detail. This behaviour may be caused by reduced observational error (Gaffen et al., 2000a), increased model realism (Gillett et al., 2000, Collins et al., 2001), or a combination of these two factors in the lower troposphere. It is not possible in the current analysis to accurately, and quantitatively, distinguish between these potential causes.

Detection traces for HadCM2 considering HadRT2.1s lower tropospheric temperatures are shown in Figure 6.8. For the GS + VOL signal combination, GS is both robustly detected and consistent in all four input LAA cases, and for nearly all truncations. The VOL signal is only very occasionally detected, and is not always a consistent explanation of the observations. The best-guess VOL signal amplitude is almost always positive, as was the case for upper-tropospheric temperatures, and systematically less than 1, implying that in the lower troposphere the volcanic signal response is again of the correct pattern but significantly overestimated in amplitude in HadCM2. Analysis of SNR values in Tables 6.8 and 6.9 show that this is highly unlikely to be due to a poorly-defined volcanic signal. There is a marked reduction in the synchronicity of changes in GS and VOL amplitude estimators for HadCM2 compared to results for the upper troposphere (c.f. Figure 6.4). The principal axis of the uncertainty ellipse for lower tropospheric temperatures in Figure 6.5 is not as aligned towards the diagonal as was the case for upper tropospheric temperatures. The implication is that the global mean change component is likely to be less

important in explaining lower tropospheric temperatures, than it is in explaining upper tropospheric temperatures.

Results for the G + S input signal combination for HadCM2 (Figure 6.7) yield that G is robustly detected, and in all four LAA diagnostics, and at almost all truncations is a consistent explanation of the lower tropospheric temperature observations. An S signal is less robustly detected, although it too is a consistent explanation of the observations in all LAA diagnostics, and at the majority of truncations. In common with results for upper tropospheric temperatures, the S response amplitude estimate is poorly constrained, most likely because of low SNRs, especially at low truncations in HadCM2 (Tables 6.8 and 6.9).

Components of the three-signal regression for HadCM2 (Figure 6.8) track their individual signal amplitude estimates from the two-signal regressions discussed above. Therefore, G is robustly detected and generally a consistent explanation of observed lower tropospheric temperatures, S is sometimes detected and nearly always consistent, and VOL is occasionally detected but overestimated in amplitude within HadCM2. This overestimation problem for VOL is reduced when G and S are allowed to vary rather than being a fixed ratio (GS), leading to a slightly increased occurrence of detection of a VOL signal. This result is seen in cases where the best-guess amplitude estimator of S is close to zero in the 10-area "smart" LAA diagnostic (although not for other input LAA diagnostics). Therefore, it may be that there is a degree of co-linearity between the S and VOL forcing responses leading to degenerate solutions. This is confirmed in Table 6.3 where three signal combinations were found to be potentially degenerate in the 10-area "smart" LAA diagnostic. Results for all input signal combinations analysed are insensitive to the choice of HadRT dataset being considered. However, there is increased failure of the test on residuals when HadRT2.1 is used, as discussed previously with regards to upper tropospheric temperatures.

GS + VOL signal combination detection traces for HadCM3 lower tropospheric temperature fields shown in Figure 6.9 indicate robust detection of a GS signal, which is a consistent explanation of the observations. Results for VOL are more uncertain, the signal hardly ever being detected when the residuals test is passed, and

are at best marginally consistent, tending to be overestimated in amplitude in HadCM3. In the preferred "smart" 10-area LAA diagnostic, the best-guess HadCM3 VOL signal amplitude 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 explain the observations at these truncations, which on purely physical grounds cannot be correct. This is unlikely to be an artefact of weak signals as SNR analysis (Table 6.8) shows that the signal is well defined. Therefore, confidence in a volcanic influence on temperatures within the lower troposphere is reduced, at least for HadCM3. In the G + S signal combination for HadCM3, G is both robustly detected and consistent, whereas S is robustly detected in "smart" LAA diagnostics but not "naïve" LAA diagnostics, although in all cases it remains a consistent explanation of the observations. The HadCM3 S signal results are very poorly constrained at low truncations, improving marginally with increasing truncation, possibly due to the signal being poorly defined (low SNR values) leading to uncertain estimators (Tables 6.8 and 6.9). Estimates of HadCM3 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 amplitude scaling estimates. Reasons for this behaviour are unknown, requiring further investigation outside the remit of this thesis, but it is unlikely to be due to signal covariance as there is no similar shift in the G amplitude estimates. This illustrates the importance of considering a broad range of pre-processing techniques to reduce any ambiguity in the results. The three-way regression results for HadCM3 mirror those for the individual components discussed above. For all three input signal combinations being considered in HadCM3, results are insensitive to which version of HadRT dataset is being used as the observations, although considering HadRT2.1 again increases the frequency of residuals consistency test failure.

Global-mean lower tropospheric temperature reconstructions are shown in Figure 6.10. Reconstructions based upon HadCM2 input fields are unable to resolve the observed minimum in the mid-1970s in all cases. This may be due to inaccuracies in the sulphate forcing history in HadCM2, as for HadCM3 reconstructions there is a large sulphate cooling effect through the 1960s and early 1970s, which is not evident in HadCM2 reconstructions. For HadCM2, volcanic forcings help to explain the observed maximum in global-mean lower tropospheric temperatures in the late

1980s. Without accounting for volcanic influences, the reconstruction for HadCM2 is of a more monotonic increase than is actually the case, implying that volcanic influences may be important in accounting for recent global mean lower tropospheric temperature trends (a finding in agreement with Santer et al., 2001). The component forcing that explains the majority of the observed lower tropospheric warming trend according to HadCM2 is well-mixed greenhouse-gases. All HadCM3 reconstructions fall within the uncertainty range 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. Most of the warming trend is derived from greenhouse-gases, although this is moderated by the cooling effects of sulphate aerosols, particularly important in explaining trends early in the period. As for upper tropospheric reconstructions, 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, whereas sulphate aerosols and volcanic influences are less necessary, as they both tend to cool the lower troposphere over this period.

6.4 Near-surface temperatures

Recently observed near-surface temperatures have been compared in numerous optimal detection studies to date with both HadCM2 and HadCM3 model output (Allen et al., 2000, 2001, Barnett et al., 1999, 2000, Hegerl et al., 2000, Stott et al., 2001, 2001a, Tett et al., 1999, 2001). Therefore, it is instructive to compare the current analysis with these previous analyses. Previous studies have consistently yielded detectable well-mixed greenhouse-gases (G) and, more tentatively, sulphate aerosol (S) influences on the near-surface temperature record for the latter half of the 20th Century. There is also evidence in some studies (where such forcings have been considered) for detectable solar and volcanic influences, although detection of these influences is shown to be sensitive to methodological considerations. All previous studies have considered the full available observed field, whereas here consideration is being given to a much more data-sparse representation. This is also the first time that both any representation of spatial patterns of change other than spherical harmonic coefficients, and any temporal sampling other than decadal resolution for

50-year or greater trend lengths, have been employed in space-time optimal regression detection studies considering near-surface temperature fields. Effects of changing observational coverage, and changes to the input field pre-formatting cannot be completely separated in the present analysis. Expectations are that reduced spatial data coverage and trend length will both tend to reduce the power of the detection algorithm by increasing uncertainty due to natural internal climate variability, against which tests for the detection statistic significance are made (Santer et al., 1995, Barnett et al., 1998). There remains scope for performing a three-step analysis, moving from decadal to five-year averages, truncating to a shorter time period, and then using LAAs with the original data coverage, before undertaking the present analysis. Under such an approach the causes of any differences in the results could be effectively separated (c.f. Gillett et al., 2001). However, this is at most of only marginal interest to the current analysis, as what is desired is a comparison with results for upper-air temperatures under a common framework, so coverage should be coincident with available HadRT coverage (Santer et al., 1999).

Detection results are shown for near-surface temperatures in Table 6.4 for both models. There exists only one version of the HadCRUTv dataset, so the number of columns in this table is reduced compared to previous free troposphere analyses (c.f. Tables 6.2 and 6.3), which additionally accounted for uncertainty due to the version of HadRT dataset being considered. The most striking feature of results detailed in Table 6.4 when compared to those in Tables 6.2 and 6.3 is that there are a far greater number of cases where the model signals are detected in free tropospheric layer average temperature diagnostics than there are in near-surface temperature diagnostics. Solely from Table 6.4, the conclusion is that a GS signal is the most likely cause of recent near-surface temperature observations in both models, possibly in its component G and S parts, at least in HadCM2. This finding is consistent with previous near-surface temperature detection studies, providing some confidence in the use of LAA input diagnostics on sparser input fields and over a reduced trend period. In both models, anthropogenic (GS) signal strength is significantly overestimated according to results from OLS regression analysis. However, for HadCM2 this may solely be caused by known systematic negative biases in the upper bounds of the OLS estimator (AS01), as results from TLS do not appear to yield

significant differences between observed and modelled near-surface temperature fields when signals are detected. Confidence in a demonstrable anthropogenic influence is reduced as results are highly dependent upon both the LAA diagnostic, and the signal combination being considered. Expectations are for a type I error (single detection of any one signal) to occur solely by chance. Therefore, the detection of LBB for HadCM3 and TLS regression seen in Table 6.4 is ignored, as it is not replicated in any other combinations, greatly reducing confidence in the presence of a demonstrable solar influence. For consistency with previous analyses for free troposphere temperatures, further consideration is given to the same three input signal combinations (GS + VOL, G + S, G + S + VOL), although from Table 6.4 expectations are that regression estimates for volcanic influences may be highly uncertain.

Detection traces for HadCM2 near-surface temperatures are given in Figure 6.11. In the GS + VOL signal combination, GS is both marginally detected and a marginally consistent explanation of the observations for some, but by no means all, truncations for each LAA diagnostic. Best-guess amplitude estimates for GS in HadCM2 are consistent between LAA diagnostics in suggesting that the model signal strength at the near-surface is being overestimated by approximately 100%. This is at odds with previous analyses (Tett et al., 1999, Stott et al., 2001, 2001a) which yielded best-guess amplitude estimates for a latter twentieth century HadCM2 derived GS signal of approximately unity. Reasons for this difference could relate to either spatial or temporal pre-processing differences in the input diagnostics, or the shorter period being considered. Also, in comparison, in the free troposphere the HadCM2 GS signal estimate is seen to be of approximately the correct magnitude, with no discernible bias (c.f. Figures 6.4 and 6.8). There is no evidence that SNRs for GS are systematically decreased compared to free troposphere temperature diagnostics (Tables 6.8 and 6.9), leading to negatively biased estimators, as might be expected given that near-surface temperatures are likely to contain many more degrees of freedom (Jones et al., 1997). Therefore, results point towards the potential existence of a systematic error in either HadCM2 or the observations, whereby model GS ensemble output is required to be scaled by very different scaling factors to fit near-surface and free troposphere temperature observations. The HadCM2 VOL signal amplitude in near-surface temperature observations is zero or close to zero in all four

LAA diagnostics, and includes negative values, implying a negligible volcanic influence. Considering G and S, both tend to be overestimated in amplitude in HadCM2 in all four LAA diagnostics and at all truncations, although S is not significantly overestimated in any LAA diagnostic or at any truncation. G is occasionally detected in all four input diagnostics, whereas S is detected only once (in the 12-area "naïve" LAA diagnostic). The S signal also exhibits very low SNR, particularly at low truncations, for all four LAA input representations (Tables 6.8 and 6.9) and, therefore, may be significantly negatively biased. Individual components of the three-way regression for near-surface temperatures track their components in the two-way regressions described above so are not explicitly discussed here.

HadCM3 detection traces for near-surface temperatures are given in Figure 6.12. As for HadCM2, GS, when considered in combination with VOL, is only occasionally detected and the signal strength overestimated, normally significantly. For HadCM3, the damping required on the GS signal estimate is even greater than that for HadCM2. This is in keeping with previous analyses at the near-surface that HadCM3 is more likely to overestimate the magnitude of the late 20th Century anthropogenic response (see Figures 7 and 11 of Tett et al., 2001, c.f. Tett et al., 1999, Stott et al., 2001, 2001a). However, the results detailed here exhibit a greater overestimation bias than previous space-time optimal detection studies (Tett et al., 2001). The systematic effect on results for both models indicates that it is most likely to relate to input field pre-processing in some manner. The major difference in observed and modelled input fields is the complete lack of high latitude Southern Hemisphere or ocean temperature data in the versions used here. This may explain the majority of the observed difference in detection results, but is not explicitly tested here. For HadCM3, in agreement with HadCM2, a VOL signal is discounted, being at or around zero scaling required in nearly all cases, the exception being small truncations for the 10-area "smart" LAA diagnostic which are positive, but where the signal is seen to be poorly defined (Table 6.8). G is only rarely detected in combination with S for HadCM3, and almost always significantly overestimated in amplitude within the model. S is generally a consistent explanation of observed near-surface temperatures, but only ever marginally detected in "smart" LAA diagnostics, and never detected in "naïve" LAA diagnostics. Results for the individual HadCM3 signals discussed above are generally insensitive to the consideration of a three-way regression. The

exception is VOL, for which the seemingly anomalous positive non-zero estimates at low truncations in the 10-area "smart" LAA input diagnostic identified earlier disappear. This reduces further any residual confidence in a demonstrable HadCM3 estimated volcanic influence on near-surface temperatures.

Global mean near-surface temperature reconstructions are shown in Figure 6.13. For both models those reconstructions including VOL can be discarded as physically plausible reconstructions, since best-guess amplitude scalings applied to this forcing are negative at the truncation considered (see Figures 6.11 and 6.12). Therefore, concentrating on the two G + S reconstructions, both capture observed trends in near-surface temperatures within 2σ, with greenhouse-gases providing the major component of the warming trend, in agreement with results for the free troposphere. However, the absolute trend over the entire period is not well captured, being underestimated by both model reconstructions. This may be due to the down-weighting required on both signals in the models in this analysis, as Stott et al (2001) and Tett et al. (2001), both of which give signal amplitude estimators closer to unity, more adequately capture the global-mean trend. Furthermore, it should be noted, that for free troposphere layer average temperature reconstructions, scalings applied were very different to those applied for near-surface temperatures. The model is being scaled to fit the observations, and by a different factor for each tropospheric temperature variable being considered. It is of interest to consider whether a scaling factor, or range of scaling factors can be chosen which satisfies all solutions such that no warping of the model other than by a constant value is required to reconstruct the observations for each observed parameter being considered. This is effectively a test for model internal consistency, and is returned to briefly in section 6.9, and in far greater detail in chapter 7.

6.5 Free troposphere lapse rates

Detection results for free troposphere lapse rates are shown in Table 6.5. For HadCM2 there are many fewer positive detections than there were for either of the individual component free troposphere layer average temperature variables (sections 6.2 and 6.3). Residuals also generally fail the test for consistency at low truncations

for HadCM2 in those cases when HadRT2.1s is used as the observational dataset, whereas they do not where HadRT2.1 is being considered. A systematic difference also exists in which signals are detected in HadCM2 between HadRT versions, with some evidence for both anthropogenic and volcanic forcings when considering HadRT2.1s, whereas there is only convincing evidence for volcanic influences when considering HadRT2.1. In no case is either a HadCM2-predicted stratospheric ozone depletion or solar influence detected, so these can be discounted. Detection results for HadCM3 in Table 6.5 are even more sensitive to the choice of HadRT dataset being considered than those for HadCM2. In no case is detection achieved at the maximum truncation of 21 for HadCM3 free troposphere lapse rate diagnostics. When signals are detected at lower truncations, results for HadRT2.1s point towards anthropogenic influences, with a single detection of a volcanic signal in combination with these. For HadRT2.1 there are only two occasions when signals are detected, and these are anthropogenic influences in three-way regression combinations, and only for "smart" 10-area LAA input diagnostics under OLS. Therefore, for both models there is great uncertainty according to these results, although in both cases the most likely explanations remain anthropogenic forcings, either on their own or in combination with volcanic influences. Hence the same three input signal combinations (GS + VOL, G + S, G + S + VOL) are considered in greater depth here as in layer average temperature diagnostics considered in sections 6.2 to 6.4.

Detection traces for free troposphere lapse rates for all three input signal combinations are shown for HadCM2, considering HadRT2.1s observed data, in Figure 6.14. Compared to previous analyses of layer average temperatures, amplitude estimator confidence limits are larger for all signals being considered, most likely at least in part because the SNR values are consistently lower in free troposphere lapse rates (Tables 6.8 and 6.9). In most cases SNR values, although lower, remain significantly greater than that expected were noise a limiting factor, and therefore OLS estimators should not be significantly negatively biased (AS01, Tett et al., 2001). Lower SNR values confirm previous analysis in chapter 3, which suggested that SNRs might be a limitation in lapse rate detection studies. This result is not model dependent, as confirmed by a consideration of HadCM3 traces (Figure 6.15), and SNR values (Tables 6.8 and 6.9). A consideration of lapse rates will be useful, however, so long as the signals can be shown not to be significantly noise

contaminated, even if large amplitude estimate uncertainties mean that results add limited power to the detection results per se. The signal estimators should still yield positive best-guess amplitude estimates if the model exhibits skill, even if the uncertainty range in these estimates is no longer significantly non-zero. Specifically, the signal amplitude estimators should still be consistent with the observations if the model is adequate. The estimators should also not be significantly different to those for layer average temperatures, where signals are more robustly detected (see section 6.9 and chapter 7).

Traces for GS in the GS + VOL signal combination in HadCM2 (Figure 6.14) show that, for all four LAA inputs, GS is a consistent explanation of free troposphere lapse rate observations, even when the residuals of the regression are inconsistent. A GS signal is also detected for some truncations in all four LAA input diagnostics for HadCM2, and the best-guess signal amplitude is always positive, giving increased confidence in the presence of a HadCM2-predicted GS signal in the observations. For the HadCM2 VOL signal, estimates are nearly always consistent with the observations and with positive best-guess amplitude estimators, and it is also occasionally detected. The best-guess signal amplitude estimators of VOL are always less than unity for HadCM2, except for the case of the 6-area "naïve" LAA input, which gives systematically higher estimates than the other three LAA diagnostics, and also has an anomalously high GS amplitude estimate. Evidence here suggests that the volcanic response is likely to be overestimated in magnitude within HadCM2, a finding consistent with previous free tropospheric layer average detection results for the HadCM2 VOL signal (sections 6.2 and 6.3). The solutions also exhibit strong covariance, confirmed by analysis of ellipses (not shown here, see earlier discussion in section 6.2). The systematic increase in amplitude estimates for both GS and VOL must be due to the different spatio-temporal representation used in 6-area "naïve" LAA diagnostics as compared to those in the other three LAA representations. This confirms that without considering a reasonable range of input pre-processings, potentially ambiguous results could arise in detection studies.

For the G + S signal combination in HadCM2 (Figure 6.14), both amplitude estimators for free troposphere lapse rates are negatively biased at low truncations in all input diagnostics except the 6-area "naïve" LAA diagnostic, which again behaves

anomalously. This negative bias may relate to the low SNR values, potentially leading to negatively biased OLS estimators, particularly for S (Tables 6.8 and 6.9). G is only sometimes detected, detection being dependent both upon truncation and input LAA diagnostic, but it is always a consistent explanation of the observations. S is also only sometimes detected, depending upon the LAA diagnostic and truncation being considered, but nearly always a consistent explanation of the observations. It is normally detected only when residuals from the regression are inconsistent. This reduces confidence in the presence of a demonstrable sulphate aerosol influence, as simulated by HadCM2, in observed free troposphere lapse rates. Confidence is further reduced, as at small truncations there is a tendency for the best-guess HadCM2 S signal amplitude estimators to be negative, although this may be due solely to the low SNRs at these truncations (Table 6.8) leading to highly negatively biased signal amplitude estimators. At higher truncations, best-guess HadCM2 S signal amplitude estimates are always positive, although they are still likely to be noise contaminated (Table 6.9) and, therefore, potentially negatively biased.

The three-way regression result for HadCM2, in common with that for layer average input temperature fields, is similar to the results described for the two-way regressions above. Amplitude estimators for S are consistently higher compared to those when only G is considered, rather than G and VOL. Therefore, a weaker S signal is required to explain the observations when both additional forcings are applied. There are similar shifts in both the G and VOL amplitude estimates implying a degree of covariance, although these shifts are small in comparison to those for S.

Contrary to the results of analysis of layer average temperatures, there is no systematic increase in the number of cases where the consistency test on the residuals passes for “smart” LAA diagnostics over “naïve” LAA diagnostics for free troposphere lapse rates. In fact, if anything, the reverse is true, at least for HadCM2. This observation is independent of the model being considered, as HadCM3 results exhibit a similar pattern (Figure 6.15). “Smart” LAA diagnostics are heavily weighted to Northern Hemisphere mid-latitude continental regions. Analysis in chapter 2 suggests that the Arctic Oscillation (Thompson and Wallace, 2000) has a strong signal on inter-annual and longer timescales in the HadRT record at these

latitudes, with the AO temperature signal consisting of cooling aloft and warming nearer the surface in its positive phase. Gillett et al. (2000a) have shown how this mode of variability is poorly defined in HadCM2, but can be estimated and removed, and how this does not affect primary conclusions of detection studies for near-surface temperatures. Here, no attempt has been made to remove the AO signal component from the observed and modelled fields. If LAA input diagnostics concentrate on those regions in which the AO projects strongly, then expectations are that the model might grossly fail to capture variability in free troposphere lapse rates. This would lead to an increased frequency of occurrence of residuals consistency test failure as the independent realisation of natural variability with which the residuals are being compared will be underestimated. "Naïve" LAA diagnostics give equal weighting to equal areas in the region from 30°S to 90°N. Any bias introduced by the inability of the models to capture AO variability and trend would, therefore, be likely to have a more limited impact on consistency test results of detection studies considering these input LAA diagnostics.

There is also a marked reduction in the frequency of consistency test failure in traces for HadCM2 when results based upon HadRT2.1 observations (not shown here) are considered instead of those based upon HadRT2.1s. This result is not model-dependent as results for HadCM3 are similar in nature. It is assumed that corrections applied between the HadRT series primarily remove gross errors in the vertical gradients of the individual records upon which the HadRT series is based. This appears reasonable given the methodology applied by Parker et al. (1997). In removing vertical discrepancies, errors in the difference field between any two levels will be decreased, even if corrections applied add spurious spatio-temporal patterns of change to the individual component fields. As a result, any application of the regression algorithm to this corrected lapse rate field will be more likely to yield residuals which resemble an independent estimate of real-world variability. Further, corrections applied between the HadRT series tend to bias all the estimators negatively in a systematic manner, although overall trend patterns remain similar. Hence all signals are both more likely to be overestimated in amplitude and, generally, less detectable in HadCM2 for HadRT2.1 than for HadRT2.1s. This should be balanced against the fact that as the residual test fails less, detections of

both S and VOL which failed because of consistency test failures on the residuals under HadRT2.1s (Figure 6.13), are detected under HadRT2.1. Corrections to observed fields may have corrected dubious values, which yielded spuriously high OLS estimators of the scalings required on the HadCM2 signals to recreate HadRT2.1s. There is no guarantee that any other observed error would bias results of the detection algorithm in a similar manner in other atmospheric variables.

