2. The HadRT temperature record: Treatment and quality control

The HadRT temperature record is the only currently available long-term globally gridded (although incomplete) upper air radiosonde temperature dataset (see Chapter 1). The versions of HadRT used in this thesis are HadRT2.1 and 2.1s, which have been corrected globally for known post-1979 inhomogeneities (Gaffen, 1996) with reference to the MSUc series (Christy et al., 1998, 2000) in a similar manner to that described in Parker et al. (1997) for Oceania. The difference between these versions is that 2.1s is corrected only within the stratosphere, whereas 2.1 is corrected throughout the depth of the troposphere as well. Here (and in the rest of this thesis) reference is primarily made to version 2.1s as there is currently some uncertainty in the MSUc series data, especially in the lower troposphere (NRC, 2000 and references therein). Data are available for the 41-year period 1958-1998 on a monthly basis, as departures from the 1971-90 climatological monthly mean temperatures (Parker et al. 1997), on a 5° latitude by 10° longitude resolution grid.

Additional information made available from the Hadley Centre includes a record of the HadRT dataset coverage statistics. These are available on a monthly basis and detail the number of stations providing data for each grid box. Infilling of grid boxes has taken place in the HadRT dataset where three or more of the eight neighbouring boxes have values for any given month (Parker et al., 1997). The Hadley Centre also provided a limited list of contributing World Meteorological Organisation (WMO) station numbers for given grid boxes where requested. In addition Dian Seidel (née Gaffen) (NOAA Air Resources Laboratory) has allowed use of her radiosonde station metadata resource (Gaffen, 1996) which details all known changes in observational techniques and instrumentation within much of the WMO radiosonde station network. However, this is not to state that all changes have necessarily been recorded, or centrally documented, to date (a caveat stressed by Gaffen, 1996).

In this chapter a spatial quality control check is proposed, resulting in a proportion of the available data being discarded. Section 2.1 describes the reformatting of the monthly HadRT records to seasonal and annual values. In section 2.2 the methodology and rationale of the quality control procedure is described in detail, whilst in section 2.3 the results are described. Section 2.4 discusses methodological considerations. In section 2.5 the resulting dataset is briefly considered, whilst section 2.6 concludes.

2.1 Averaging to seasonal and annual values

The monthly HadRT record is first averaged to seasonal, and then to annual, values to enable comparison with the annual resolution HadCM2 and HadCM3 model fields. To allow sensitivity studies to be carried out in future detection studies, three datasets are constructed whereby the temporal coverage criterion for inclusion becomes progressively stricter. In all cases annual averages are calculated if three or more seasons have values in a given year. Version 1 (henceforth refered to as V1) requires the presence of values for only a single month in any season for a seasonal value to be calculated. Version 2 (V2), which is the preferred version, requires the presence of two months for inclusion. Version 3 (V3) requires full temporal coverage in any given season. The choice of V2 as the default version is a balance between the high spatial and temporal coverages desired in the dataset. V1 is likely to exhibit unacceptably high variance, at least on an individual grid-box scale in some cases, and V3 to give too little spatial coverage.

Figure 2.1 shows global and hemispheric annual average temperatures and coverage statistics using the three versions at the 500hPa level. The three versions of the temperature series agree at this height for these large scale averages, with slight departures (of the order of a couple of tenths of a degree Celsius) for individual years. Comparisons of this type are repeated throughout the depth of the troposphere and into the lower stratosphere. Above 50hPa there are larger differences between the versions of the order of half a degree to one degree Celsius early in the record, particularly for Southern Hemisphere averages. This is most likely due to a higher rate of balloon burst before reaching these heights at this time (Parker and Cox, 1995). The coverage statistics exhibit a relatively small reduction between V1 and V2, with a larger reduction between V2 and V3 at all levels, especially for stratospheric levels. This provides justification for the use of V2 as the default version in this thesis, as a consideration of spatial as well as temporal patterns of change is desired.

2.2 Quality control procedures

Previous climate change detection studies using the HadRT dataset have applied it solely in some zonally averaged sense (Santer et al., 1996a, Tett et al., 1996, Allen and Tett, 1999, Gillett et al., 2000, Hill et al., 2001, Tett et al., 2001, G. S. Jones et al., 2001). In these previous studies the three-dimensional (xyz) spatial field is zonally averaged to form a latitude-height (yz) two-dimensional field. In such a procedure much information is averaged which could, potentially, be useful in climate change detection studies. Here an attempt is made to quantify the suitability of using the HadRT temperature dataset to consider a fully three-dimensional spatial analysis. Following the discussion in chapter 1 describing the HadRT dataset, it is realistic to expect erroneous grid box values in at least a few cases. In detection and attribution studies one is less dependent upon the absolute values within a given dataset, but critically dependent upon the spatio-temporal pattern (Santer et al., 1993, Allen and Tett, 1999). Therefore it is important to consider how the field behaves as a whole, rather than individual grid-box or station series.