Detection traces for HadCM3 give more ambiguous results for free troposphere lapse rates than those for HadCM2 (Figure 6.15 c.f. Figure 6.14). In the GS + VOL signal combination, both estimators are highly uncertain for all four input LAA diagnostics for HadCM3. GS best-guess amplitude estimates are generally positive when the test on the residuals is passed. GS is only ever detected in the 6-area "naïve" LAA diagnostic, and generally HadCM3 overestimates the GS signal strength in all LAA diagnostics, often significantly. The VOL signal is even more uncertain in HadCM3, the amplitude estimators being negative at a number of truncations for each input LAA diagnostic, greatly reducing confidence in a HadCM3-predicted volcanic influence on free troposphere lapse rates. Tables 6.8 and 6.9 show that the VOL SNR is low in free troposphere lapse rates and, therefore, these results may be due to known negative biases under the OLS approach. If the HadCM3 VOL signal is present in the observations, then it too is likely to be overestimated, possibly significantly. Considering G + S HadCM3 signals, G best-guess amplitude estimators are positive in all cases except for the 5-area "smart" LAA diagnostic, although G is only detected in the 6-area "naïve" LAA diagnostic, and tends to be overestimated in amplitude in HadCM3, sometimes significantly. Signal strength estimators for S from HadCM3 are highly uncertain, including large negative amplitude best-guess estimates at some truncations in all four LAA diagnostics. This may solely be due to the low SNR for S, at least in some LAA diagnostics (Tables 6.8 and 6.9). The S signal is only detected once, at low truncation, where low SNR may be a limiting factor (Table 6.8), in the 10-area "smart" LAA diagnostic. In most cases, the S signal is an inconsistent explanation of the observations, being significantly overestimated in amplitude within HadCM3. The result for S remains in the three-way regression, and hence confidence in the presence of a sulphate aerosol influence as predicted by HadCM3 in observed free troposphere lapse rate fields is very low. G and VOL signals exhibit much the same behaviour in the three-way regression for HadCM3 as

they did in their respective two-way regressions. Primary conclusions are independent of whether HadRT2.1 is used in place of HadRT2.1s, although S becomes even less plausible an explanation.

Global-mean free troposphere lapse rate temperature reconstructions for both models (Figure 6.16) show that the model-estimated natural internal variability in the observations is greatly reduced compared to that for layer-average temperatures (c.f. Figures 6.7, 6.10, and 6.13). Intuitively, on purely physical grounds and considering the annual timescales considered in the present study, it is expected that upper and lower tropospheric temperatures would tend to exhibit a large degree of covariance. Reduced model-based variability estimates, when compared to other temperature variables considered, may, therefore, be correct, but this is not implicitly tested here. The overall observed global-mean trend in free troposphere lapse rates is almost zero for the period 1960-1995 in both representations of the observations, with projection onto the leading modes of HadCM2 yielding greater overall inter-period variability than that for HadCM3. This increased variability estimate is likely, at least in part, to explain why the residuals test fails less for HadCM2 than HadCM3. It implies that HadCM3 might be grossly underestimating natural internal climate variability for free troposphere lapse rates. This gives added weight to conclusions in section 6.2 that HadCM3 may also be grossly underestimating upper tropospheric temperature variability. All three input signal combinations considered adequately capture the observed changes for HadCM2. The same is true for HadCM3, but these can all be discounted as at least one of the component signals being used in the reconstruction is scaled by a negative (unphysical) scaling factor in each case at the truncation considered (Figure 6.15).

6.6 Entire troposphere lapse rates

Detection results for entire troposphere lapse rates detailed in Table 6.6 yield very few signal detections for either model. For both models, results are highly sensitive to choice of LAA input diagnostic, regression algorithm, and HadRT version used to derive the upper tropospheric layer average temperature estimate. For HadCM2, detection only occurs for HadRT2.1s observed upper tropospheric temperatures, and

considering 10-area "smart" LAA diagnostics under a TLS regression approach. In the few cases of positive detection, GS or its component G + S parts are detected. There are two input combinations which yield detectable signals for HadCM3 entire troposphere lapse rates according to Table 6.6. For HadRT2.1s upper tropospheric series, and considering a 5-area "smart" LAA diagnostic under an OLS regression approach, there is strong evidence for a volcanic influence in HadCM3. The other case is considering HadRT2.1s upper tropospheric temperatures under a TLS approach for 10-area "smart" LAA input diagnostics, where anthropogenic forcings are detected in HadCM3. The lack of positive detection results for both models in the majority of input signal combinations considered in Table 6.6 may solely be due to the lower SNRs (Tables 6.8 and 6.9). For consistency with analyses in previous sections, therefore, the same three input signal combinations (GS + VOL, G + S, G + S + VOL) are considered in further detail here for each model.

Detection traces for HadCM2 entire troposphere lapse rates (Figure 6.17) yield failures of the consistency test on the residuals solely for "naïve" LAA diagnostics, in marked contrast to behaviour of the statistic for free troposphere lapse rates, but in agreement with that for individual layer average analyses. This result is independent of the HadRT version used to derive the observed upper tropospheric temperature field. Given that there are no truncation failures in any cases for HadCM3 entire troposphere lapse rates signal combinations considered here (Figure 6.18), it is difficult to confirm whether this is an artefact of the model or a more general pattern. This behaviour is repeated for lower troposphere lapse rates for HadCM2 (Figure 6.20). HadCRUTv is likely to be well constrained in comparison to the HadRT record, being based upon relatively constant observing practices, and having been more rigorously investigated for potential sources of uncertainty (section 1.2). For free troposphere lapse rates, the difference field is between two HadRT layer averages, both of which are likely to contain similar residual error structures that essentially cancel each other out. In those lapse rate diagnostics including a HadCRUTv component, the difference field will yield large erroneous values for those locations where the HadRT tropospheric temperature record remains in error. This will likely resemble nothing seen in the models, at least in their leading modes of variability and, therefore, will tend to yield inconsistent residuals. HadRT records are more likely to be in error in poorly sampled regions, where consistency checks

with respect to near-neighbours have not been possible (chapter 2). The effects of HadRT residual errors are, therefore, likely to cause increased consistency test failures, for "naïve" LAA diagnostics in particular. Effects of model deficiencies in capturing leading modes of variability such as the AO, which is likely to be important in free troposphere lapse rate diagnostics, are unlikely to be the primary cause of consistency test failure in those lapse rate diagnostics which incorporate a near-surface temperature component.

For HadCM2 detection traces for entire troposphere lapse rates (Figure 6.17), there are very few cases of signal detection in any of the input signal combinations being considered. This is likely to be due at least in part to the low SNRs evident for all the signals (Tables 6.8 and 6.9) leading to inflated uncertainty estimates compared to earlier temperature variables considered. The GS signal best-guess amplitude estimate in the GS + VOL input combination for HadCM2 is always positive, and GS is almost always a consistent explanation of the observations for "smart" LAA diagnostics. Results for "naïve" LAA diagnostics are more ambiguous, containing negative GS best-guess amplitude estimators, although most of these occur at truncations where the residuals are inconsistent, reducing confidence in these seemingly erroneous estimates. Best-guess VOL signal amplitude estimates for HadCM2 are always positive for the 10-area "smart" LAA diagnostic, and nearly always consistent, but never detected. For other LAA diagnostics, the best-guess VOL signal amplitude estimate for HadCM2 is essentially zero, and includes negative values. This reduces confidence in the presence of a demonstrable volcanic influence on entire troposphere lapse rates for HadCM2. Both G and S best-guess amplitude estimates for HadCM2 are positive at all truncations for "smart" LAA diagnostics, although very small at high truncations for the 5-area "smart" LAA diagnostic. The signal amplitude estimates are also consistent in most cases with the observed entire troposphere lapse rates. Results for "naïve" LAA diagnostics are more ambiguous, with negative best-guess amplitude estimates for both G and S HadCM2 signals, although these occur mainly when the test on the residuals fails, reducing confidence in these estimates. Therefore, evidence for demonstrable HadCM2 G + S signal influences is at best ambiguous in free troposphere lapse rates. Results for the three-way regression for HadCM2 are similar to those described above for the component forcings, with slightly greater confidence in a VOL signal

for both “smart” LAA input diagnostics than was the case in the two-way regression of GS + VOL. Details of detection traces for HadCM2 are not dependent upon the HadRT dataset used to derive the observed upper tropospheric temperature field employed in calculating the lapse rate. In agreement with free troposphere layer average analyses, there is an increased frequency of residual test failures when HadRT2.1 is used as the observed dataset.

Detection traces for HadCM3 signals for entire troposphere lapse rates (Figure 6.18) are similar to those for HadCM2, with signal detection being very rare. The seemingly strong detection of a VOL signal for HadCM3 in Table 6.6 is seen to be solely an artefact of truncation choice for the 5-area “smart” LAA diagnostic. Furthermore, in other input LAA diagnostics VOL is never detected. The VOL signal suffers from problems of low SNR at all truncations (Tables 6.8 and 6.9). This could possibly cause the best-guess OLS estimator to be periodically negative for all four input LAA diagnostics. These factors taken together greatly reduce confidence in the seemingly quite strong VOL detection result in Table 6.6. This reinforces the need to consider detection traces, rather than results for single truncations, as well as numerous pre-processing choices. Based solely upon the relevant column in Table 6.6, an unjustifiably strong conclusion as to the likely presence of a significant volcanic influence would have been reached. GS amplitude estimates for HadCM3 are always positive in “smart” LAA input diagnostics, but the signal is overestimated in amplitude within HadCM3, significantly so at latter truncations. For “naïve” LAA diagnostics the GS signal yields negative best-guess estimators at some truncations in each case, reducing confidence in the presence of a HadCM3 GS forcing signal in the observed entire troposphere lapse rates. Analysis of Tables 6.8 and 6.9 shows that the GS SNR values for HadCM3 are consistently lowest in this lapse rate variable out of all six temperature variables being considered. However, SNR values are still large enough not to be significantly noise contaminated, and therefore OLS results are unlikely to be significantly negatively biased. Splitting this HadCM3 GS signal into its G + S components yields estimators for G which are positive at all truncations for the 10-area “smart” LAA diagnostic, and all but the last truncation for the 5-area “smart” LAA diagnostic, but more uncertain for “naïve” LAA diagnostics. The G signal tends to be overestimated in amplitude within HadCM3 in all LAA diagnostics, sometimes significantly. Results for the HadCM3 S signal are more

ambiguous, with negative best-guess amplitude estimates for at least some truncations in each of the LAA input diagnostics. Hence, confidence in a demonstrable S influence is low in this entire troposphere lapse rate diagnostic for HadCM3. However, SNR analysis (Tables 6.8 and 6.9) shows that S is significantly noise contaminated in all four LAA diagnostics, and therefore analysis of OLS estimators may be yielding highly negatively biased and uncertain results. Caution is therefore advised against deriving too much from these plots. Results for individual signals are seen to be insensitive to splitting the signals further to perform a three-way regression, or which version of the HadRT upper tropospheric temperatures is being used to derive the lapse rate. In common with most other temperature variables, there is an increased frequency of residual consistency test failure when HadRT2.1 is considered.

The overall observed global-mean trend in entire troposphere lapse rates (Figure 6.19) for the period 1960-95 has been one of relative upper tropospheric cooling according to the projection of the observations onto the leading modes of variability of each model. This trend has been discontinuous, with warmings and coolings between 5-year periods, especially in the late 1970s (and more so for the HadCM3 projection), when there is a distinct cooling evident. Uncertainty estimates are much greater than those for free troposphere lapse rate temperatures. The boundary layer is likely to act as a semi-permeable boundary to the communication of temperatures between the free troposphere and the near-surface, particularly at night (Hurrell et al., 2000), and in mid- to high-latitude winters when surface warming is minimal. All reconstructions fall within the uncertainty range for both versions of the Hadley Centre GCM. Both models fail, however, to capture the full magnitude of the decreasing trend, or its overall temporal evolution. The reconstructed relative upper tropospheric cooling trend is primarily due to increases in anthropogenic greenhouse gases for both models, with little contribution from either sulphate aerosols, or volcanic eruptions. This may appear counter-intuitive, but the diagnostics considered here concentrate upon extra-tropical regions, whereas the majority of the relative tropospheric warming under anthropogenic forcings in the full-field global means in both models arises in the tropics. Relative trends in other regions are more uncertain and may be negative on a global mean basis.

6.7 Lower troposphere lapse rates

There has been much recent interest in lower tropospheric (both MSU and radiosonde) and near-surface temperatures, and more specifically differences between their global mean series over the period 1979 to date (NRC, 2000, Santer et al., 2000, 2001). The near-surface has been warming at a faster rate than the lower troposphere, at least in a global-mean sense, over this period, a finding contrary to most, but not all (Santer et al., 2001), GCM predictions of the response to anthropogenic forcings. Previous studies which have considered this difference have attained closer agreement between individual level trends by accounting for factors including coverage differences, natural internal climate variability, volcanic influences, and ENSO influences, amongst others (see Santer et al., 1999, 2000, 2001 for example). Despite these efforts to date, a degree of ambiguity remains between the two temperature series that cannot be entirely satisfactorily explained. In this section the detection algorithm is applied to the difference field between a version of the lower tropospheric temperature series, over the longer period of 1960 to 1995, and the near-surface temperature series. Over this longer period the discrepancy between the series is reduced, at least in global-mean terms (Angell, 2000, Gaffen et al., 2000, Jones et al., 1997), implying relative lower tropospheric warming early in the period. The detection algorithm also considers spatio-temporal changes solely at the largest space and longest timescales and, therefore, may have advantages over previous approaches to reconciling observed differences in only considering those space and time scales at which there is confidence in model skill (Stott and Tett, 1998, AS01).

Results from detection analyses for lower troposphere lapse rates in Table 6.7 yield essentially no useful information. A sulphate aerosol signal is detected under a TLS approach for a subset of input signal combinations in HadCM2, whilst there are absolutely no cases of any signals being detected in HadCM3. Even though signals are not detected, they may still be present in the observations despite being statistically indistinguishable from natural internal climate variability, most likely due to low SNRs (Tables 6.8 and 6.9). In other tropospheric temperature variables, anthropogenic, and to a lesser extent, volcanic influences have been found to be

important in explaining the observations. The same input variable combinations are therefore considered in greater depth here as in earlier analyses (GS + VOL, G + S, G + S + VOL), as it is believed that they are most likely to explain observed trends in lower troposphere lapse rates.

Detection traces for lower troposphere lapse rates derived from HadRT2.1s lower tropospheric layer average temperatures are shown for HadCM2 in Figure 6.20. For GS, when considered in combination with VOL, there is detection at small truncations for the 10-area "smart" LAA diagnostic, and the signal is generally consistent with the observations, although tending to be overestimated in amplitude by HadCM2. This result is not repeated for the other three LAA diagnostics considered, which yield a negligible or negative best-guess amplitude estimate for the HadCM2 GS signal. This casts doubt upon the seemingly encouraging result for the preferred 10-area "smart" LAA diagnostic. There is no associated decrease in the SNR values (Tables 6.8 and 6.9), so this is highly unlikely to be due to any systematic bias in the estimator between the LAA diagnostics. In none of the LAA diagnostics considered is there any evidence for a HadCM2 predicted volcanic signal in observed lower troposphere lapse rates. This result is at odds with the analysis of Santer et al. (2001), which points towards a definite volcanic effect on the difference field from 1979 to present. Nor can any confidence even be attached to the observations being consistent with the HadCM2 VOL signal. It may be that the timescales being considered in this study are sub-optimal for detecting volcanic effects (Stott et al., 2001). This should be obvious in analyses for other input temperature diagnostics. It is difficult to make definitive statements on this point which are objective rather than subjective in nature without undertaking a sensitivity study similar to that of Stott et al. (2001). Consideration is also being given here to a longer timescale than that of Santer et al. (2001), for which before 1979 there was relatively little explosive volcanic activity (only Agung in 1963). Therefore, the HadCM2 VOL signal may be significantly noise contaminated in some of the 5-year periods considered, especially early in the period, even though the overall SNR values are large (Tables 6.8 and 6.9). In splitting the HadCM2 GS field into its G + S components, results for the 10-area "smart" LAA diagnostic yield positive best-guess amplitude estimators for both signals which are always consistent explanations of the observations, but never detected. Results from other LAA diagnostics cast doubt

upon this result as they yield estimators that are essentially zero in most cases for both HadCM2 signals, especially at higher truncations. Evidence from SNR values again points to this systematic difference being real (Tables 6.8 and 6.9). Principal results remain unchanged if all three signals are considered simultaneously within the regression. The choice of HadRT version employed to derive the lower tropospheric temperatures used in the calculation also has little impact upon the results for HadCM2, although there is an increase in failure of the consistency test on the residuals for HadRT2.1.

The analysis of detection traces for lower troposphere lapse rates is repeated for HadCM3 fields considering lapse rates derived using HadRT2.1s lower tropospheric temperatures in Figure 6.21. In agreement with lower troposphere lapse rates analysis for HadCM2, results for the 10-area "smart" LAA diagnostic are anomalous compared to the other three input LAA diagnostics. For the 10-area "smart" LAA diagnostic there is some confidence in the presence of GS, VOL, G, and S signals in the observations. Although none of these signals are detected, they are in the majority of cases consistent explanations of the observations and generally have positive best-guess amplitude scaling estimates. All the signal responses tend to be overestimated in amplitude for HadCM3 signals according to this 10-area "smart" LAA diagnostic. For all other input LAA diagnostics, best-guess amplitude estimates for each of the HadCM3 signals are essentially zero, and the signals can be effectively discounted. Both S and VOL signals are indistinguishable from noise according to SNR values in most cases (Tables 6.8 and 6.9) and, therefore, results from OLS regression for these signals may be significantly negatively biased. The lack of agreement between the LAA diagnostics reduces confidence in the presence of any of the HadCM3 signals being considered in the observations. Results are similar when HadRT2.1 fields are used in place of HadRT2.1s fields as the lower tropospheric temperature series being used to derive the observed lower troposphere lapse rate estimates, although with more cases of inconsistent residuals.

Global-mean reconstructions of lower troposphere lapse rates for both models are shown in Figure 6.22. Projection of the observations onto the leading modes of variability in each model yields a discontinuous relative cooling of the lower troposphere, after an initial warming (which is based upon early HadRT records in

which there is lower confidence and, therefore, may be spurious). There is evidence for a large scale cooling of lower troposphere lapse rates in the late 1970s, in agreement with entire troposphere lapse rate analysis. Those HadCM2 reconstructions which incorporate a volcanic forcing component can be discounted as an adequate explanation, as the volcanic forcing scaling factor is negative (unphysical) at the truncation considered in both reconstructions (Figure 6.20). The G + S reconstruction for HadCM2 yields an adequate explanation of the observations, although it fails to pick out details of the temporal evolution of lower troposphere lapse rates. HadCM3 generally better captures observed trends than HadCM2, especially when volcanic effects are incorporated, which helps reproduce some of the details, giving limited support to the findings of Santer et al. (2001). The majority of the observed relative lower tropospheric cooling according to this analysis in both models arises from the effects of greenhouse gases. This appears contrary to commonly reported results from most GCMs, which indicate that greenhouse gases should cause relative tropospheric warming. The reconstruction here based upon 10-area "smart" LAAs is highly geographically biased towards mid- to high-latitude continental regions where the differential warming rate due to greenhouse-gases is less certain and may well be of opposite sign to the entire model field global mean trend which is dominated by relative lower tropospheric warming in the tropics.

6.8 Discussion

In sections 6.2 to 6.7, results were considered in isolation for each of the six tropospheric temperature variables in turn. This section is an attempt to tie all these individual strands together to form a balanced opinion as to the most likely causes of recently observed tropospheric temperature changes according to both HadCM2 and HadCM3. It also begins to address residual uncertainties. In section 6.8.1, potential reasons for the strong biases observed in the results of consistency tests on the residuals by choice of HadRT dataset version are considered. Section 6.8.2 summarises the results of tropospheric temperature detection studies for each model in turn, to assess what meaningful conclusions regarding both the causes of recent climate change, and the ability of the models can be made.

6.8.1 The most likely causes of differences in detection results between HadRT versions

Two troposphere only versions of the HadRT dataset were considered in this study; HadRT2.1s which is uncorrected for suspected tropospheric inhomogeneities in individual grid-box records, and HadRT2.1 where these have been corrected post-1979 with reference to MSUc (Christy et al., 1998) temperature records. Any correction applied simply involves a shift in the means of any individual grid-box series identified as erroneous to create a "homogenised" grid-box series (Parker et al., 1997). No attempt is made to implicitly maintain any spatial field consistency measure within the corrections. It may, therefore, be expected that the two different HadRT versions considered would yield somewhat different detection results in any spatio-temporal detection study for the various temperature diagnostics within which the HadRT series is present. In fact, analyses yield that the signal amplitude estimates resulting from an OLS regression approach are remarkably similar between the HadRT versions for all input temperature variables, and for both models. The sole noticeable difference in results between the HadRT versions is the frequency with which the consistency test on the residuals of the regression fails.

For those temperature variables including a HadRT component, except free troposphere lapse rates, the observed input data are either HadRT, or its difference field with the relatively well-constrained HadCRUTv near-surface temperature dataset. For these temperature variables, HadRT2.1 (corrected) observed input data fails the consistency test on the residuals more often than HadRT2.1s (uncorrected). The most likely reason for this increased failure rate is that, at least some of the corrections applied to HadRT2.1 must be completely different from any of the leading modes of spatio-temporal variability in either HadCM2 or HadCM3. Such corrections would inflate the residuals, leading to an increased frequency of consistency test failure. Quality control procedures applied solely in the temporal sense (Parker et al., 1997) have, in all probability, reduced the spatio-temporal coherency of the HadRT dataset. Future quality control analysis and subsequent corrections applied to the next generation of radiosonde products should aim to

address the full spatio-temporal field consistency in a more rigorous manner. This is left to future work.

For free troposphere lapse rates the result is reversed, with applications of the OLS algorithm using HadRT2.1s observed data yielding a greater frequency of residual test failure than those using HadRT2.1. In this case only, the input field is the difference between two levels of HadRT data. Corrections applied by Parker et al. (1997) are likely to have mainly reduced vertical errors within individual grid-box series, rather than spatial (latitude-longitude) errors in the record. Expectations must be that free troposphere lapse rates would, therefore, fail the consistency test on the residuals less frequently for HadRT2.1 than for HadRT2.1s. This would continue to be the case even if those individual temperature records from which the difference field was calculated were in error, as long as the error structures were similar between the two fields, essentially cancelling each other out.

Additional non-negligible errors are likely to remain in both versions of the HadRT dataset considered here. According to this analysis, this is especially true for upper tropospheric temperatures, in agreement with previous studies of the likely sources of errors in the radiosonde record (Parker and Cox, 1995, Gaffen et al., 2000a). However, the presence of model errors at this height cannot be entirely ruled out as an explanation for at least part of the increased frequency of residuals test failure, especially for HadCM3 (section 6.2).

6.8.2 A summary of principal tropospheric temperature detection study results for two state-of-the-art climate models

A subjective ranking system is applied to summarise the results gained in sections 6.2 to 6.7 for each of the models, to yield six sets of estimates for each model, one for each tropospheric temperature variable considered in this study. Rankings range from VL to VH, with VL implying very little or no confidence, and VH implying very high confidence. Although the summary is subjective in nature, it is built upon the objective analysis in sections 6.2 to 6.7. Therefore, although others may choose to differ over exact values, the general pattern would remain constant if more

individuals were to repeat the exercise. Analysis relates solely to those signal input combinations that were considered in further detail in each section. Both stratospheric ozone depletion, and solar influences, are effectively discounted in this summary as they were only ever detected with a very low confidence at best, or, for stratospheric ozone depletion, when combined in a fixed ratio with other signals and in any of the input temperature variables. Potential causes of the recently observed changes in tropospheric temperature variables considered are therefore GS (a fixed ratio of G and S), G, S, and VOL. Two properties are considered for each signal: whether the signal is detected, and whether it is a consistent explanation of the observations. A successful detection and attribution study (one undertaken using an adequate model) would place high confidence in both of these evaluating to true for those forcings causing the observed changes. Detection need not occur with high confidence in all atmospheric variables being considered, as in some cases the uncertainties in the OLS estimators may be sufficiently large for the amplitude estimate to be indistinguishable from zero. However, in all cases the signals should be consistent explanations of the observations if the model is truly adequate. Two further properties are considered for each model: whether residuals of the OLS regression are generally consistent, and how well observed global-mean temperature trends are reconstructed. An adequate model would again place high confidence in both of these properties.

6.8.2.1 HadCM2

Results for the criteria outlined above are summarised for HadCM2 in Table 6.10 for all six input tropospheric temperature variables. The probability of successful signal detection exhibits marked differences between input temperature variables for all signals. Confidence in signal detection is lower for lapse rate diagnostics, and particularly those which include a near-surface temperature component. This is likely to be due primarily to lower SNRs leading to more uncertain, and possibly negatively biased, estimators in these lapse rate diagnostics (Tables 6.8 and 6.9). Importantly, some (albeit small) evidence remains for the presence of detectable anthropogenic forcing influences in all three lapse rate diagnostics for HadCM2. On the basis of an overall assessment for all six tropospheric temperature variables, confidence is:

- high in the presence of a well-mixed greenhouse-gases signal (either alone or when combined with sulphate aerosols in a fixed ratio)
- medium to low in the presence of a detectable sulphate aerosol signal
- low to very low in that of a volcanic signal
- very low to zero in solar and stratospheric ozone depletion effects

Advancing to consider whether the signals are consistent explanations of observations, for layer average temperature variables there is a marked difference in results between the free troposphere and near-surface temperatures. Within the free troposphere, GS (or its component parts G + S) is (are) a consistent explanation of the observations, whereas in contrast at the near-surface HadCM2 is tending to significantly overestimate the amplitude of the observed temperature response. This pattern is repeated, although more weakly, in lapse rate diagnostics, with free troposphere lapse rates being correctly diagnosed within HadCM2, but those involving a near-surface component yielding possible model signal amplitude overestimates. Reasons for this systematic bias could relate to:

- signal degeneracy
- natural internal climate variability and other sources of uncertainty within the detection algorithm
- residual observational / model errors
- lower SNRs leading to negatively biased estimators

Explicit tests made for signal degeneracy imply that this should not be a problem, so this explanation is heavily discounted. Expectations are that OLS regression solutions for the different input variables will not perfectly overlap even for a perfect model due to natural internal climate variability, finite ensemble effects, and optimisation and hypothesis testing being based upon finite sections of control, amongst others. Therefore the observed discrepancy, in terms of signal consistency with the observations, might be nothing more than a result of these uncertainties. It should be possible to design a quantitative test that could determine whether the observed separation in the solutions was significant compared to that expected by chance alone. Chapter 7 provides an indicative framework for one such approach. Although lower SNRs may be responsible in some cases for at least part of the

negative bias, analysis in sections 6.2 to 6.7 suggests that it cannot account for this discrepancy entirely.

If it could be proven that the causes of the discrepancy in the estimators were significant observational or model error, then it might be difficult to differentiate between them, at least without having access to an order of magnitude more independent models, and independent realisations of the observations. Significant observational error has been hypothesised in the near-surface temperature record as explaining recently-observed global-mean warming relative to that in the free troposphere (see NRC, 2000 and references therein for a summary of these previous criticisms). The oft-quoted hypothesis in such cases is that the near-surface temperature record exhibits spurious warming due primarily to urban heat island effects, which have not been factored into the analysis, whilst upper air temperatures from radiosonde and MSU records are correct. Results from this detection study suggest that HadCM2 is significantly overestimating temperature changes at the near-surface, but adequately capturing them in the free troposphere. If observational error is the cause of this systematic bias in results, then near-surface temperatures should have responded with a greater magnitude to anthropogenic forcings, i.e. a larger net warming at the surface, than that observed. This result, were it able to be proven that observational error were the cause, is at odds with previous studies which have attempted to reconcile recently observed differences between these levels. These studies tend to still yield a residual discrepancy between the series, with unexplained relative near-surface warming, over the shorter period of 1979 to date (see for example Santer et al., 2000, 2001).

Differences exist between this and previous studies which have attempted to address observed discrepancies between the free troposphere and near-surface temperature series. It should be stressed that in the present study reference is made to the longer period upper-air radiosonde temperature record. Over the entire 1960-1995 period this exhibits much less, if any, overall global-mean trend discrepancy with the near-surface temperature record (Jones et al., 1997, Gaffen et al., 2000a, Angell, 2000, Figure 6.22). This study also employs a truncated spatial representation of input temperature fields, which is further optimised such as to maximise the SNR, whereas previous inter-comparison studies have considered the raw data fields (Santer et al.,

2000, 2001 for example). Geographical coverage is also limited in the present analysis to continental, and in particular Northern Hemisphere mid-latitude continental regions. There exist known significant discrepancies between MSU and radiosonde records only in the tropics (Simon Tett, personal communication, 2001); a region effectively discounted in the current analysis, at least for "smart" LAA diagnostics.

If the observed discrepancy in regression results between the different temperature variables were solely to arise as a result of natural internal variability and other sources of residual uncertainty alone, then regression traces for different temperature variables should agree within some reasonable bounds. This holds true under the assumption that the model is an internally consistent (adequate) explanation of recently observed climate changes. Detection traces for the "smart" 10-area LAA diagnostic for all six tropospheric temperature variables are shown in Figure 6.23. Only this "smart" 10-area LAA diagnostic is considered, as it is the preferred input diagnostic. This figure provides an indication as to whether the results are in any meaningful way consistent, although it is far from perfect. There is some discrepancy between free troposphere layer average temperature diagnostic results and those for all other diagnostics for both GS and G signals. For these signals, free troposphere layer average temperature signal best-guess amplitude estimates are approximately consistent with the observations, whilst all others imply that HadCM2 is overestimating the signal response amplitude by approximately a factor of two. **There is a degree of overlap in the individual solutions, however, which implies that the observed discrepancy in results might reasonably have arisen solely by chance**. For both G and GS signals, this overlap region is predominantly bounded by zero and unity, implying that HadCM2 is likely to contain a detectable greenhouse-gas signal, but its amplitude is likely to be overestimated within the model. Results for S indicate that the signal is likely to be slightly overestimated in HadCM2, whilst those for VOL suggest it is significantly overestimated in amplitude in HadCM2. For both S and VOL, the overlap in solutions implies that HadCM2 is also a potentially internally consistent explanation of the observations considering these forcings. For these two signals, the overlap in solutions may be underestimated in its upper bounds due to low SNR values in some of the temperature variables being considered.

The observed discrepancy in amplitude estimates between the near-surface and free troposphere may alternatively be solely an artefact of the OLS estimators being employed. AS01 have previously demonstrated how for noisy or poorly defined (based upon small ensemble population) signals, OLS results, and particularly their upper confidence limits, can be negatively biased, and how the use of a TLS approach negates this problem. There is good theoretical reason to believe that at least some near-surface temperature and lapse rate signal fields might be poorly defined in HadCM2, when compared to free troposphere layer average temperature fields, due to their lower SNR (Tables 6.8 and 6.9, Jones et al., 1997, chapter 3, sections 6.2 to 6.7). The analysis of detection traces is, therefore, repeated for the three layer-average temperature diagnostics in Figure 6.24 for both OLS and TLS approaches, to investigate whether the observed discrepancy in the OLS detection results could arise solely due to this bias. Lapse rates are not considered, both because they add little information for detection purposes, and because they are more likely to contain significant observational error. AS01 show theoretically how under a TLS approach if the noise in the observations is not equivalent to that in the model then highly biased and variable estimators are likely to arise. Results for GS, G and VOL signals in HadCM2 (Figure 6.24) imply that OLS techniques are unlikely to yield highly biased signal amplitude estimators for these signals in the layer-average temperature diagnostics considered in this study. The observed discrepancy in best-guess amplitude estimates in fact increases when TLS results are considered, although there is an associated increase in respective amplitude estimator uncertainty limits when implicitly taking into account signal uncertainty under a TLS approach (not shown here). Therefore, the TLS estimators continue to overlap despite the increased difference between the component best-guess estimators, and the model remains a potentially internally consistent explanation of the observations. For the HadCM2 S signal, TLS results are much more uncertain than OLS results in the two free troposphere layer average diagnostics. At higher truncations, there is evidence under a TLS approach that the S signal strength may be grossly underestimated in amplitude in HadCM2. However, confidence in this result is greatly reduced as the variability in the best-guess amplitude estimates implies that TLS results for HadCM2 are highly likely to be affected by non-negligible observational errors in the S signal direction.

Returning to Table 6.10, there is generally high confidence that the residuals from an OLS regression approach are consistent for nearly all temperature variables. The exception is free troposphere lapse rates, for which confidence is much lower. This is likely to relate to non-negligible residual observational errors in individual grid-box records, and is ameliorated by consideration of HadRT2.1 observations, which have been corrected on a grid-box basis for suspected inhomogeneities within the troposphere (see sections 6.5 and 6.8.1).

Finally, from Table 6.10, it is seen that HadCM2 is generally able to reproduce the observed global mean trends in all six tropospheric temperature variables to a reasonable degree. This analysis can be extended to consider full spatio-temporal fields, as shown in Figure 6.25 for near-surface temperatures for the G + S signal combination. The regression aims to recreate the optimised (pre-whitened) observations, which in this graphical representation have been re-projected onto the original spatial field. Optimisation tends particularly to dampen down temperature anomalies in Northern Hemisphere high latitudes as well as spreading them out zonally, especially for those diagnostics including a free troposphere temperature component. This latter effect is likely to relate to the reduced zonal heterogeneity of HadCM2 fields compared to the HadRT observations (chapter 3). The reconstruction (Figure 6.25) tends to reasonably capture the optimised observations' spatio-temporal near-surface temperature structure, although there are some exceptions, most notably in the late 1970s and early 1980s, when the large step-change in Northern Hemisphere mid-latitude near-surface temperatures is not captured. This may be related to observed behaviour of the AO, which is poorly defined within HadCM2 (Gillett et al., 2000a). In the late 1970s, the reconstruction also fails to capture the timing of the sign change in Southern Hemisphere temperature anomalies. Results for other layer average and lapse rate temperature variables (not shown) exhibit similar general trend agreement, but in all cases some discrepancies can be identified. Although the level of agreement of such spatio-temporal reconstructions is not explicitly quantified in any objective manner in the present analysis, these results provide some added confidence in the veracity of the overall detection of a HadCM2-predicted anthropogenic influence in the observations. However, for each temperature variable considered, the scaling on the signals to reproduce both global-mean series and full spatio-temporal fields is different. If

HadCM2 were truly adequate then there must exist a single scaling, or range of scalings, which could be applied to all the variables considered here and still adequately recreate the observations.

6.8.2.2 HadCM3

Results for HadCM3 detailed in sections 6.2 to 6.7 are summarised in Table 6.11 in the same manner as those for HadCM2 were in Table 6.10. In common with results from HadCM2, confidence in the detection of any signal is seen to decrease for HadCM3 when lapse rates are considered. Results are also broadly consistent between models in terms of the overall confidence with which the individual signals are detected, namely:

- high confidence in a detectable greenhouse-gases signal
- medium to low confidence in a detectable sulphate aerosol influence
- low to very low confidence in a detectable volcanic influence
- very low to zero confidence in detectable solar and stratospheric ozone depletion signals

Modelled anthropogenic signal responses tend to be overestimated in amplitude for HadCM3, noticeably more so than HadCM2 (Table 6.11 c.f. Table 6.10). There remains a marked difference in amplitude estimators between the near-surface and the free troposphere, with free tropospheric layer average temperatures as simulated by HadCM3 tending to be more likely to be a consistent explanation of the observations. Figure 6.26 repeats the analysis of Figure 6.23 for HadCM3, and yields very similar results. In common with HadCM2, detection results for 10-area "smart" LAA diagnostic free troposphere layer average temperatures for both G and GS HadCM3 signals consistently yield amplitude estimators that are close to unity. Results based upon all other temperature variables suggest that HadCM3 is overestimating the amplitude of the response to both of these forcings in the observations. The region of overlap in the individual solutions is greatly reduced compared to that for HadCM2, leading to lower confidence in HadCM3 being an internally consistent explanation of the observations. Estimators for S are highly uncertain compared to those for HadCM2, with very limited evidence that they are a

consistent explanation of the observations. Analysis of SNR values (Tables 6.8 and 6.9) shows that the S signal is far more likely to be significantly noise polluted in HadCM3, and so confidence in this result is much lower, being potentially an artefact of biases in the OLS estimator. Results for VOL suggest that HadCM3 significantly overestimates the amplitude of the response, if there is indeed a detectable volcanic influence upon the observations. As is the case for the S signal, the VOL signal is likely to be noise polluted (Tables 6.8 and 6.9) and, therefore, confidence in this result is reduced. The difference between layer average anthropogenic signal amplitude estimators, is in common with results for HadCM2, independent of whether an OLS or a TLS regression algorithm is employed (not shown here). Therefore, the discrepancy is likely to be at least partly real.

Compared to results for HadCM2 fields, HadCM3 is markedly less likely to yield consistent residuals for those temperature variables which include an upper tropospheric component (Table 6.11, c.f. Table 6.10). The implication is that HadCM3 is likely to be grossly mis-representing natural internal climate variability in this region, most likely by underestimating the magnitude of variability in its leading modes (see earlier discussions in sections 6.2 and 6.6, also Collins et al., 2001). In all other temperature variables, confidence is either high, or very high, that the residuals of the regression are consistent.

From Table 6.11 it can also be seen that HadCM3 tends to adequately recreate the observed global mean trends in all temperature variables considered. The exception is for free troposphere lapse rates, where at the truncation being considered the OLS signal estimators are negative for at least one signal in each input signal combination considered, which on purely physical grounds cannot be correct. In common with HadCM2, spatio-temporal reconstructions of the observed input temperature fields are seen to be generally adequate, although in all cases at least some discrepancies can be found. Again, free troposphere lapse rate fields can be dismissed as being adequate explanations of the observed spatio-temporal changes for the same reasons as given above for global mean plots.

6.9 Conclusions

In this study, a number of tropospheric temperature climate indicators have been considered simultaneously under a common detection and attribution approach for the first time. Output from HadCM2 and HadCM3 models has been used to assess the likely impact of model uncertainty upon the results. Numerous sensitivity studies were additionally applied in an attempt to ensure against ambiguous conclusions being reached.

Corrections applied to the HadRT2.1 upper-air radiosonde temperature series used here (Parker et al., 1997) have been found to be sub-optimal in not explicitly considering the full spatio-temporal consistency of the resulting dataset. Evidence points towards at least some of these corrections not resembling the leading modes of spatio-temporal variability in either HadCM2 or HadCM3. However, the corrections are likely to have reduced the vertical errors within at least some individual grid-box series.

For both HadCM2, and HadCM3, there is high confidence in a detectable anthropogenic, and in particular a well-mixed greenhouse-gases, influence upon a range of tropospheric temperature variables over the period 1960-1995. Evidence for a sulphate aerosol influence is slightly more ambiguous. There is also much lower confidence in a detectable volcanic forcing influence. No compelling evidence exists for either a detectable solar or stratospheric ozone depletion effect. Results are to first order model independent in terms of the signals detected, providing increased confidence in these results, although the two versions of the Hadley Centre model employed are not entirely independent. It would be desirable to repeat the analysis here on other model output to confirm the principal findings.

There is a marked tendency for the response to anthropogenic forcings, especially well-mixed greenhouse-gases, to be overestimated in amplitude within both models, but particularly HadCM3, for at least some tropospheric temperature variables. This is highly unlikely to be related to known biases within the preferred detection algorithm employed in this study. If a volcanic influence does exist in the

observations, then both models also tend to significantly overestimate the amplitude of this response.

Global-mean and spatio-temporal patterns of change are reasonably recreated for each tropospheric input temperature variable considered by both HadCM2 and HadCM3 when anthropogenic (and volcanic) influences are considered. The exception is HadCM3 fields for those diagnostics involving upper tropospheric temperatures. The most likely reason is that HadCM3 underestimates the magnitude of natural variability within this portion of the troposphere (Collins et al., 2001). It should be noted, however, that the scalings applied to the model fields to reconstruct the observed fields differ between the temperature variables being considered. Effectively model output is being warped by differing factors to recreate these observations for each temperature variable considered.

Assuming that the models are adequate (internally consistent explanations of the observations), then there must exist a single "true" scaling factor, or range of scaling factors, which satisfies each forcing for every tropospheric temperature variable. Any residual differences in the individual signal amplitude estimators between temperature variables for an adequate model would then be solely due to natural internal variability of the climate system, and other sources of residual uncertainty. An unbiased estimator of this "true" scaling must be realised and tested to ensure that such a solution is plausible given the component estimators, in order to be able to make meaningful quantitative statements as to the internal consistency of any model. This is left to others to advance, although an outline of one potential approach is given in chapter 7.

The current analysis was by necessity limited to a qualitative assessment of internal tropospheric model consistency, albeit based upon the quantitative measures resulting from OLS regression. The analysis was also limited to the preferred 10-area "smart" LAA diagnostic for each model. There are differences in amplitude estimates between the individual tropospheric temperature variables considered in each model, both in terms of their best-guess amplitudes, and their uncertainty ranges. However, there is no compelling evidence of any fundamental discrepancy at least for well-mixed greenhouse-gases, which are seen to be the major forcing

component in the observed global mean changes in all tropospheric temperature variables considered for both models. Therefore, for the first time this study has shown that GCM predictions are, at least potentially, internally consistent explanations of recently observed tropospheric temperature changes when forced with anthropogenic forcings. This increases confidence in the ability of the models considered in this thesis to predict future climate changes under increasing anthropogenic forcings at the largest scales.

Model signals considered

Signal	HadCM2	HadCM3
G	Well-mixed greenhouse gases considered as CO_2 equivalent concentrations	Well-mixed greenhouse-gases considered as constituent gases and with inter-active chemistry.
S	Direct effect of sulphate aerosols	Direct and first indirect effect of sulphate aerosols. Tropospheric ozone changes.
O	Stratospheric Ozone depletion	Stratospheric Ozone depletion
LBB	Response to Lean et al. (1995) solar output variations	Response to Lean et al. (1995) solar output variations
VOL	Response to Sato et al. (1993) volcanic forcing series	Response to Sato et al. (1993) volcanic forcing series

Table 6.1 Summary of the individual candidate forcings considered in this study and their acronyms (HadCM3 forcings are aliased to the HadCM2 forcings). For more information on how these forcings are prescribed within the models see chapter 1 (section 1.3), Mitchell et al. (1995), and Tett et al. (2001).

Model	HadCM2						HadCM3					
Observed Series	HadRT2.1s			HadRT2.1			HadRT2.1s			HadRT2.1		
LAA diagnostic	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area
Regression type	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS
signals												
GSO	GSO	GSO	GSO*	GSO	GSO	GSO*	GSO	GSO	GSO	GSO	GSO	GSO
GS	GS	GS	GS	GS	GS	GS	GS	GS$	GS	GS	GS$	GS
G	G$	G$	G	G$	G$	G	G$	G$		G$	G$	G$
LBB							LBB	LBB*		LBB		
VOL												
GSO + LBB	GSO	GSO		GSO	GSO		GSO	GSO	GSO*	GSO	GSO	GSO*
GSO + VOL	GSO	GSO	GSO*	GSO	GSO	GSO*	GSO + VOL	GSO + VOL	GSO* + VOL	GSO + VOL	GSO	GSO* + VOL
GS + O	GS	GS		GS	GS		GS	GS		GS	GS$	
GS + LBB	GS	GS		GS	GS		GS	GS	GS*	GS	GS	GS*
GS + VOL	GS	GS	GS	GS	GS	GS	GS + VOL	GS + VOL	GS + VOL	GS + VOL	GS + VOL	GS + VOL
G + S	G + S	G + S		G	G + S		G + S	G$		G + S	G$	
G + SO	G + SO	G + SO	G* + SO*	G + SO	G + SO	G* + SO*	G + SO	G		G + SO	G$	
G + LBB	G	G		G	G$			G$	G		G$	
G + VOL	G	G$	G + VOL	G	G + VOL	G + VOL		G$	G + VOL		G$	G + VOL
LBB + VOL		LBB*			LBB*		LBB*	LBB*		LBB*	LBB*	
GSO + LBB + VOL	*GSO*	*GSO*		*GSO*	*GSO*		GSO	*GSO*		GSO + VOL	*GSO*	
GS + O + LBB	*GS*	*GS*		*GS*	*GS*		GS	*GS*		GS	*GS*	
GS + O + VOL	*GS*	*GS*		*GS*	*GS*		GS + VOL	*GS + VOL*		GS + VOL	*GS + VOL*	
GS + LBB + VOL	*GS*	*GS*		*GS*	*GS*		GS + VOL	*GS*		GS	*GS*	
G + S + O	*G + S*	*G + S*		*G + S*	*G + S*		G + S	*G*		G + S + O	*G$*	
G + SO + LBB	*G + SO*	*G + SO*		*G + SO*	*G + SO*		G + SO	*G + SO*	G* + SO	G + SO	*G + SO*	G* + SO
G + SO + VOL	*G + SO*	*G + SO*	*G* + SO*	*G + SO*	*G + SO*	*G* + SO*	G + SO	*G + SO + VOL*		G + SO + VOL	*G + VOL*	G* + SO* + VOL
G + S + LBB	*G + S*	*G + S*		*G*	*G + S*		G + S	*G + S*	G + S*	G + S	*G + S*	G + S*
G + S + VOL	*G + S*	*G + S*		*G*	*G + S*		G + S + VOL	*G + S + VOL*	G + S* + VOL	G + S	*G + S + VOL*	
G + LBB + VOL	*G*	*G*		*G*	*G + VOL*			*G$*			*G*	

* = Signal sig. under-estimated (beta > 1)
$ = Signal sig. over-estimated (beta < 1)
Italics indicate degenerate combinations.
Orange indicates truncation failure before 21

Table 6.2 Detection results for all possible 1-, 2-, and 3- input signal combinations for upper tropospheric temperatures. . If a signal is detected in any particular input combination then it is recorded in this table, along with whether it is significantly over-($) or under-(*) estimated in its amplitude (inconsistent) within the model. In those cases where signals are detected but residuals are inconsistent before the maximum truncation of 21 they are highlighted in orange. Other cases of early residual consistency test failure are not indicated here.