The radiosonde network has been used primarily for operational meteorological purposes. Only recently has attention been placed upon its potential use for climate change studies. The question of residual systematic errors in the radiosonde record remains and may explain part or all of the observed trends. Much previous work has been undertaken identifying and correcting for inhomogeneities in the radiosonde temperature records, either for individual stations (Gaffen et al. 2000a and references therein), or for individual grid-box series in the HadRT gridded product (Parker et al., 1997). In neither case is there a spatial consistency requirement imposed. Gaffen et al. (2000a) describe a number of potential sources of errors and consider a variety of techniques for identifying and removing spurious trends from individual station records. They conclude that fully automated techniques are potentially dangerous, given that climate variations may occur in a step like fashion and therefore be removed by break-point analysis. Such an analysis is also complicated as, depending upon the threshold criteria, the resulting ‘homogenised’ series could be significantly different from one another. Gaffen et al. (2000a) therefore suggest that a hybrid approach be taken, whereby available metadata are used to elucidate statistics which, if significant, are used to correct the individual station series. Such an idealised approach is not possible with the HadRT record as in many cases, grid-box values are the average of more than one station record. The contributing stations may not have contemporaneous changes in observational practice, and the station values are weighted according to distance from the gridbox centre (Parker et al., 1997), making any comparisons non-trivial.

No studies have previously explicitly considered the spatial homogeneity characteristics of the available radiosonde temperature records. In the current analysis, spatial quality control is undertaken firstly with reference to near-neighbour grid boxes, making the assumption that at least on a seasonal and annual basis there will be a high degree of similarity between neighbouring grid boxes. It can be seen from visual inspection of V2 of the raw HadRT2.1s data (see for example Figure 2.11) that, as expected, temperatures vary smoothly and coherently on a seasonal and annual basis over much of the globe. However, there are a number of grid boxes which, in a qualitative sense, appear to be highly anomalous when compared to their nearest neighbours. Any near-neighbour comparisons can only be made in those regions of the globe rich in data. It is likely that this problem is minor in comparison to the uncertainties associated with alternative potential spatial quality control procedures using the MSU record, GCMs, or operational reanalysis datasets such as NCEP (see chapter 1).

To avoid bias, those grid boxes which have been infilled in the HadRT2.1 and 2.1s datasets are masked out in the versions used in this thesis. In the original dataset an artificial correlation between neighbouring boxes occurs in such areas by construction (Parker et al., 1997). Use of such an infilled product is hence of questionable value, particularly in climate change detection and attribution studies, as it artificially alters the covariance statistics of the observations. Use of the infilled product may also lead to single grid-box errors propagating over larger areas, yielding regionally erroneous values. Figure 2.2 shows the effect of this masking procedure upon both the global and hemispheric averages and global coverage statistics at the 500hPa level (as for Figure 2.1, for which the green line is equivalent to the red line used here) using V2 of the HadRT2.1s dataset. The major impact is upon the global coverage, which is reduced by up to 50% at all levels by this procedure. There are also some minor effects of up to a tenth of a degree on global, and particularly hemispheric, averages for individual years at all pressure levels. Purely qualitatively there are no major trend differences between the two datasets however, in line with expectations. If infilling had led to qualitatively significant differences between series for such large-scale averages then confidence in the overall spatio-temporal consistency of the observations would be extremely low.

To perform near-neighbour checks a dataset of near-neighbour values, y, on both a seasonal and an annual basis is constructed. The criteria for the calculation of y at any grid box and at any time is that HadRT2.1s data exist for the grid box, and at least four of the eight surrounding grid boxes, at the pressure level under consideration. The value calculated is a simple average; no attempt is made to weight the values according to grid box area or the number of stations contributing to each of the surrounding grid boxes. It is believed that the effects of instigating a more complex method would be minimal, although this is not explicitly tested. Comparisons are then made on a grid-box basis between y and the HadRT 2.1s V2 observations, x, where y exists (a subset of x by construction).

The first test is a simple Z-Score value for each grid box over time in order to try to isolate individual erroneous values. The Z-Score used here is defined as follows:

$$Z - Score = \left| \frac{x - y}{\sigma_x} \right|$$

Where σ_x = Standard deviation of the observed grid-box field.

Individual values are flagged if they have an absolute Z-score of greater than three. Assuming that the data are normally distributed this corresponds to an approximately 99% confidence interval that the value is significantly different to that of the near neighbour average. As a sensitivity study, the σ term is also considered in terms of y and x-y. The Z-scores are also averaged over the entire record and any value greater than an absolute value of 1 is flagged as potentially erroneous, showing on average a very low degree of correspondence between the individual points in the series.