Model	HadCM2						HadCM3					
Observed Series	HadRT2.1s			HadRT2.1			HadRT2.1s			HadRT2.1		
LAA diagnostic	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area
Regression type	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS
signals												
GSO	GSO	GSO	GSO*	GSO	GSO	GSO*	GSO	GSO	GSO*	GSO	GSO	GSO*
GS	GS	GS	GS	GS	GS$	GS	GS	GS	GS	GS	GS	GS
G	G$	G$	G	G	G$	G	G	GS	G	G	G	G
LBB							LBB*	LBB*	LBB*	LBB	LBB*	LBB*
VOL												
GSO + LBB	GSO	GSO	GSO*	GSO	GSO	GSO*	GSO	GSO	GSO*	GSO	GSO	
GSO + VOL	GSO	GSO	GSO*	GSO	GSO	GSO*	GSO	GSO + VOL	GSO* + VOL	GSO	GSO + VOL	GSO* + VOL
GS + O	GS	GS		GS	GS		GS	GS		GS	GS	
GS + LBB	GS	GS	GS + LBB	GS	GS	GS + LBB	GS	GS	GS*	GS	GS	
GS + VOL	GS	GS	GS + VOL	GS	GS	GS + VOL	GS	GS + VOL	GS	GS	GS + VOL	
G + S	G + S	G	G + S*	G + S	G	G + S*	G + S	G	G + S*	G	G	
G + SO	G + SO	G	G* + SO*	G + SO	G	G* + SO*	G + SO	G	G* + SO*	G + SO	G	
G + LBB	G$	G$		G	G			G		LBB	G	
G + VOL	G + VOL	G	G* + VOL	G + VOL	G	G* + VOL		G			G	
LBB + VOL		LBB			LBB			LBB* + VOL*		LBB	LBB* + VOL*	
GSO + LBB + VOL	*GSO*	*GSO*	*GSO**	*GSO*	*GSO*	*GSO**	*GSO*	GSO		*GSO*	GSO + VOL	
GS + O + LBB	*GS*	*GS*		*GS*	*GS*		*GS*	GS		*GS*	GS	
GS + O + VOL	*GS*	*GS*		*GS*	*GS*		*GS*	GS + VOL		*GS*	GS + VOL	
GS + LBB + VOL	*GS*	*GS*	*GS + VOL*	*GS*	*GS*	*GS + VOL*	*GS*	GS		*GS*	GS + VOL	
G + S + O	*G + S*	*G*		*G + S*	*G*		*G + S*	G		*G + S*	G	
G + SO + LBB	*G + SO*	*G*	*G* + SO**	*G + SO*	*G*	*G + SO**	*G + SO*	G + SO	*G* + SO*	*G + SO*	G + SO	
G + SO + VOL	*G*	*G*		*G*	*G*	*G* + SO**	*G + SO*	G + SO + VOL	*G*+SO* + VOL*	*G + SO*	G + SO + VOL	
G + S + LBB	*G + S*	*G*		*G + S*	*G*		*G + S*	G + S	*G + S**	*G + S*	G	
G + S + VOL	*G + VOL*	*G*		*G + S*	*G*		*G + S*	G + S + VOL	*G + S**	*G + S*	G + S + VOL	
G + LBB + VOL	*G + VOL*	*G*	*G + VOL*	*G + VOL*	*G*	*G* + VOL**		G			G	

* = Signal sig. under-estimated (beta > 1)
$ = Signal sig. over-estimated (beta < 1)
Italics indicate degenerate combinations.
Orange indicates truncation failure before 21

Table 6.3 As Table 6.2 except considering lower tropospheric temperatures.

Model	HadCM2			HadCM3		
Observed Series	HadCRUTv			HadCRUTv		
LAA diagnostic	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area
Regression type	OLS	OLS	TLS	OLS	OLS	TLS
signals						
GSO				GSO$		GSO$
GS	GS$	GS$	GS	GS$		GS$
G				G$		G$
LBB						LBB
VOL						
GSO + LBB				GSO$		
GSO + VOL				GSO$		
GS + O		GS$		GS$		
GS + LBB		GS$	GS			
GS + VOL		GS$		GS$		
G + S		G$	G + S	G$		
G + SO				G$		G
G + LBB						
G + VOL						
LBB + VOL						
GSO + LBB + VOL				*GSO$*		
GS + O + LBB		GS$		*GS$*		
GS + O + VOL		GS$		*GS$*		
GS + LBB + VOL						
G + S + O		G$		*G$*		
G + SO + LBB				*G$*		
G + SO + VOL				*G$*		
G + S + LBB		G$	S			
G + S + VOL			S			
G + LBB + VOL						

* = Signal sig. under-estimated (beta > 1)
$ = Signal sig. over-estimated (beta < 1)
Italics indicate degenerate combinations.
Orange indicates truncation failure before 21

Table 6.4 As Table 6.2 except considering near-surface temperature series.

Model	HadCM2						HadCM3					
Observed Series	HadRT2.1s			HadRT2.1			HadRT2.1s			HadRT2.1		
LAA diagnostic	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area
Regression type	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS
signals												
GSO									GSO			
GS	GS	GS	GS			GS		GS	GS			
G												
LBB												
VOL			VOL*	VOL		VOL						
GSO + LBB								GSO				
GSO + VOL				VOL								
GS + O												
GS + LBB	GS	GS						GS				
GS + VOL	GS	GS + VOL	GS + VOL*	VOL		VOL						
G + S	G + S	G			S		G + S					
G + SO							G + SO					
G + LBB												
G + VOL			VOL*	VOL		VOL						
LBB + VOL				VOL								
GSO + LBB + VOL		*GSO*		*VOL*	*GSO*							
GS + O + LBB								*GS*				
GS + O + VOL	*VOL*	*VOL*		*VOL*								
GS + LBB + VOL	*GS*	*GS + VOL*		*VOL*				*GS*				
G + S + O					*S*		*S*					
G + SO + LBB					*SO*		*G + SO*			*G + SO*		
G + SO + VOL				*VOL*	*SO*		*G* + SO* + VOL*					
G + S + LBB	*G + S*	*G + S*			*S*		*G + S**	*G + S*				
G + S + VOL	*G + S + VOL*	*G + S* + VOL*	*G + S* + VOL*	*VOL*	*S*	*G + S + VOL*	*G + S*			*G + S*		
G + LBB + VOL				*VOL*								

* = Signal sig. under-estimated (beta > 1)
$ = Signal sig. over-estimated (beta < 1)
Italics indicate degenerate combinations.
Orange indicates truncation failure before 21

Table 6.5 As Table 6.2 except considering free troposphere (UT-LT) lapse rates.

Model	HadCM2						HadCM3					
Observed Series	HadRT2.1s - HadCRUTv			HadRT2.1 - HadCRUTv			HadRT2.1s - HadCRUTv			HadRT2.1 - HadCRUTv		
LAA diagnostic	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area
Regression type	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS
signals												
GSO												GSO
GS			GS									GS
G												G
LBB												
VOL								VOL				
GSO + LBB												GSO
GSO + VOL								VOL				
GS + O			GS									
GS + LBB												GS
GS + VOL			GS					VOL				
G + S												
G + SO												
G + LBB												G
G + VOL								VOL				
LBB + VOL								VOL				
GSO + LBB + VOL								*VOL*				
GS + O + LBB												
GS + O + VOL								*VOL*				
GS + LBB + VOL								*VOL*				
G + S + O												
G + SO + LBB												
G + SO + VOL								*VOL*				
G + S + LBB			S									
G + S + VOL			G					*VOL*				
G + LBB + VOL								*VOL*				

* = Signal sig. under-estimated (beta > 1)
$ = Signal sig. over-estimated (beta < 1)
Italics indicate degenerate combinations.
Orange indicates truncation failure before 21

Table 6.6 As Table 6.2 except considering entire troposphere (UT-Surf) lapse rates.

Model	HadCM2						HadCM3					
Observed Series	HadRT2.1s - HadCRUTv			HadRT2.1 - HadCRUTv			HadRT2.1s - HadCRUTv			HadRT2.1 - HadCRUTv		
LAA diagnostic	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area	10-Area	5-Area	10-Area
Regression type	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS	OLS	OLS	TLS
signals												
GSO												
GS												
G												
LBB												
VOL												
GSO + LBB												
GSO + VOL												
GS + O												
GS + LBB												
GS + VOL												
G + S			S			S						
G + SO												
G + LBB												
G + VOL												
LBB + VOL												
GSO + LBB + VOL												
GS + O + LBB												
GS + O + VOL												
GS + LBB + VOL												
G + S + O			S			S						
G + SO + LBB												
G + SO + VOL												
G + S + LBB												
G + S + VOL			S			S						
G + LBB + VOL												

* = Signal sig. under-estimated (beta > 1)
$ = Signal sig. over-estimated (beta < 1)
Italics indicate degenerate combinations.
Orange indicates truncation failure before 21

Table 6.7 As Table 6.2 except considering lower tropospheric (LT-Surf) lapse rate

10-area "smart" LAA

	HadCM2				HadCM3			
Variable	**GS**	**G**	**S**	**VOL**	**GS**	**G**	**S**	**VOL**
UT	6.61	7.93	1.50	5.24	14.17	21.11	2.31	1.54
LT	5.47	7.69	1.78	3.19	10.52	9.64	1.72	2.66
Surf	4.30	6.15	2.47	3.00	14.17	12.90	1.63	1.54
UT-LT	2.67	3.50	1.60	2.56	3.18	4.94	2.35	1.78
UT-Surf	1.76	3.98	2.76	1.95	3.13	2.01	1.01	0.52
LT-Surf	3.32	4.81	3.17	3.22	3.97	3.03	0.68	0.52

5-area "smart" LAA

	HadCM2				HadCM3			
Variable	**GS**	**G**	**S**	**VOL**	**GS**	**G**	**S**	**VOL**
UT	6.52	8.70	1.36	5.36	21.10	23.01	2.81	4.76
LT	7.72	9.28	1.48	3.09	16.41	21.15	3.40	3.65
Surf	7.15	9.25	1.46	2.96	16.05	18.29	2.32	2.12
UT-LT	3.16	4.75	1.62	3.40	5.62	8.59	3.67	1.50
UT-Surf	1.88	4.01	2.67	1.98	3.57	2.24	1.32	0.40
LT-Surf	5.31	8.20	2.40	4.71	6.10	4.88	0.84	0.68

12-area "Naïve" LAA

	HadCM2				HadCM3			
Variable	**GS**	**G**	**S**	**VOL**	**GS**	**G**	**S**	**VOL**
UT	5.75	6.47	2.18	4.64	15.70	16.51	2.12	5.07
LT	6.01	9.52	2.91	6.02	17.29	17.53	2.92	4.35
Surf	6.67	8.44	2.60	5.33	14.26	14.24	2.11	2.88
UT-LT	2.23	3.59	1.23	2.27	4.84	7.03	3.36	2.51
UT-Surf	3.23	4.01	3.99	3.14	3.30	2.20	0.87	0.56
LT-Surf	6.59	7.04	2.29	5.92	6.49	5.67	0.78	0.88

6-area "Naïve" LAA

	HadCM2				HadCM3			
Variable	**GS**	**G**	**S**	**VOL**	**GS**	**G**	**S**	**VOL**
UT	5.13	6.13	1.95	4.37	14.47	14.81	1.75	4.87
LT	4.73	6.62	1.33	4.03	19.52	18.73	2.85	5.48
Surf	5.37	6.69	1.86	4.57	13.88	13.23	1.86	3.29
UT-LT	2.24	2.54	1.80	2.86	4.68	5.52	0.63	1.90
UT-Surf	2.63	2.32	2.84	2.28	3.51	2.77	0.70	0.59
LT-Surf	5.69	5.51	2.03	4.30	6.04	5.13	0.76	0.73

Table 6.8 SNR values at truncation 11 for the individual signals being considered. Those cases where the signal is indistinguishable from noise at the 90% confidence interval are highlighted in orange.

10-area "smart" LAA

	HadCM2				HadCM3			
Variable	**GS**	**G**	**S**	**VOL**	**GS**	**G**	**S**	**VOL**
UT	4.35	5.33	1.80	4.18	13.40	14.49	1.72	3.19
LT	4.09	4.98	2.15	3.12	7.72	8.38	1.85	1.98
Surf	3.05	4.49	2.55	2.87	8.74	7.36	1.52	1.03
UT-LT	2.01	4.27	2.49	2.26	3.66	3.85	1.98	1.37
UT-Surf	1.44	2.68	2.73	1.40	3.03	1.88	0.99	0.55
LT-Surf	2.46	3.85	2.80	3.01	4.18	3.53	0.61	0.34

5-area "smart" LAA

	HadCM2				HadCM3			
Variable	**GS**	**G**	**S**	**VOL**	**GS**	**G**	**S**	**VOL**
UT	4.68	7.02	2.46	5.19	12.20	13.87	3.04	3.33
LT	4.64	6.50	2.50	4.86	11.17	14.17	3.15	3.49
Surf	4.34	6.01	2.41	4.44	10.18	12.08	2.14	2.80
UT-LT	1.92	3.65	1.65	2.74	4.91	5.36	3.30	1.35
UT-Surf	1.75	3.58	2.60	1.89	2.49	2.38	1.47	0.48
LT-Surf	3.62	5.49	2.51	4.56	4.89	5.42	0.93	0.90

12-area "Naïve" LAA

	HadCM2				HadCM3			
Variable	**GS**	**G**	**S**	**VOL**	**GS**	**G**	**S**	**VOL**
UT	4.26	5.84	2.36	5.35	10.55	11.73	2.47	3.41
LT	4.44	6.44	2.66	4.93	10.32	10.23	2.45	3.44
Surf	4.54	6.72	2.96	4.62	9.55	9.60	2.61	2.83
UT-LT	1.57	2.45	1.71	1.84	4.30	4.45	3.03	2.36
UT-Surf	2.62	4.12	2.67	2.32	2.38	2.11	1.31	0.72
LT-Surf	4.00	5.94	2.81	4.95	4.84	3.92	0.96	0.87

6-area "Naïve" LAA

	HadCM2				HadCM3			
Variable	**GS**	**G**	**S**	**VOL**	**GS**	**G**	**S**	**VOL**
UT	4.12	5.31	2.02	3.95	8.89	8.96	1.47	3.29
LT	3.62	5.78	2.33	3.39	11.28	10.79	2.31	3.78
Surf	3.57	5.29	2.22	3.66	9.27	7.86	2.05	2.34
UT-LT	1.53	1.91	1.39	1.88	3.91	4.31	1.81	1.73
UT-Surf	2.15	3.16	2.66	1.78	2.96	2.19	1.25	0.68
LT-Surf	3.44	4.56	2.30	3.57	4.32	3.43	0.91	0.62

Table 6.9 SNR values at truncation 21 for the individual signals being considered. Those cases where the signal is indistinguishable from noise at the 90% confidence interval are highlighted in orange

Tropospheric temperature parameter	HadCM2																	
	Detection				Consistent explanation of the observations												Consistent residuals	Reproduces Global mean trends
					GS			G			S			VOL				
	GS	G	S	VOL	>	eq.	<	>	eq.	<	>	eq.	<	>	eq.	<		
Upper tropospheric temperatures	VH	VH	M	L	VL	VH	VL	VL	VH	VL	VL	VH	VL	H	VL	VL	H	VH
Lower tropospheric temperatures	VH	VH	L	L	VL	VH	VL	VL	VH	VL	VL	VH	VL	M	L	VL	VH	M
Near-surface temperatures	H	H	L	VL	M	M	VL	M	M	VL	M	M	VL	VL	VL	VL	VH	M
Free troposphere lapse rates	M	M	L	M	VL	VH	VL	L	H	VL	VL	VH	VL	H	L	VL	M	H
Entire troposphere lapse rates	M	L	L	VL	L	H	VL	L	H	VL	L	H	VL	VL	L	VL	H	M
Lower troposphere lapse rates	L	L	L	VL	L	VL	VL	L	VL	VL	VL	L	VL	VL	VL	VL	H	H

VL	No to very low confidence
L	Low confidence
M	Medium confidence
H	High confidence
VH	Very high confidence

>	Model signal response significantly overestimated
eq.	Model signal response consistent with the observations
<	Model signal response significantly underestimated

Table 6.10 Summary of principal results for HadCM2. The scoring system is subjective, but is based upon objective OLS detection analyses, and therefore the overall pattern shown is likely to be robust to other choices of scoring mechanism over a reasonable range. Table considers signal detection and consistency (attribution), regression residuals consistency, and how well the temporal trends for 1960-1995 are reproduced at the global scale.

Tropospheric temperature parameter	HadCM3																	
	Detection				Consistent explanation of the observations												Consistent residuals	Reproduces Global mean trends
					GS			G			S			VOL				
	GS	G	S	VOL	>	eq.	<	>	eq.	<	>	eq.	<	>	eq.	<		
Upper tropospheric temperatures	VH	VH	M	M	M	M	VL	H	L	VL	VL	H	VL	M	M	VL	L	H
Lower tropospheric temperatures	VH	VH	M	VL	VL	VH	VL	L	H	VL	VL	H	VL	H	L	VL	H	VH
Near-surface temperatures	H	H	L	VL	H	VL	VL	VH	VL	VL	L	H	VL	VL	VL	VL	VH	M
Free troposphere lapse rates	L	L	VL	VL	M	L	VL	M	L	VL	VL	VL	VL	VL	VL	VL	L	VL
Entire troposphere lapse rates	VL	VL	VL	VL	M	L	VL	M	L	VL	VL	VL	VL	VL	VL	VL	VH	M
Lower troposphere lapse rates	L	L	L	L	M	VL	VL	M	VL	VL	L	L	VL	L	L	VL	VH	H

VL No to very low confidence
L Low confidence
M Medium confidence
H High confidence
VH Very high confidence

> Model signal response significantly overestimated
eq. Model signal response consistent with the observations
< Model signal response significantly underestimated

Table 6.11 As Table 6.10, except considering results for HadCM3.

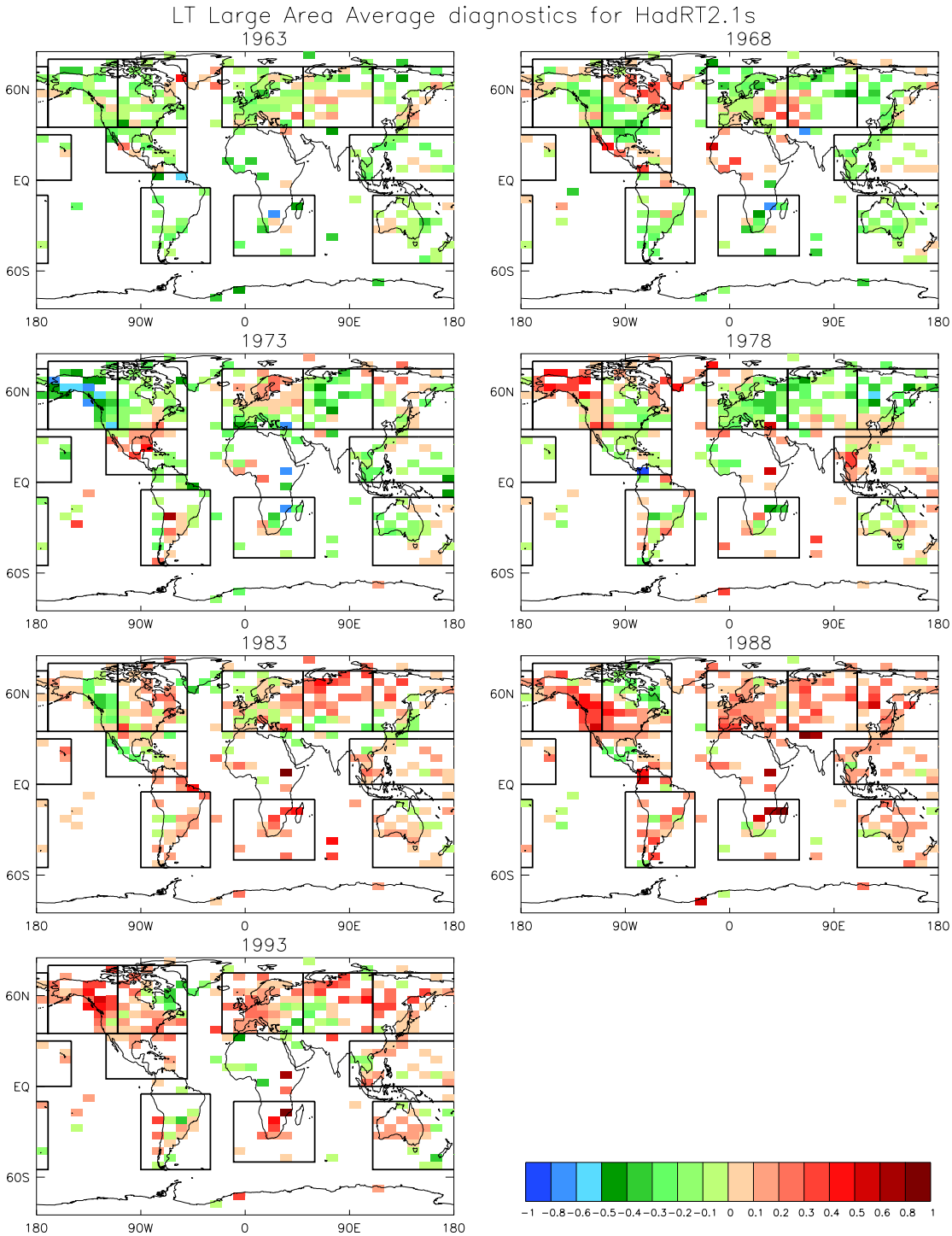

Figure 6.1 Raw lower tropospheric HadRT2.1s V2 input temperature fields. The superimposed boxes indicate the choice of regions in the preferred 10-area "smart" LAA input diagnostic used in the current study.

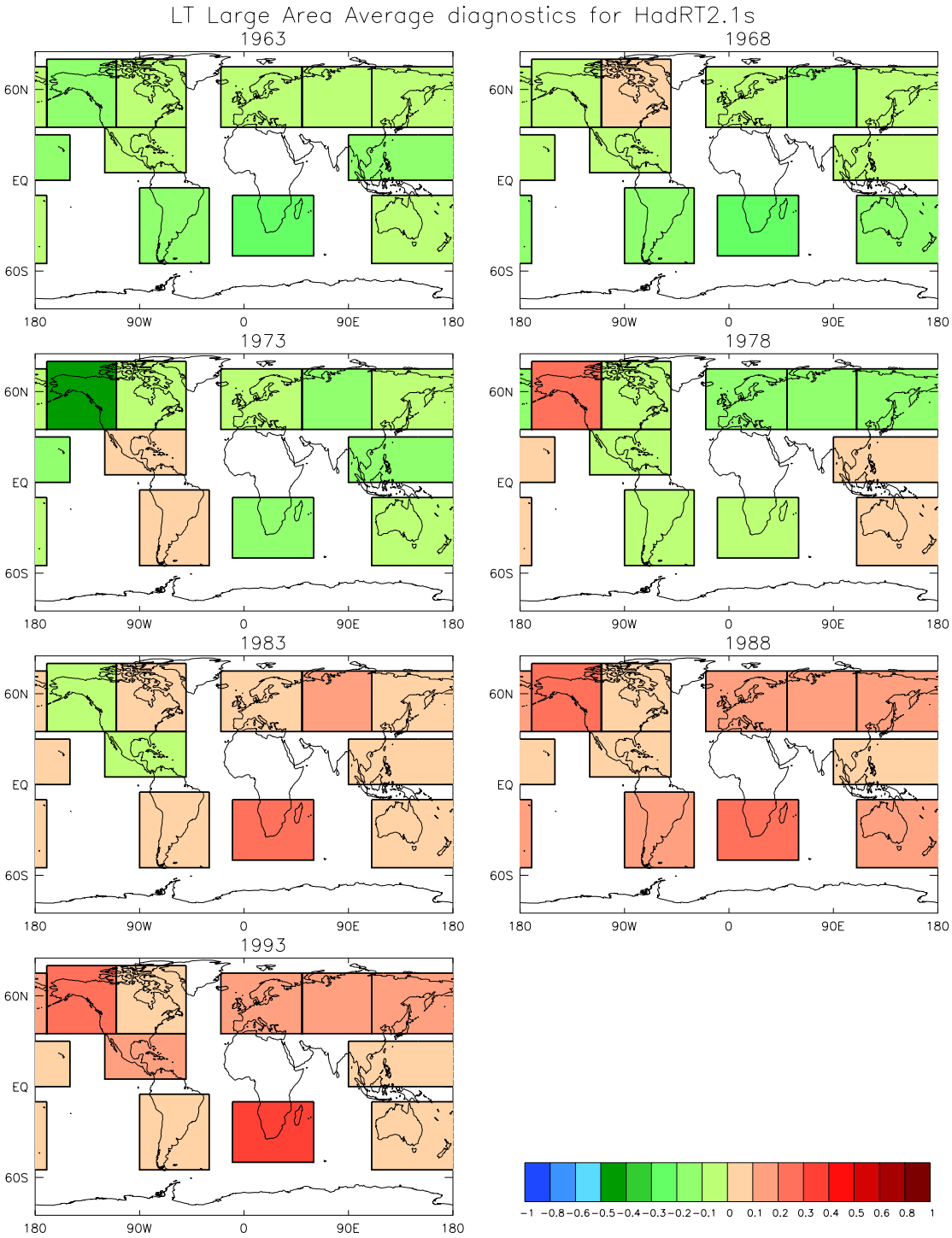

Figure 6.2 10-area "smart" LAA input diagnostic for Lower Tropospheric HadRT2.1s V2 temperatures.

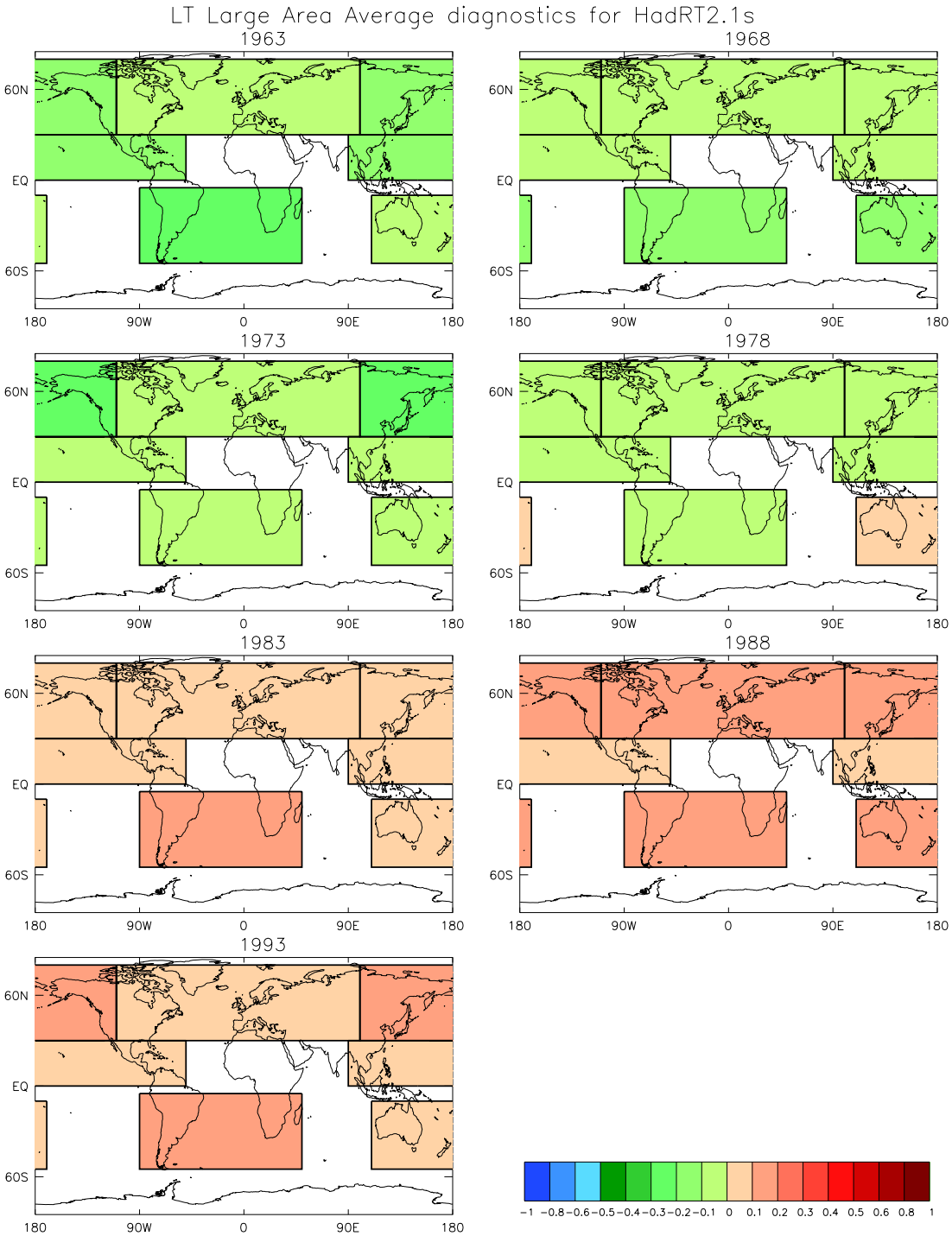

Figure 6.3 5-area "smart" LAA input diagnostic for Lower Tropospheric HadRT2.1s V2 temperatures.

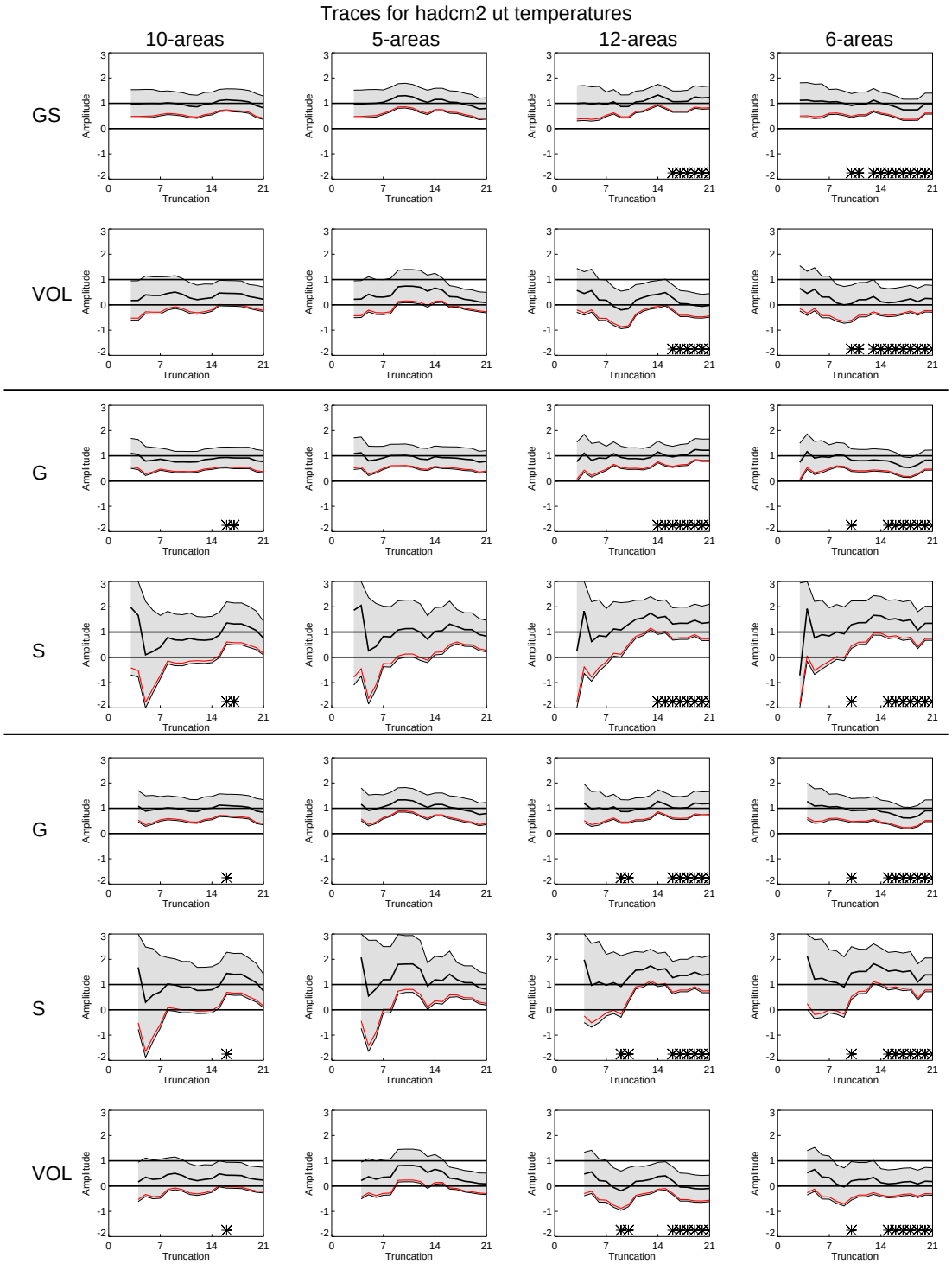

Figure 6.4 Changing beta estimates with increasing truncation for HadCM2 upper tropospheric temperatures. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

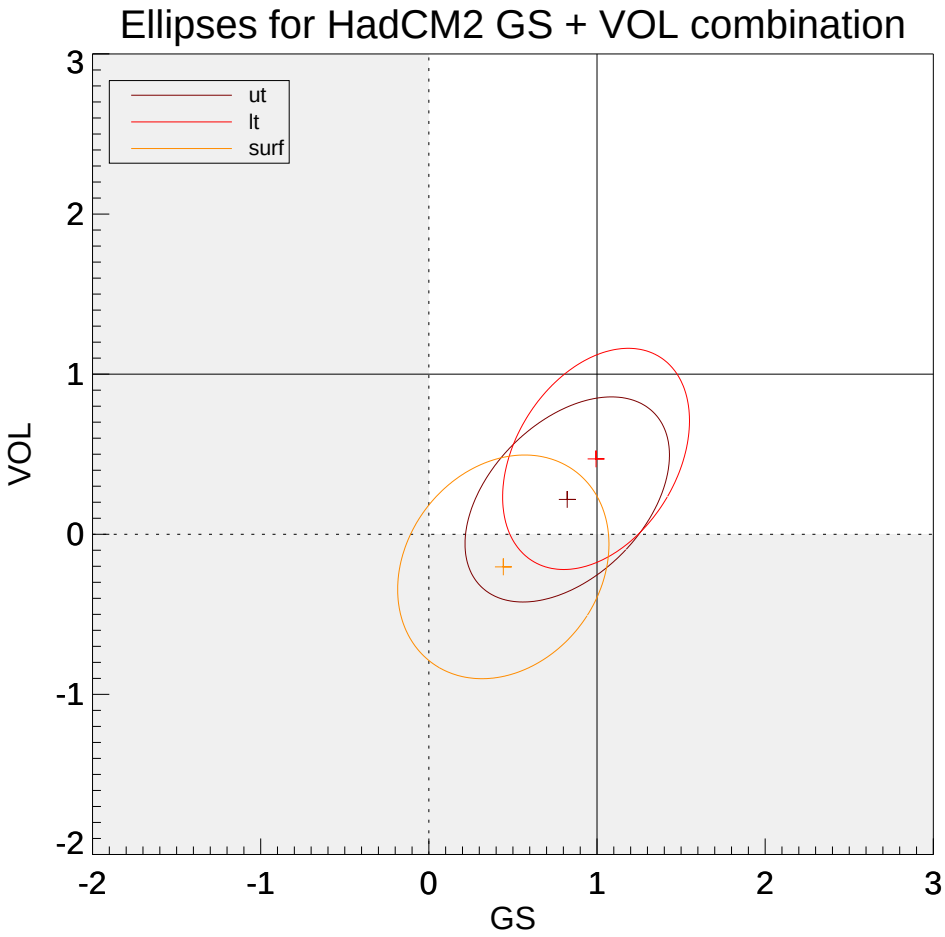

Figure 6.5 OLS regression ellipses for layer average temperature diagnostics for HadCM2 at truncation 21. The cross represents the best-guess amplitude estimator in each case, and the ellipse the 90% confidence interval as to the potential value of the true solution.

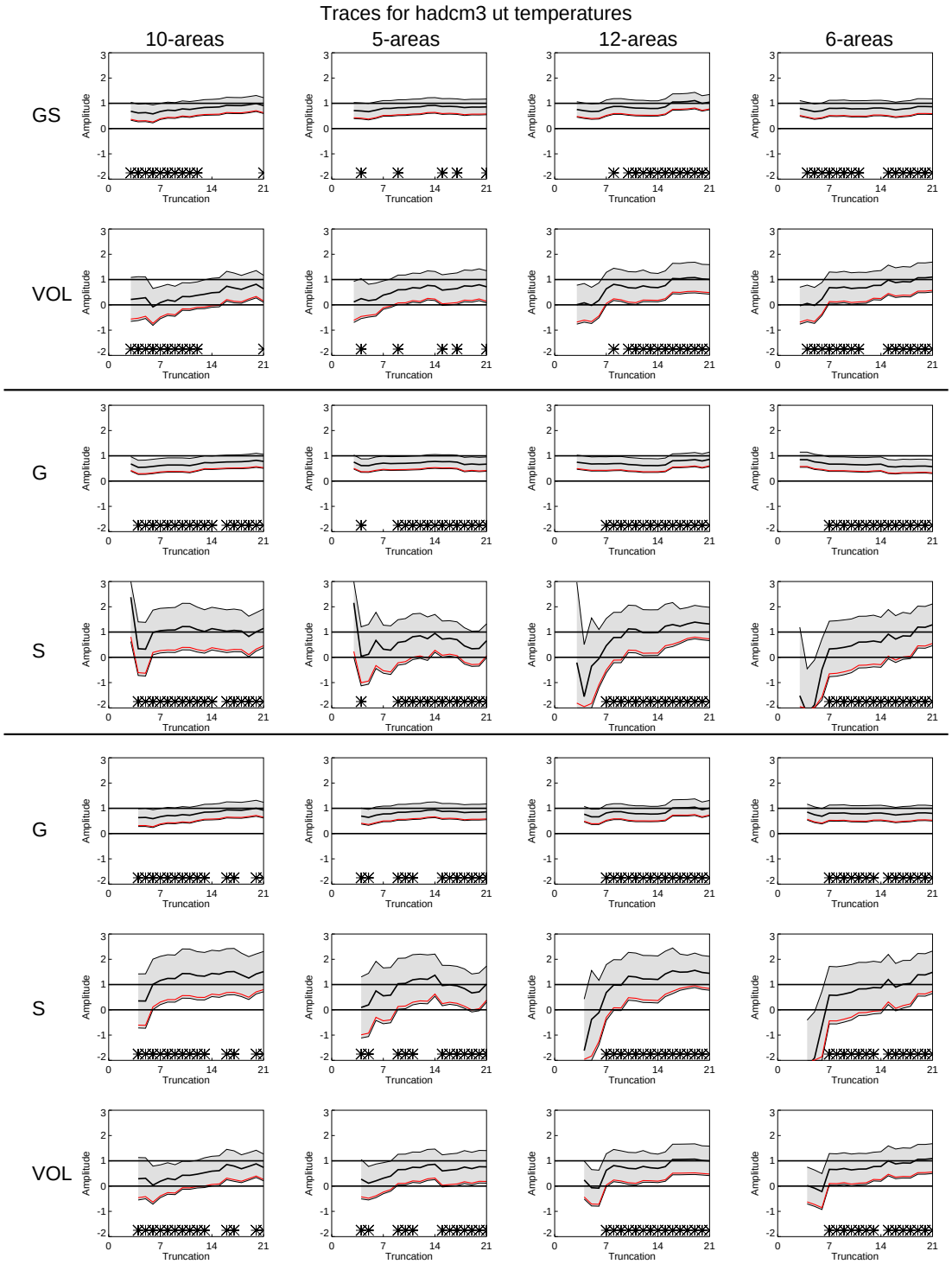

Figure 6.6 Changing beta estimates with increasing truncation for HadCM3 upper tropospheric temperatures. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

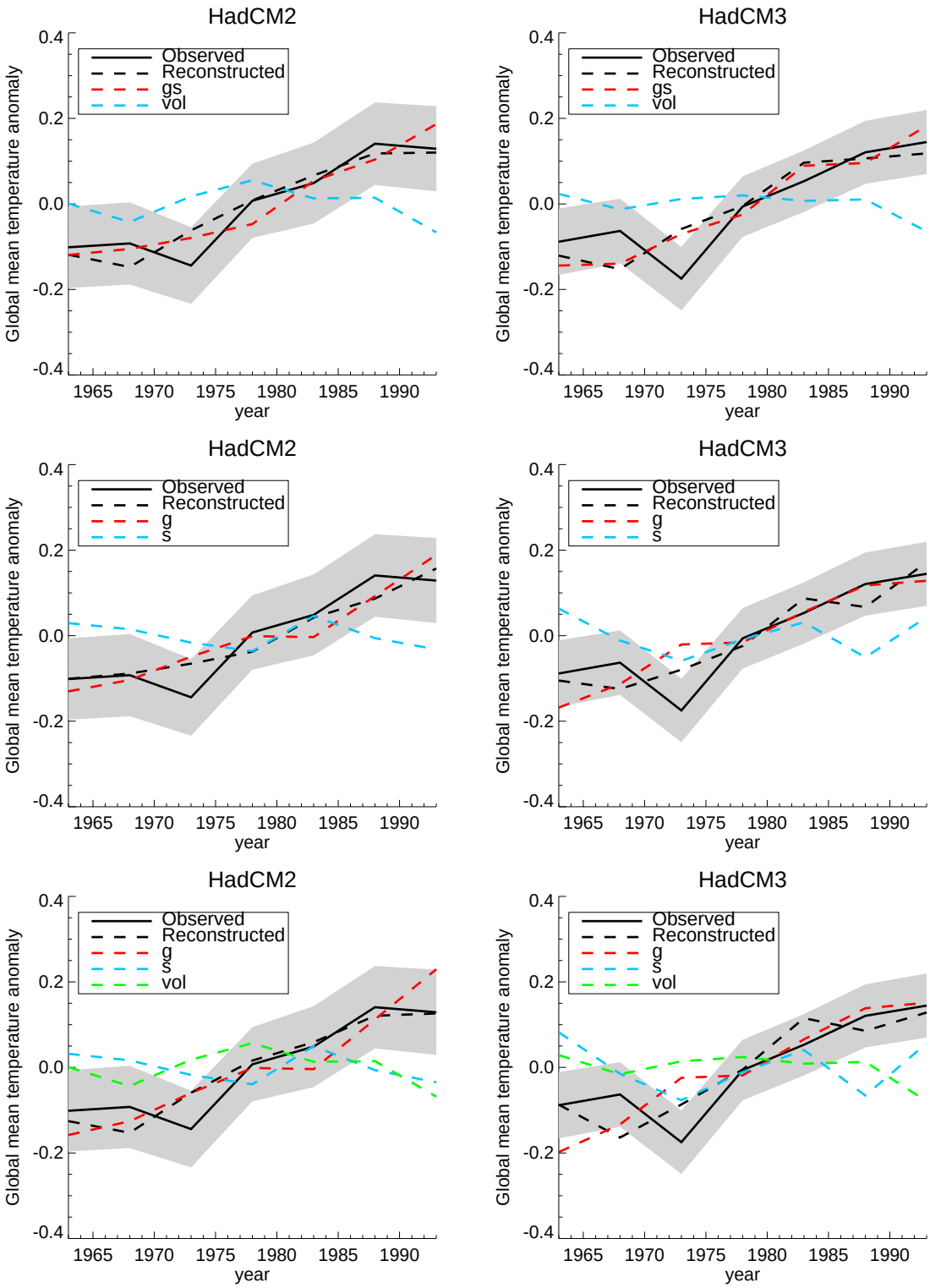

Figure 6.7 Reconstructed global mean temperature trends of upper tropospheric temperatures for HadCM2 and HadCM3. The "observations" are projections onto leading modes of model simulated internal variability, and therefore differ between models. In each case the reconstruction is based upon the signals multiplied by their best-guess amplitude estimates.

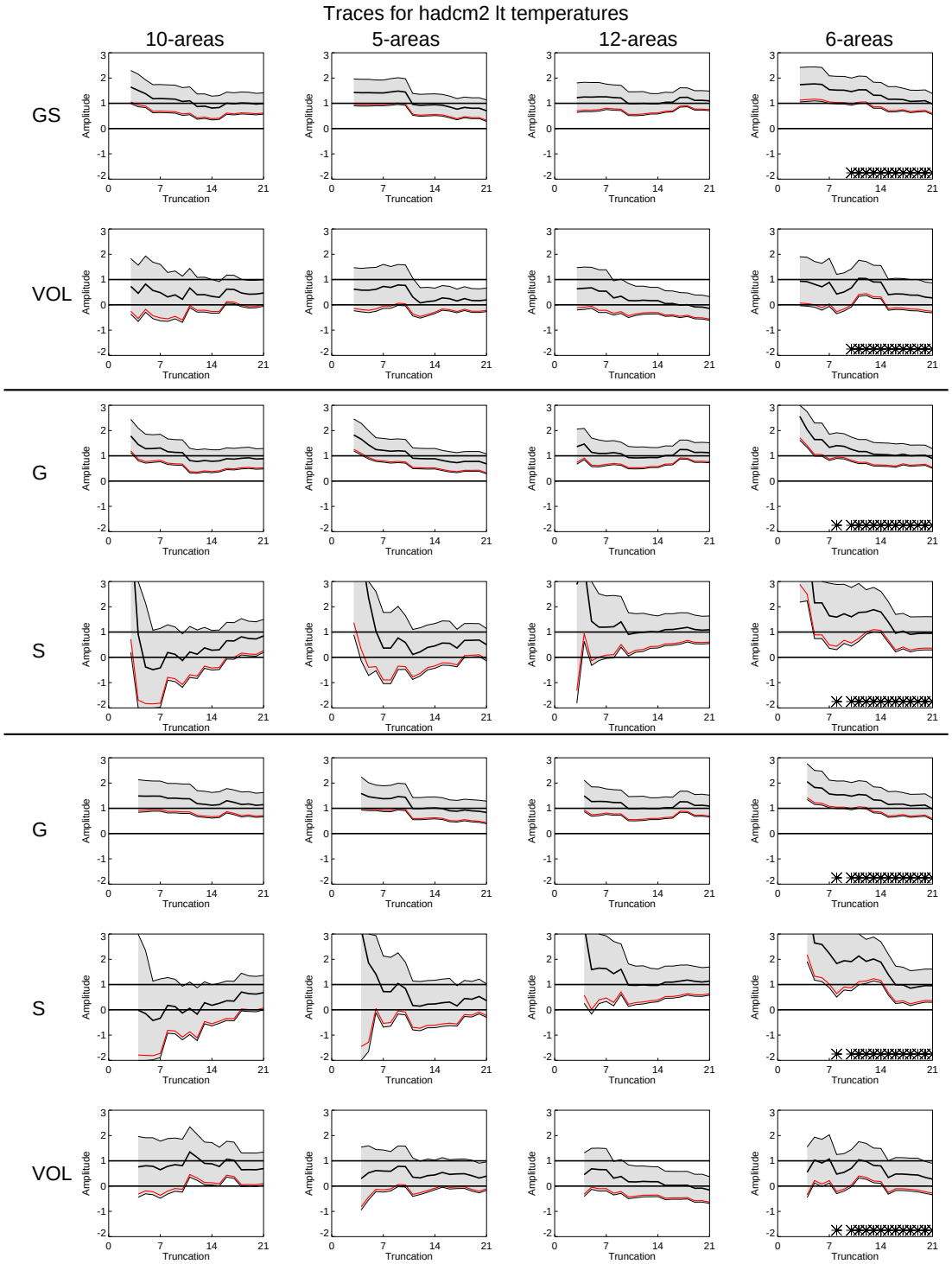

Figure 6.8 Changing beta estimates with increasing truncation for HadCM2 lower tropospheric temperatures. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

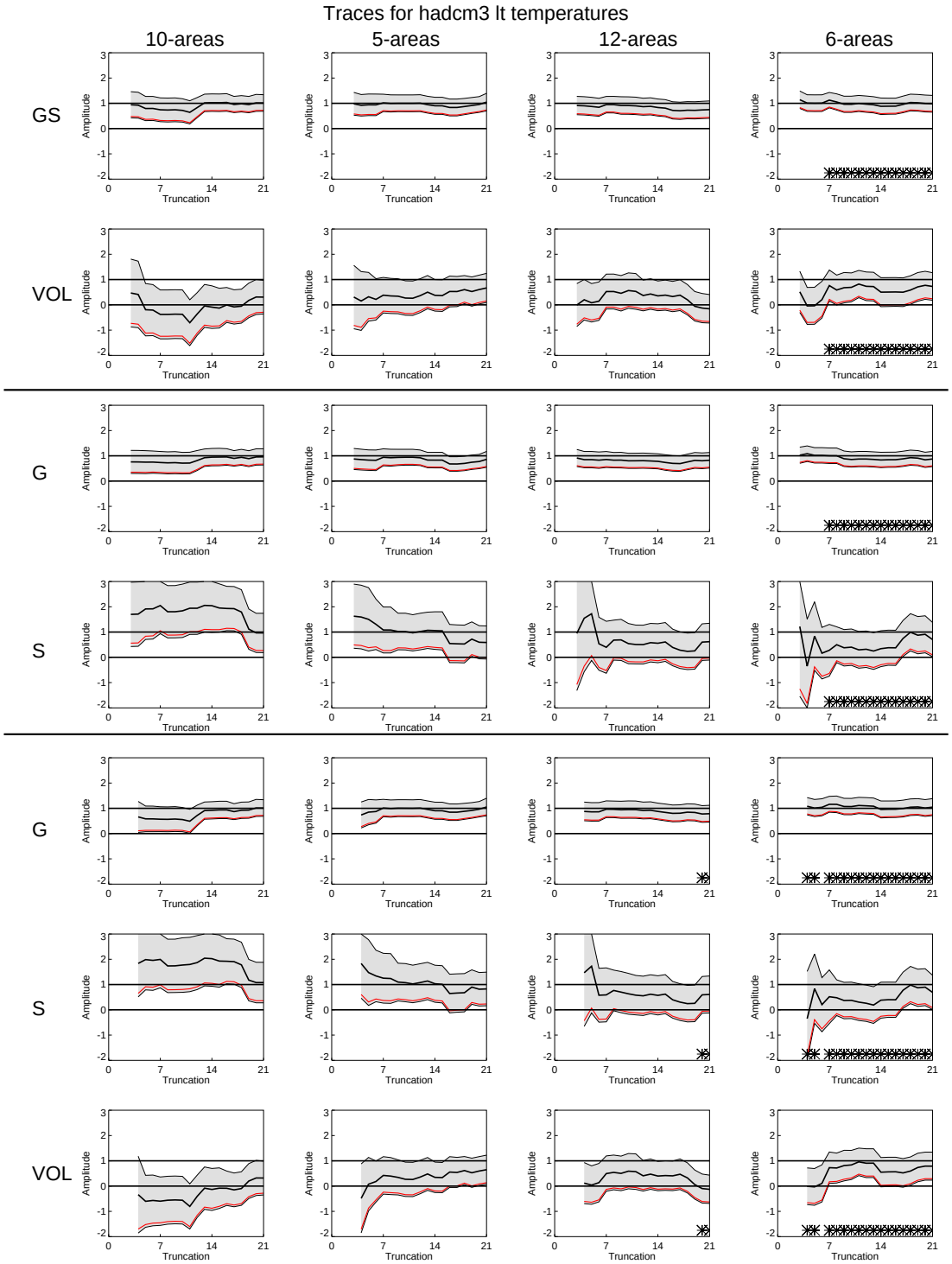

Figure 6.9 Changing beta estimates with increasing truncation for HadCM3 lower tropospheric temperatures. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

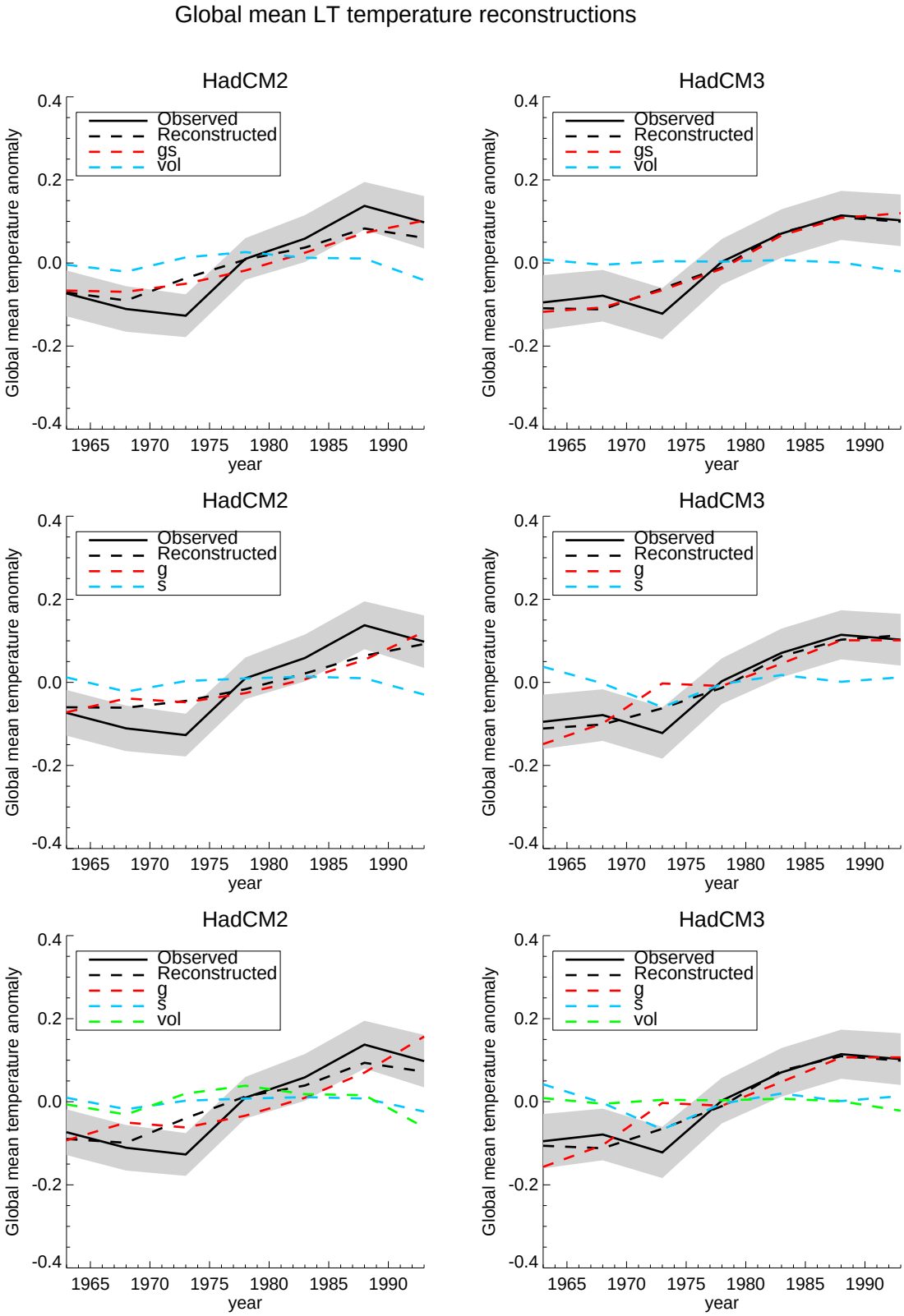

Figure 6.10 Reconstructed global mean temperature trends of lower tropospheric temperatures for HadCM2 and HadCM3. The "observations" are projections onto leading modes of model simulated internal variability, and therefore differ between models. In each case the reconstruction is based upon the signals multiplied by their best-guess amplitude estimates.

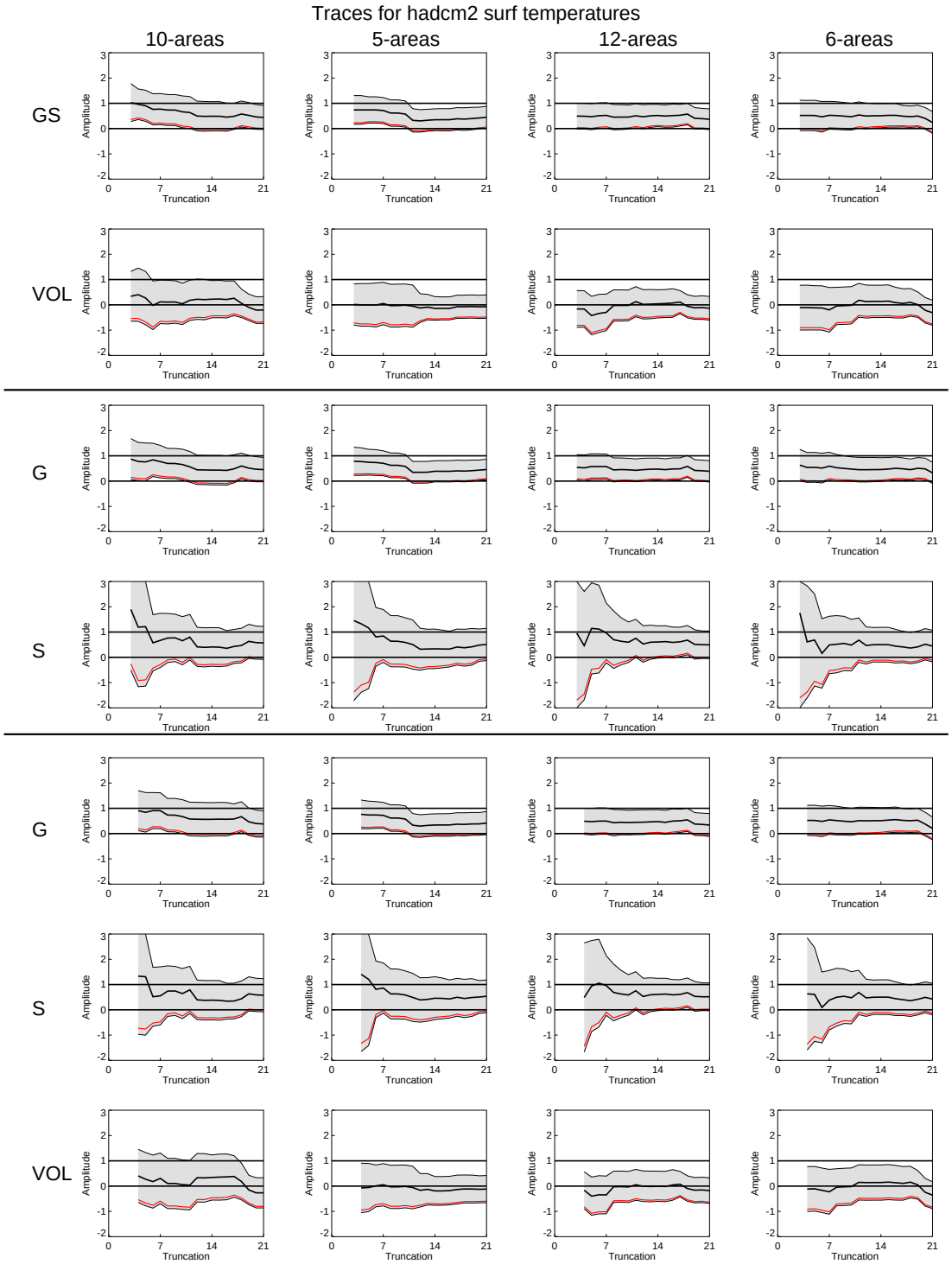

Figure 6.11 Changing beta estimates with increasing truncation for HadCM2 near-surface temperatures. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

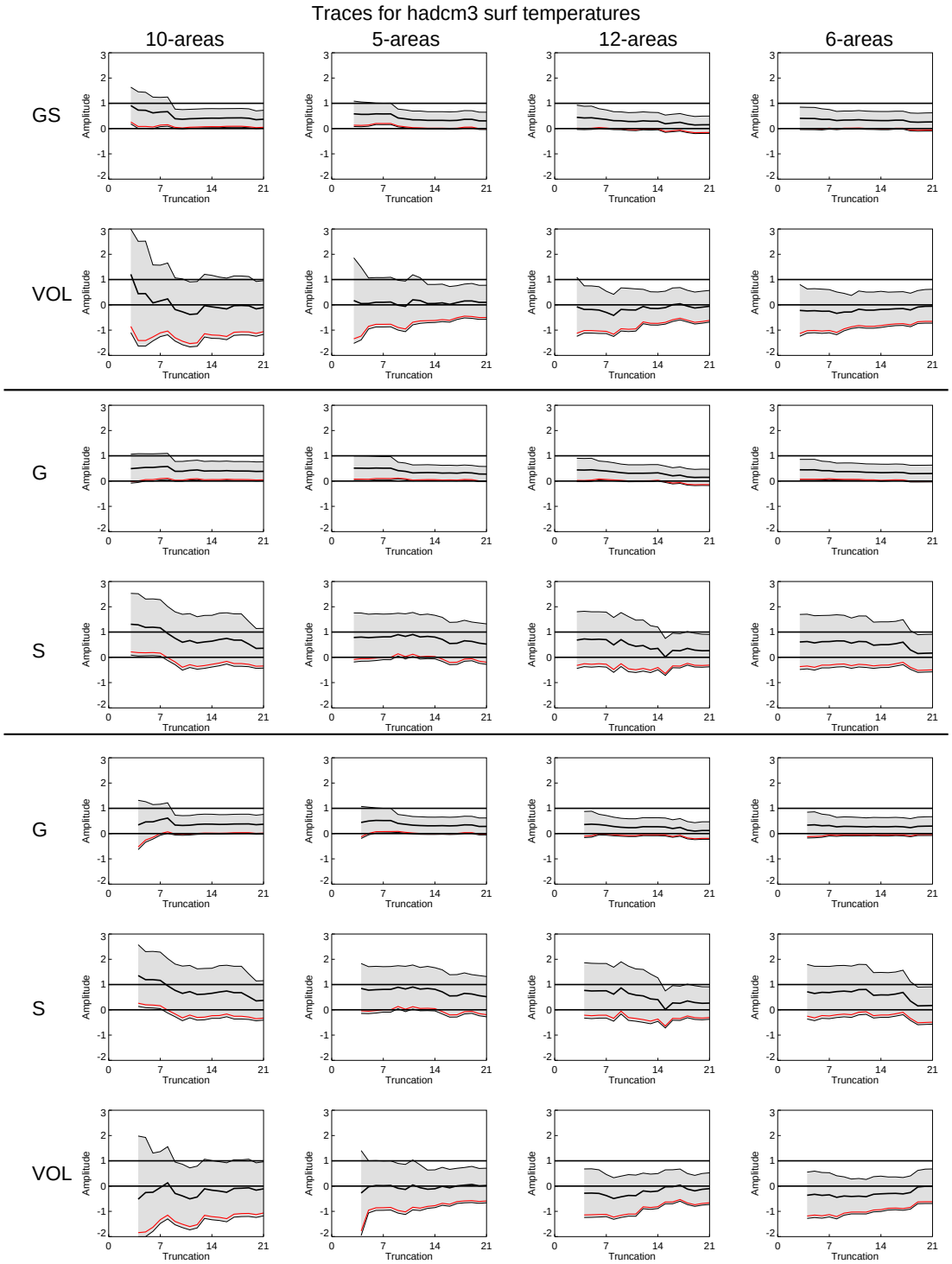

Figure 6.12 Changing beta estimates with increasing truncation for HadCM3 near-surface temperatures. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

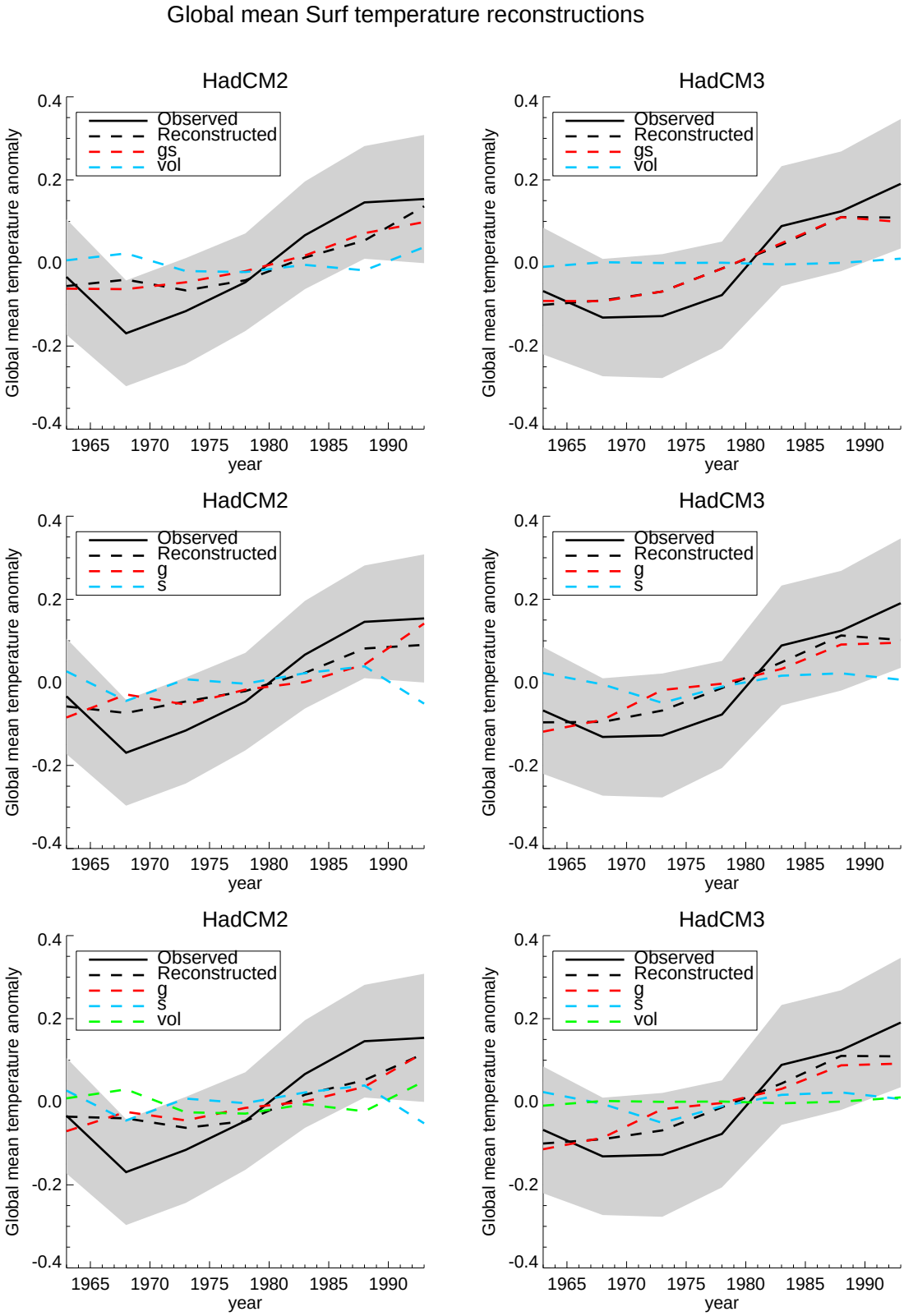

Figure 6.13 Reconstructed global mean temperature trends of near-surface temperatures for HadCM2 and HadCM3. The "observations" are projections onto leading modes of model simulated internal variability, and therefore differ between models. In each case the reconstruction is based upon the signals multiplied by their best-guess amplitude estimates.

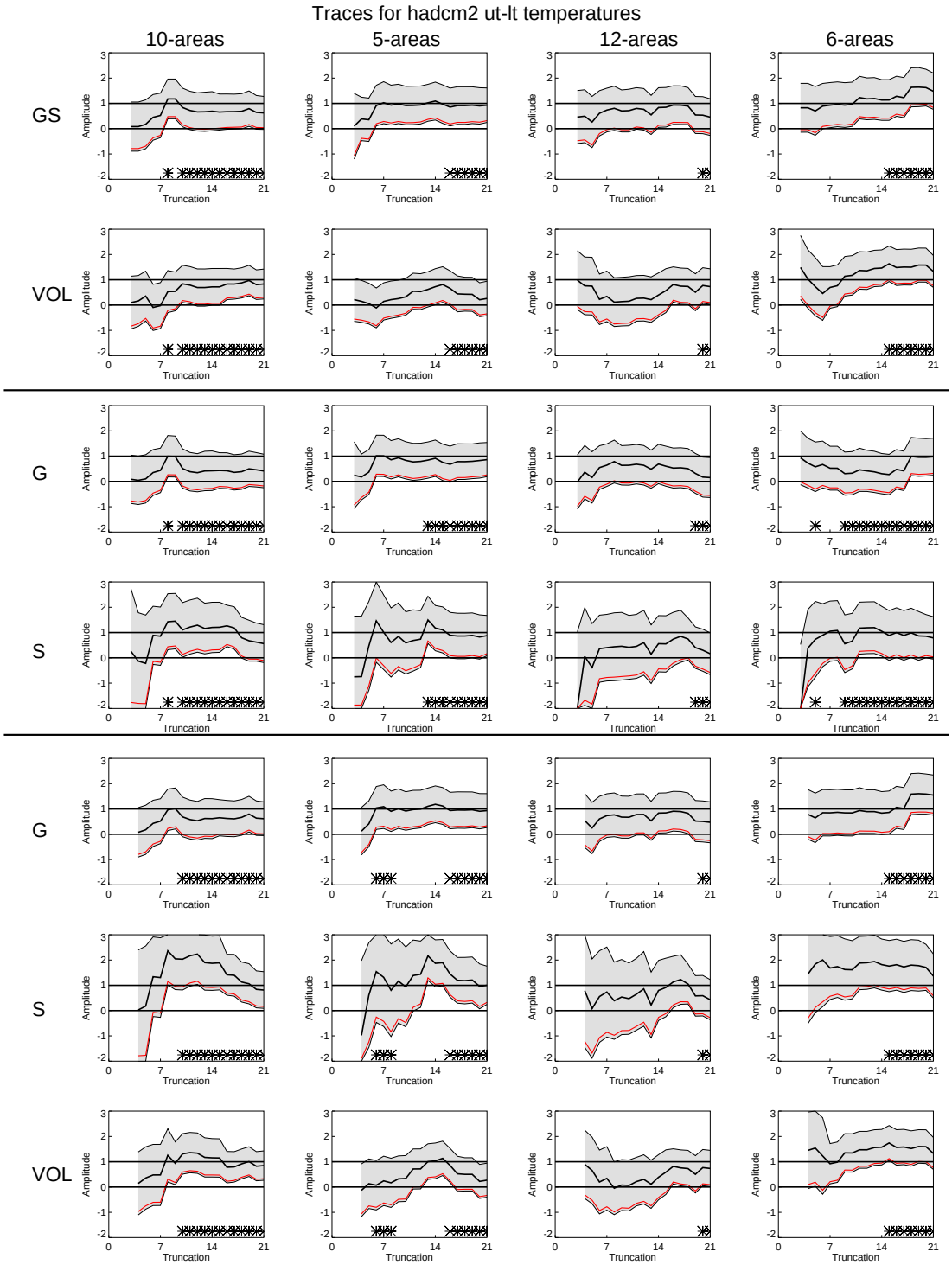

Figure 6.14 Changing beta estimates with increasing truncation for HadCM2 free troposphere lapse rates. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

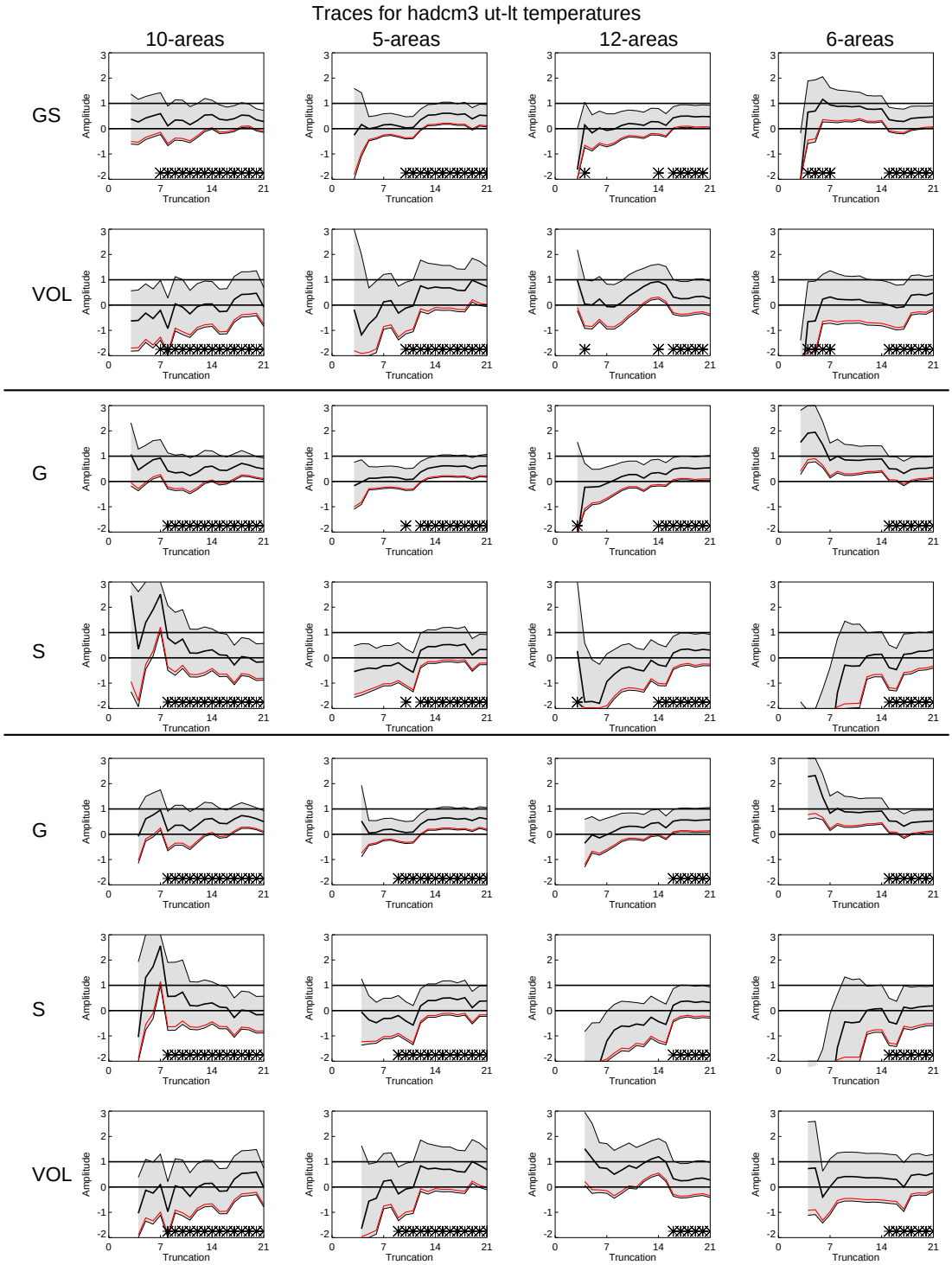

Figure 6.15 Changing beta estimates with increasing truncation for HadCM3 free troposphere lapse rates. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

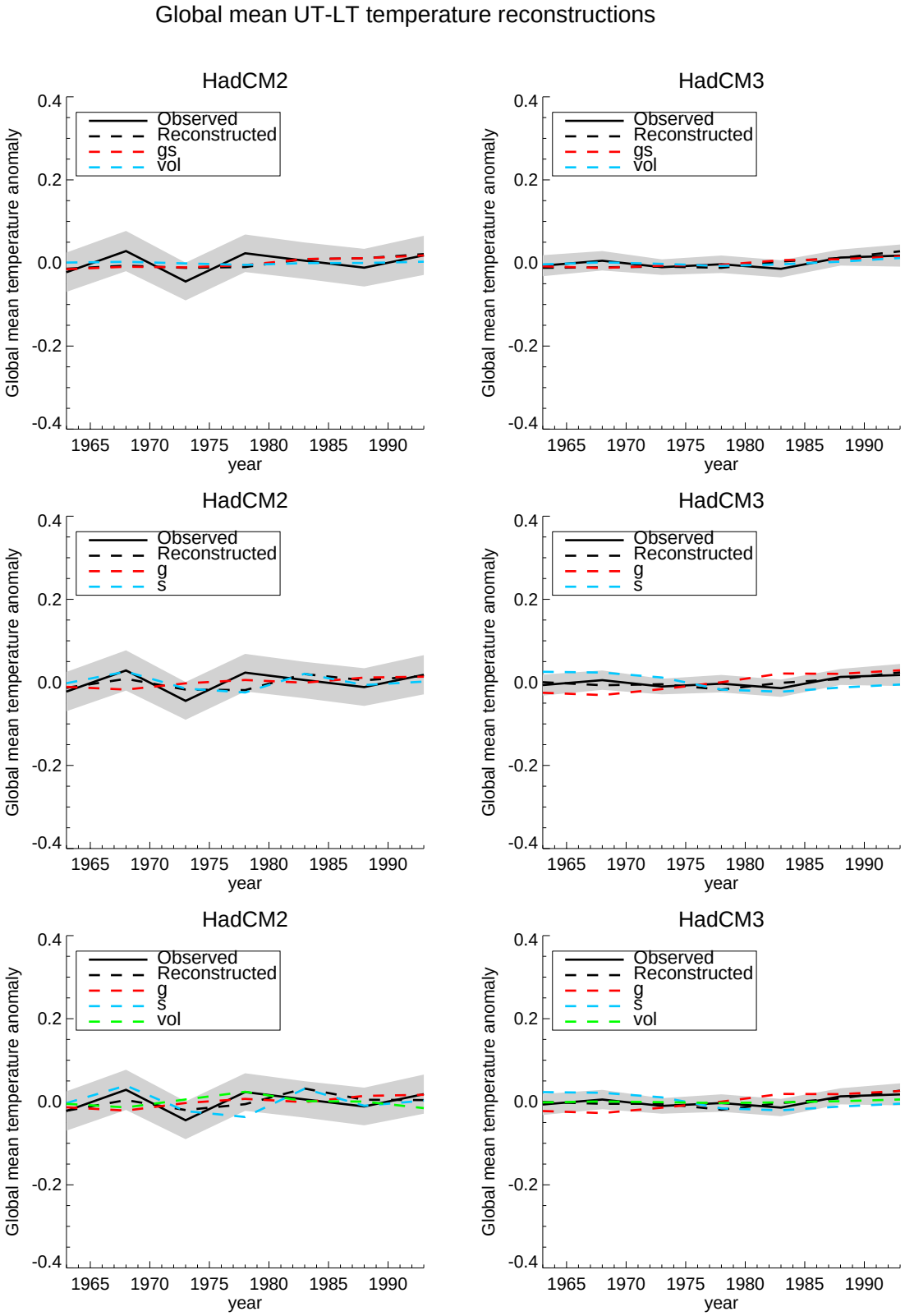

Figure 6.16 Reconstructed global mean temperature trends of free troposphere lapse rates for HadCM2 and HadCM3. The "observations" are projections onto leading modes of model simulated internal variability, and therefore differ between models. In each case the reconstruction is based upon the signals multiplied by their best-guess amplitude estimates.

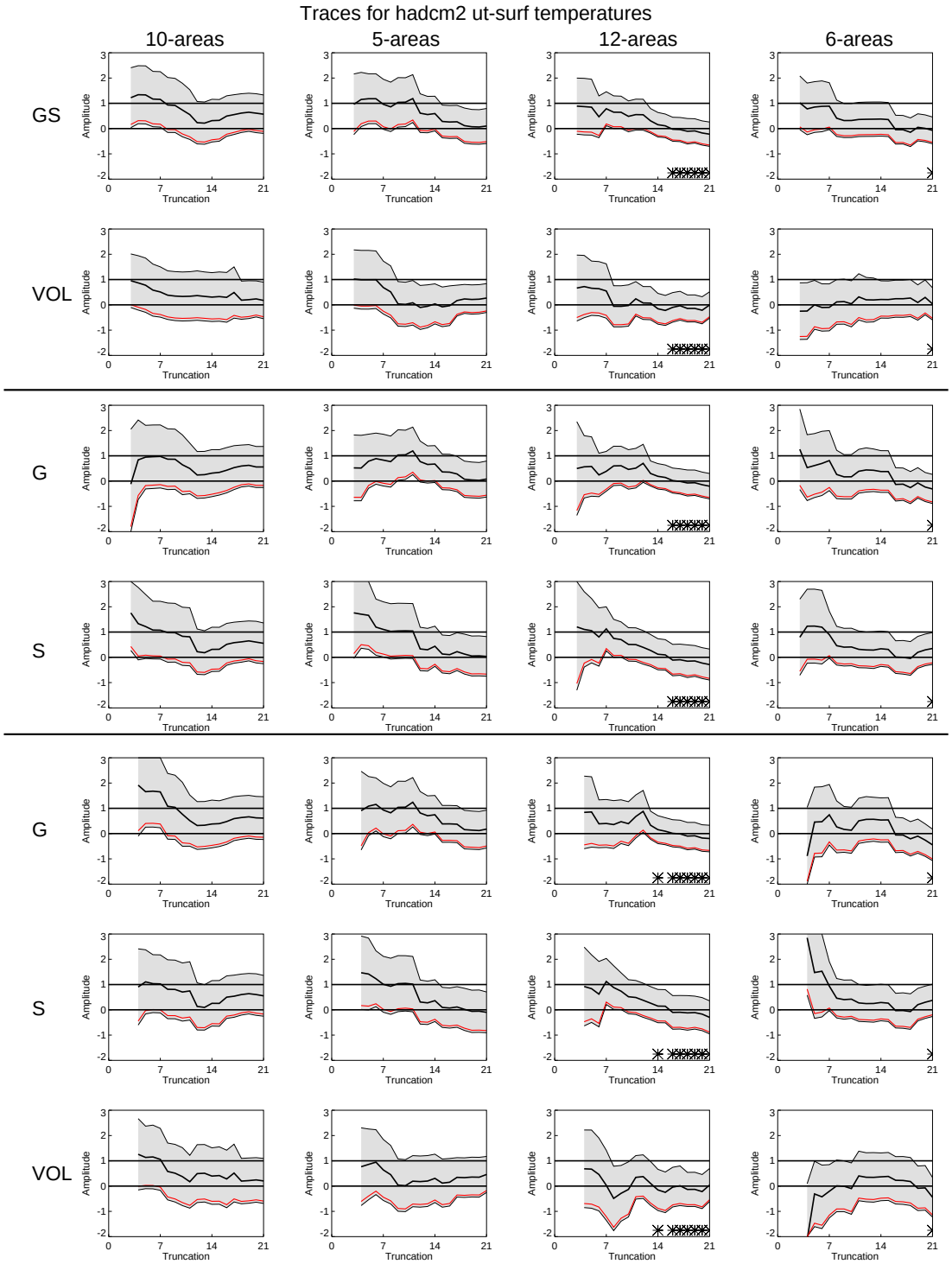

Figure 6.17 Changing beta estimates with increasing truncation for HadCM2 entire troposphere lapse rates. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

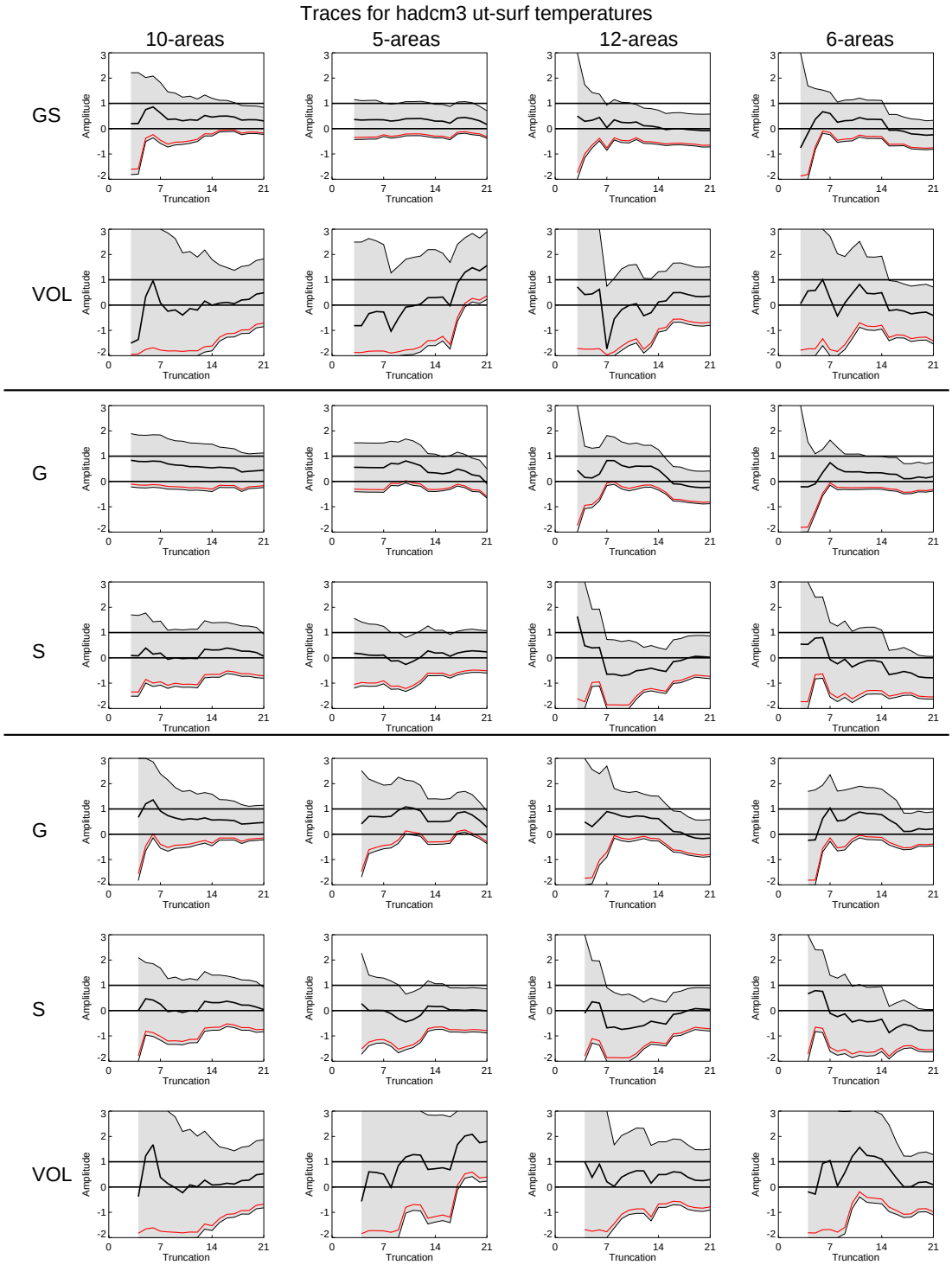

Figure 6.18 Changing beta estimates with increasing truncation for HadCM3 entire troposphere lapse rates. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

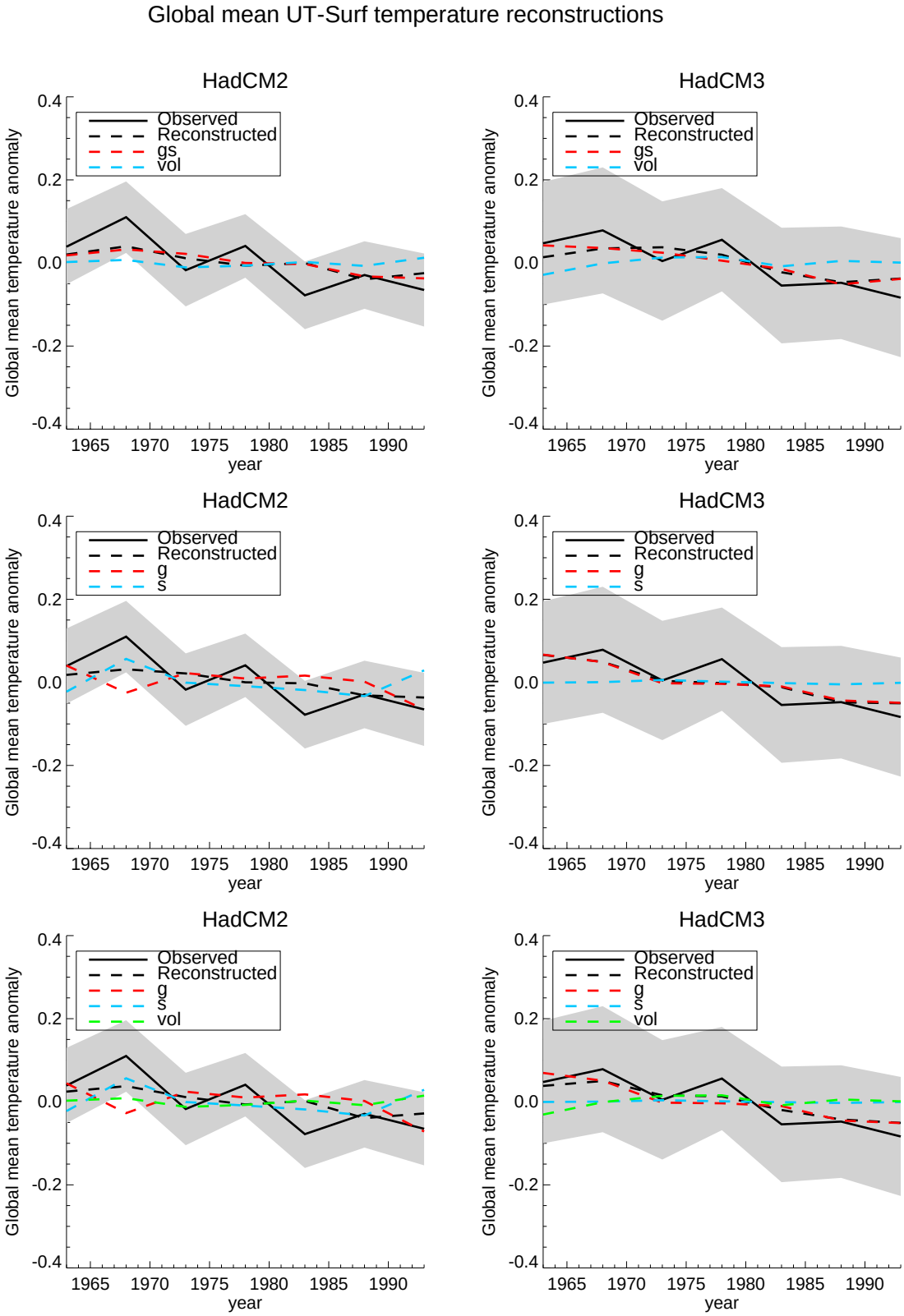

Figure 6.19 Reconstructed global mean temperature trends of entire troposphere lapse rates for HadCM2 and HadCM3. The "observations" are projections onto leading modes of model simulated internal variability, and therefore differ between models. In each case the reconstruction is based upon the signals multiplied by their best-guess amplitude estimates.

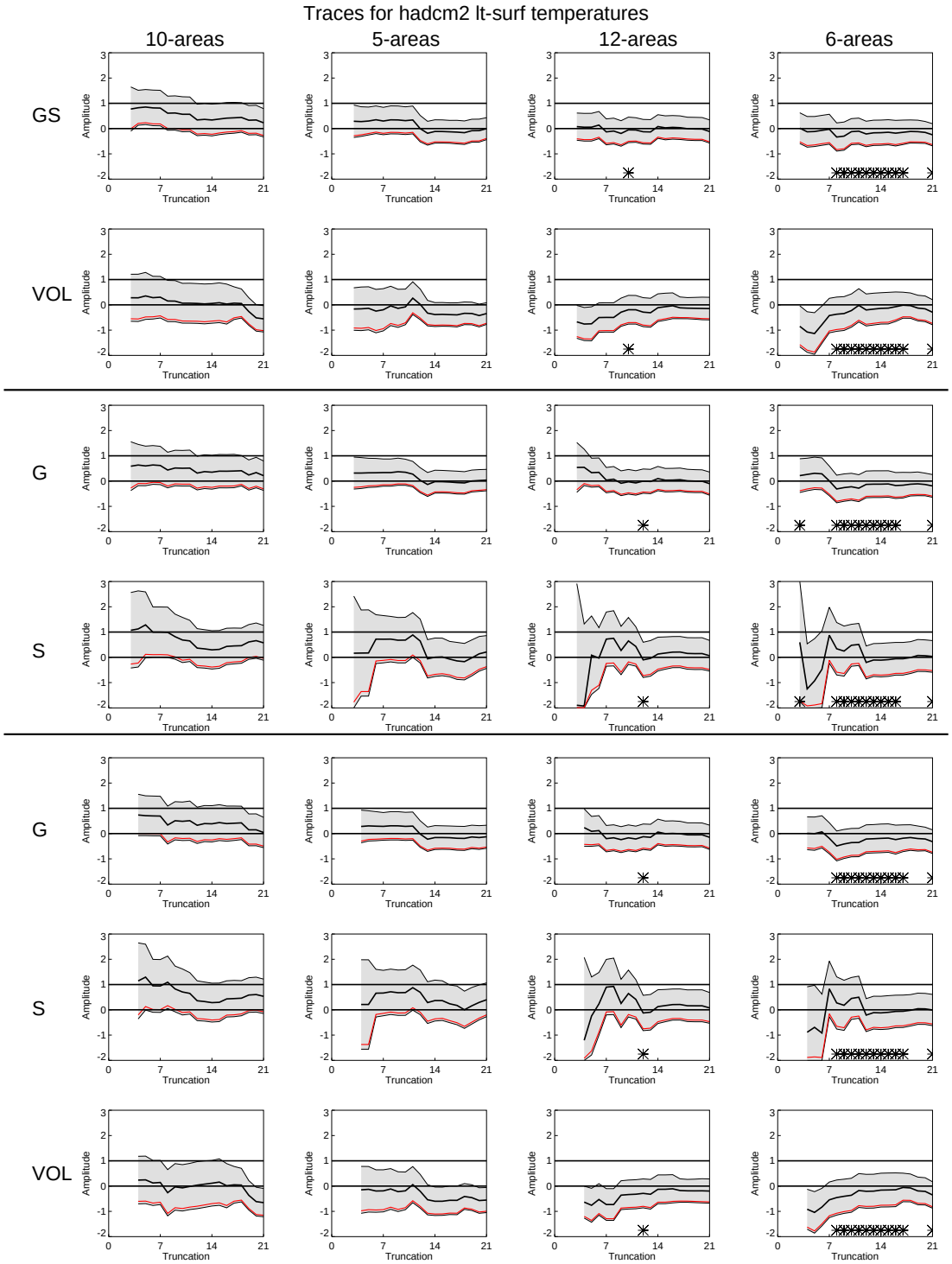

Figure 6.20 Changing beta estimates with increasing truncation for HadCM2 lower troposphere lapse rates. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

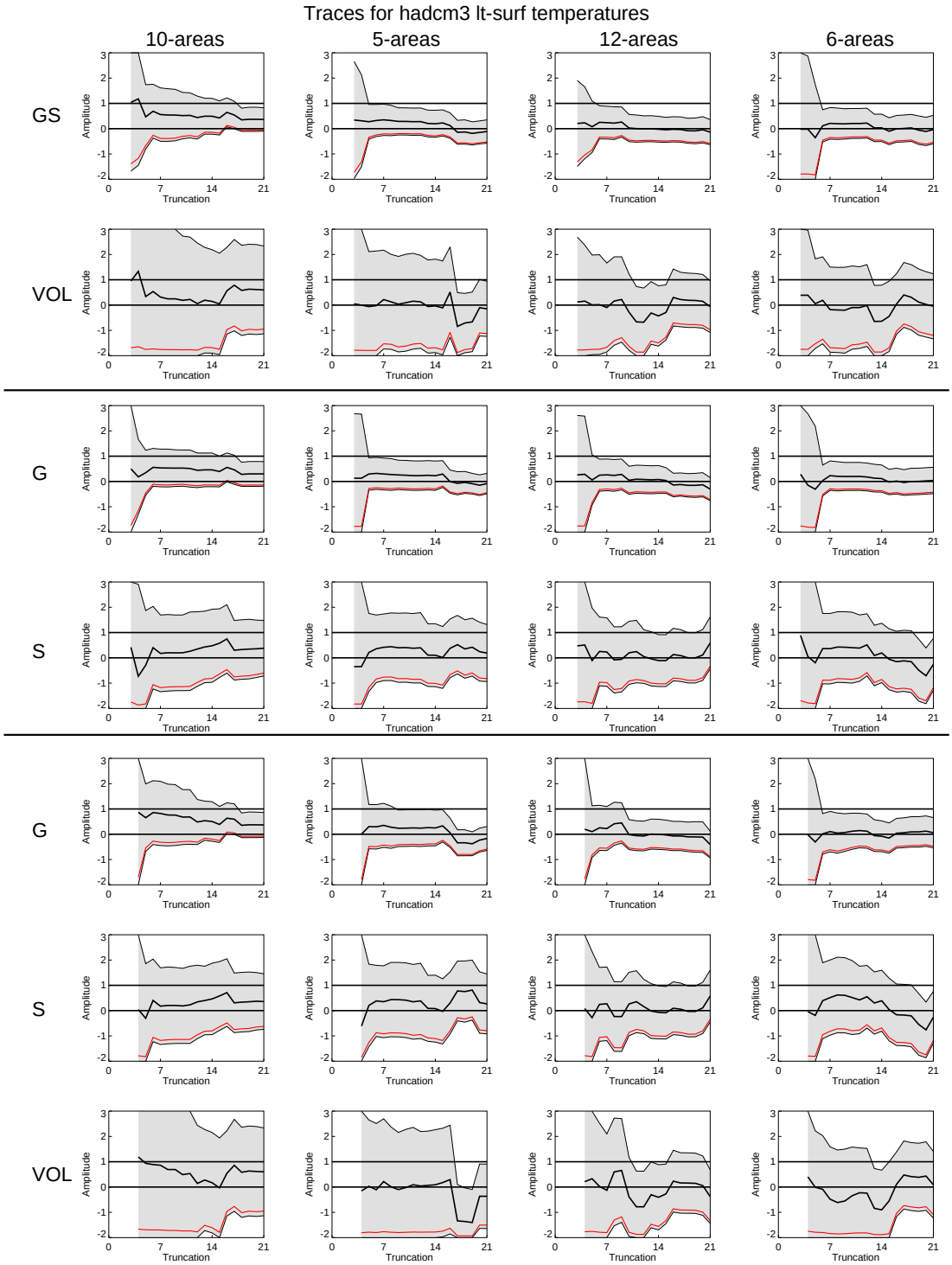

Figure 6.21 Changing beta estimates with increasing truncation for HadCM3 lower troposphere lapse rates. The best-guess model signal amplitude estimate in the observations is given by the bold line, with 90% uncertainty ranges denoted by grey shading. Detection confidence limits are denoted by a red line. Where the residuals are inconsistent this is marked by an asterisk. Results are considered for three input signal combinations, and four choices of LAA diagnostic.

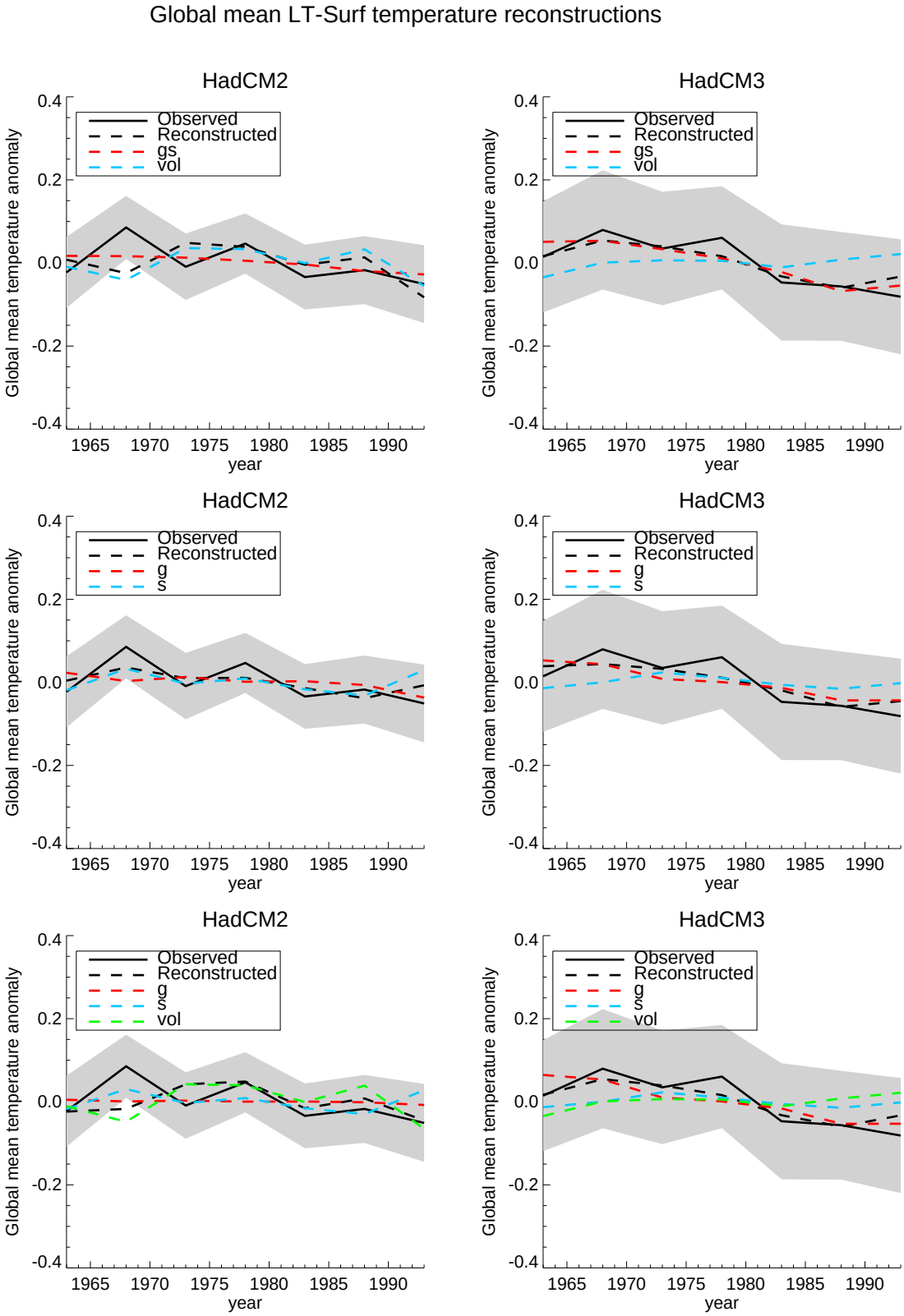

Figure 6.22 Reconstructed global mean temperature trends of lower troposphere lapse rates for HadCM2 and HadCM3. The "observations" are projections onto leading modes of model simulated internal variability, and therefore differ between models. In each case the reconstruction is based upon the signals multiplied by their best-guess amplitude estimates.

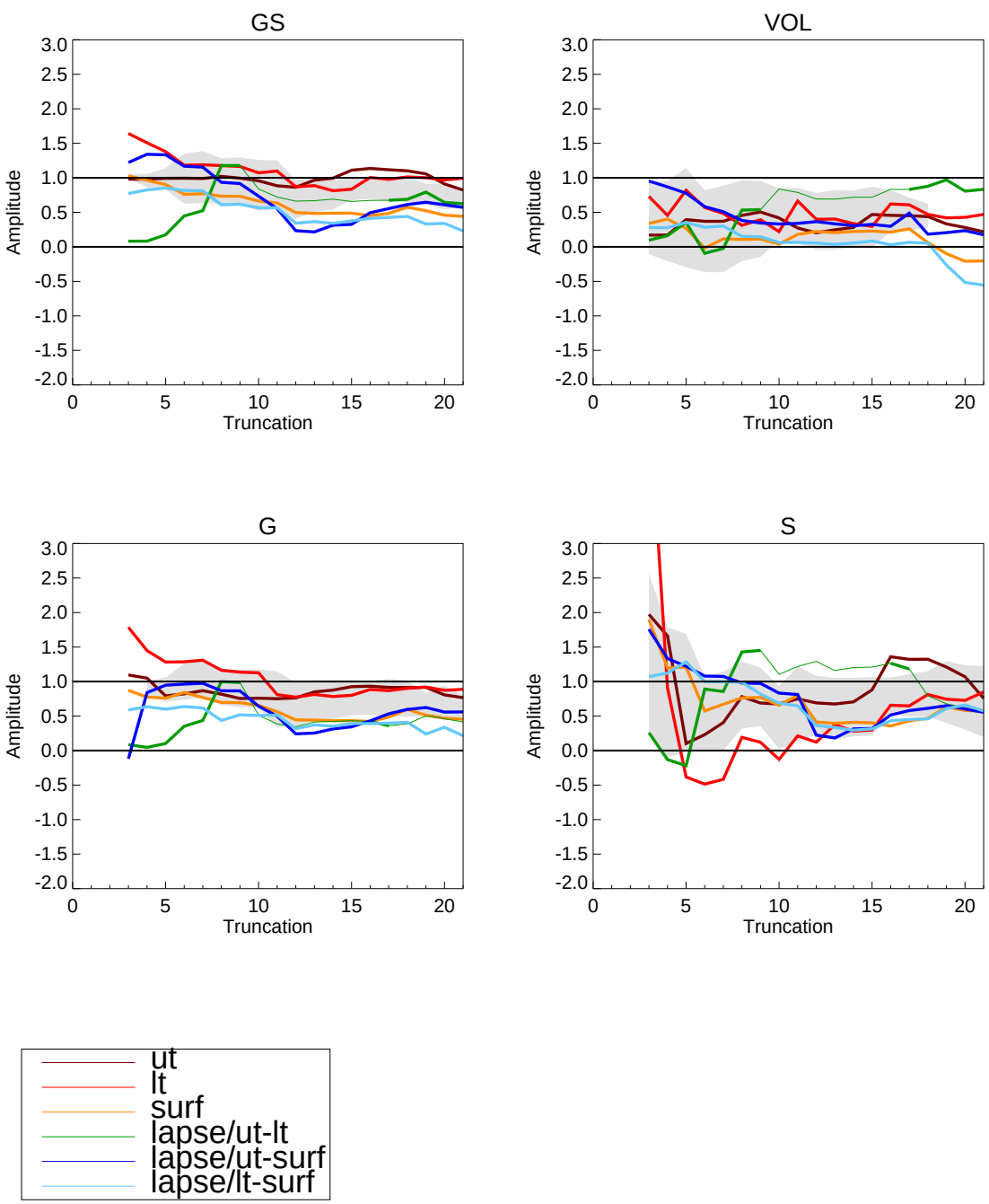

Figure 6.23 Changing beta estimates with increasing truncation for HadCM2 considering only the "smart" 10-area LAA diagnostic for all six input temperature variables. For each variable only the best-guess estimate is shown (see key for variables). The area of grey shading indicates regions where all six uncertainties overlap entirely. When the residuals are found to be inconsistent for any variable, the line is shown in feint.

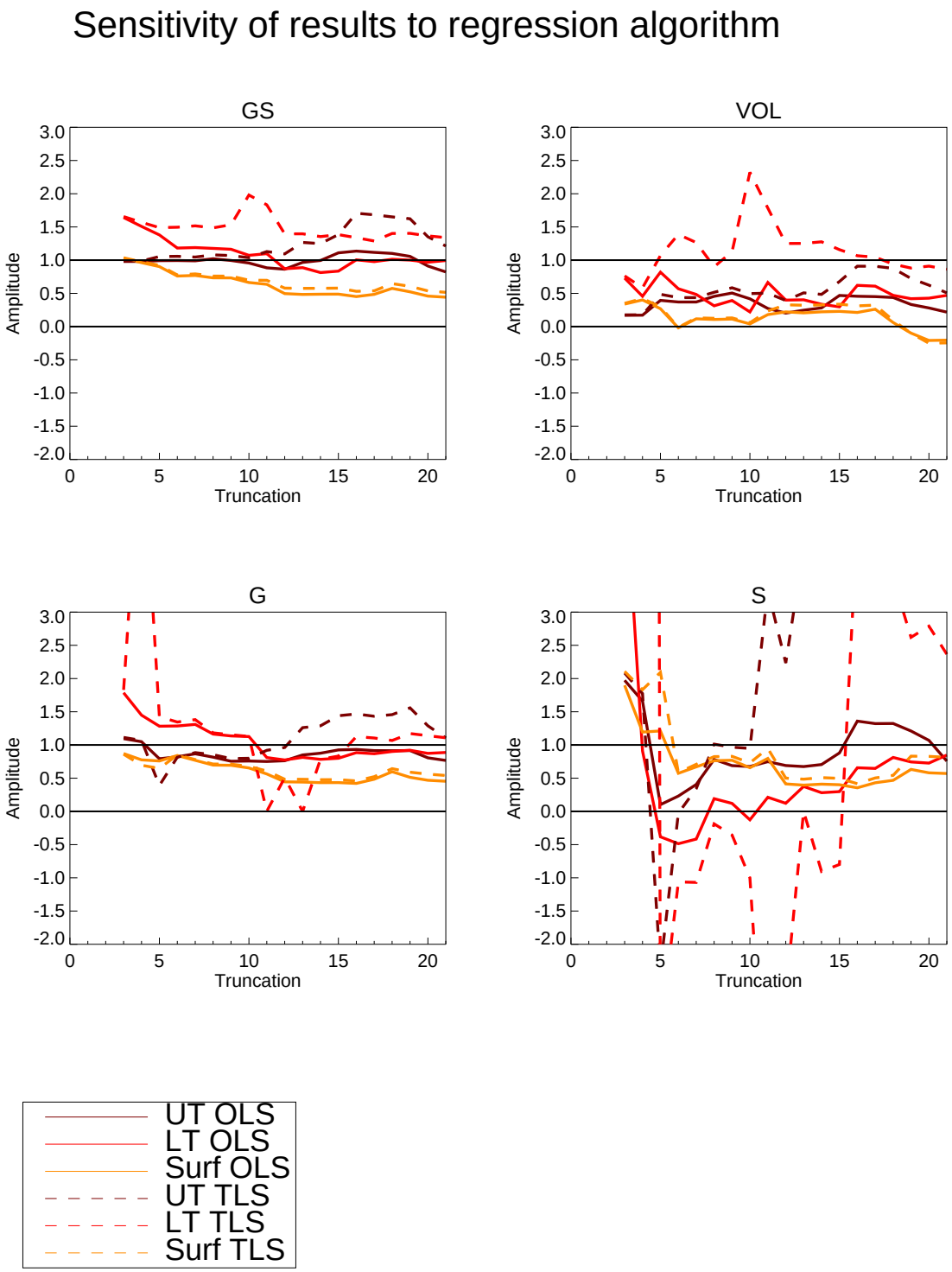

Figure 6.24 Plot assessing the likely sensitivity of principal results for HadCM2 for the three layer average temperature variables to choice of regression algorithm. OLS results are shown as solid lines and TLS as dashed.

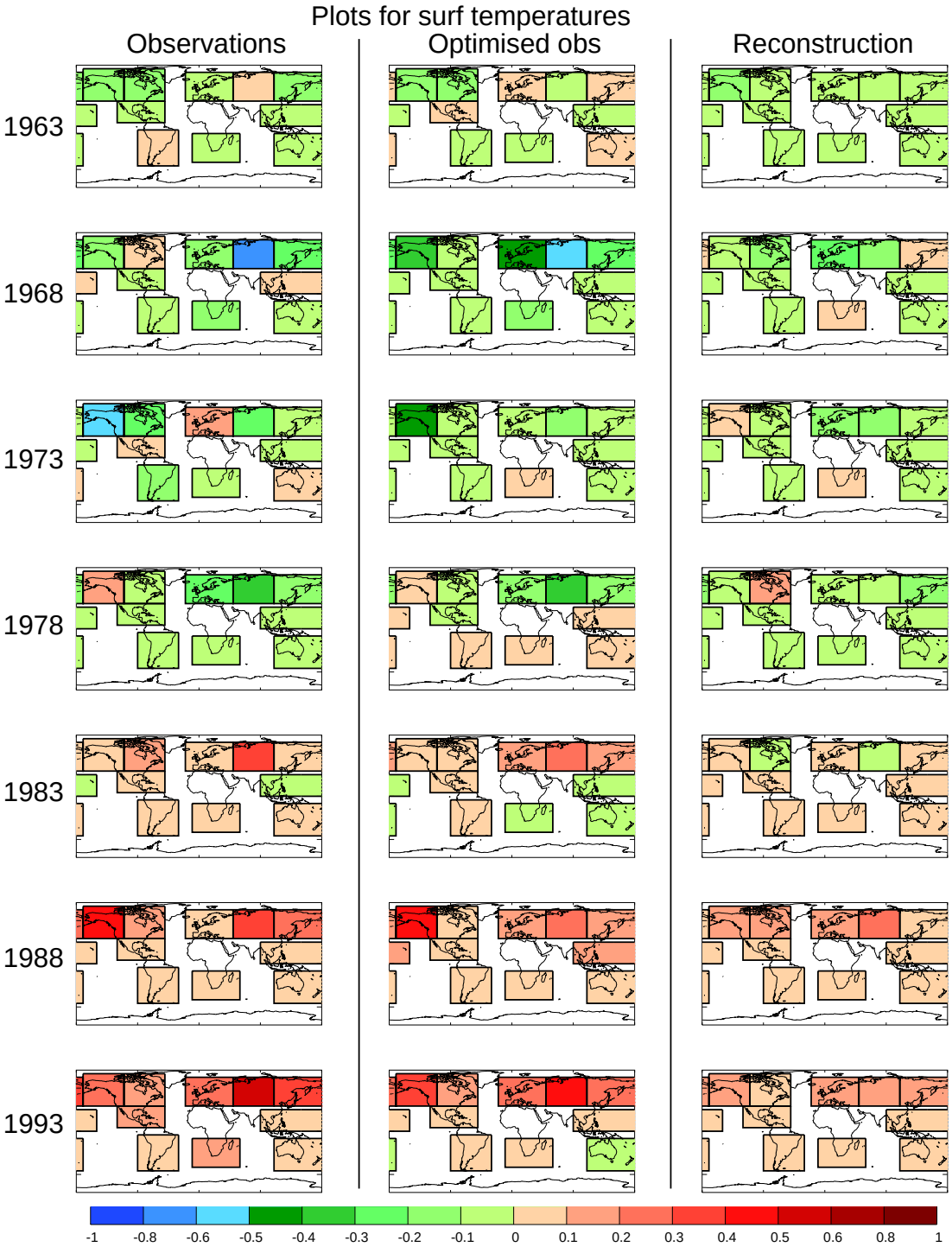

Figure 6.25 Raw observational input fields, optimised observational data, and best-guess reconstruction for near-surface temperatures. The reconstruction is based upon the G + S forcing combination for HadCM2.

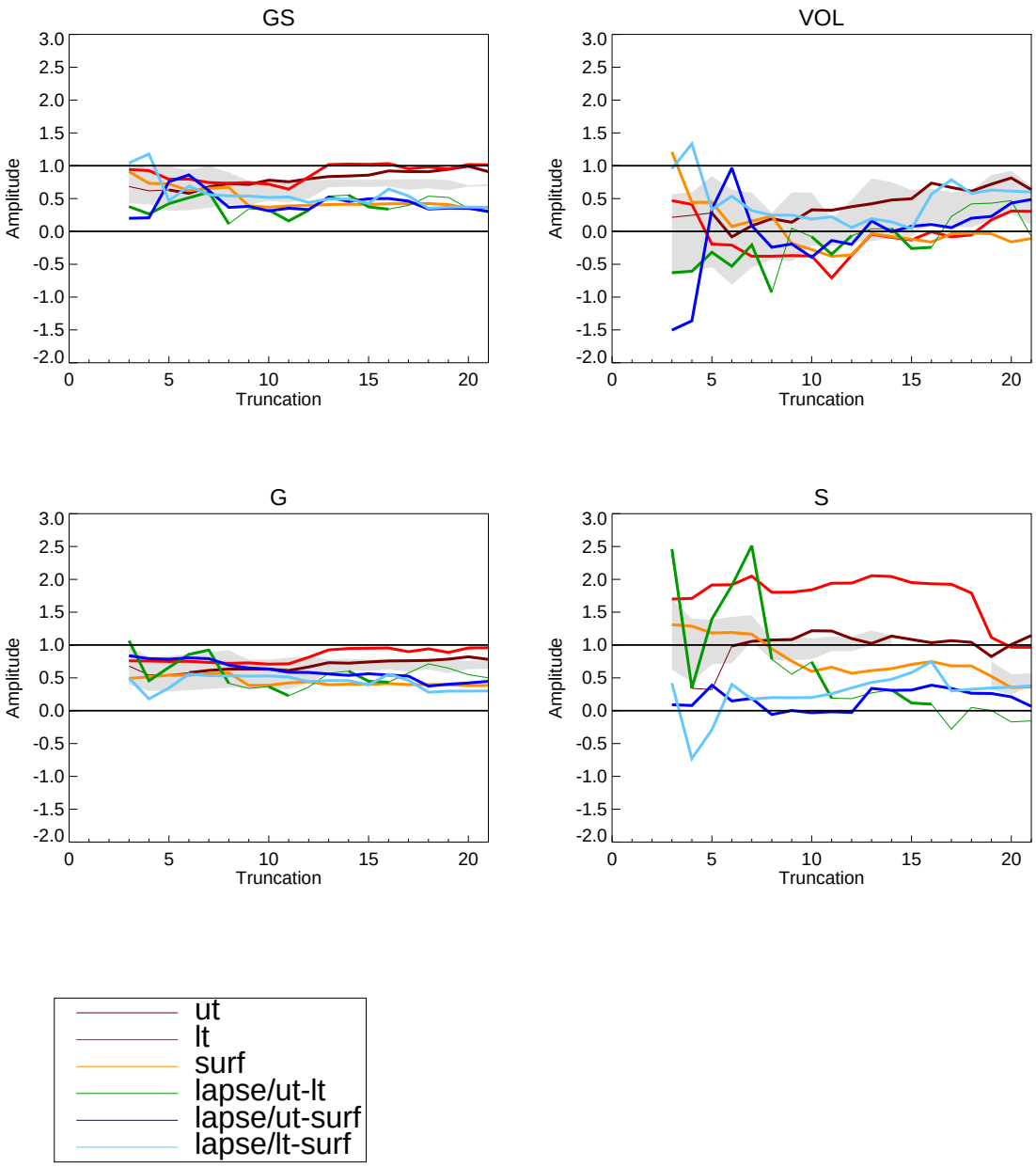

Figure 6.26 Changing beta estimates with increasing truncation for HadCM3 considering only the "smart" 10-area LAA diagnostic for all six input temperature variables. For each variable only the best-guess estimate is shown (see key for variables). The area of grey shading indicates regions where all six uncertainties overlap entirely. When the residuals are found to be inconsistent for any variable, the line is shown in feint.