The other statistical test used in near-neighbour comparisons is the correlation coefficient between the observed series (x) and near-neighbour average series (y). Expectations are that the two series should be highly positively correlated (have a value of close to +1) on the timescales considered here. Therefore, any values lower than 0.25 are flagged for further consideration. The typical sample size is of the order of 30, although this varies widely (from 1 to 41). To ensure the sample size is not unrealistic and, therefore, that the standard error of the resulting correlation coefficient does not imply spuriously low values, a minimum sample size of five values is required for the calculation of the correlations. Neglecting standard errors, a correlation coefficient of 0.25 would yield an r^2 value of 0.06 (y explaining only 6% of the variation in x) and therefore justifies flagging of the grid box for further consideration. Spatial and temporal auto-correlation effects will tend to increase the calculated value of r, at least for good grid-box series, and therefore the approach is conservative.

The individual timeseries of x and y for those grid boxes identified as potentially dubious using these two statistical indicators are examined in detail for obvious potential erroneous values and break points. Reference is then made to the station metadata series (Gaffen, 1996), in an attempt to identify physically plausible reasons for these potential errors. If this exercise finds a physically plausible reason for the suspected error then a decision regarding editing of the series is made; in most cases this involves the discarding of whole levels or entire grid boxes from the dataset. If no metadata exist then the data will be retained unless it is seen to be obviously highly dubious, to minimise the chances of removing any true data.

Before deleting any data identified as being dubious in the near-neighbour comparisons, they are used to define critical values of both the Absolute Maximum First Difference Series (AMFDS) and standard deviation (S.D) at each pressure level. These are then applied to the entire HadRT 2.1/2.1s dataset so that potentially spurious data-points in more data-sparse regions can be identified. The mean AMFDS and S.D values for those grid boxes known to be in error from near-neighbour checks are used as the critical value for these simple statistics. The mean rather than minimum is used as, in data sparse regions, there is no y series with which to compare x, and therefore errors will have to be greater in magnitude to be identified with any degree of confidence. Grid boxes are flagged if two or more levels exhibit values of either AMFDS or S.D greater than these critical values. This reduces the dataset size and ensures that consideration is given only to those locations most likely to be in error, as errors are likely to occur at a number of levels for any dubious grid-box. These locations are subsequently considered with reference to the available station metadata in the same manner as in near-neighbour comparisons.

2.3 Results from the Quality Control exercise

Table 2.1 details the locations of those grid boxes which on an annual and seasonal basis fail the near-neighbour comparisons at any pressure level (see also Figure 2.3 for a graphical representation of the annual errors). Single Z-scores greater than 3 are indicated by an X, and average Z-scores greater than 1, or correlations less than 0.25, are indicated by an X. These latter two measures are considered more important in the context of climate change studies as they purport to identify gross differences throughout the series. The majority of those locations found to be in error are not necessarily in error in all seasons and annually, or to the same degree. Further, a large number of the grid boxes have only one or two individual erroneous Z-scores, which is expected by chance in approximately 1% of all cases, assuming a normal distribution, although this may be higher due to spatial correlation in the vertical. Therefore, in subsequent analysis, consideration will be limited solely to those grid boxes indicated with an X on an annual basis. To ensure that these results are not solely an artefact of temporal sampling differences, temporal coverage characteristics are tested by F-test and T-test statistics for those grid box series flagged, against those series not flagged (Table 2.2). T-test results indicate that at least on an annual basis they have a consistent average number of temporal data points. On a seasonal basis this link is more tenuous as, at all times, the average temporal sampling in erroneous grid boxes is slightly less than that for remaining grid-box series. This pattern is consistent regardless of the version of data used in the statistical analysis. The significant F-test result on an annual basis implies that there are at least some flagged boxes that are atypical in terms of their temporal sampling, leading to a significant difference in the distributions between the two populations.

The errors in annual temporal resolution HadRT2.1s fields are illustrated graphically in Figure 2.3, summed over all levels. The values denoted by X in the annual column of Table 2.1 are represented by shades of red in this graphical representation. Flagged grid boxes tend to cluster together. Major errors are seen in South East Asia, South Africa, southern North America, the Caucasus, and the Iberian Peninsula. The frequency of flagged grid boxes is also seen to vary seasonally, being most prevalent in autumn (SON) and least so in spring (MAM) and winter (DJF) (see Table 2.1). Clustering of flagged grid boxes suggests that either regional observational errors are large or that, in certain cases, there may be a single erroneous grid-box leading to a number of flagged series.

A complete listing of errors found, likely causes and actions taken for the entire HadRT2.1 series at an annual resolution in near-neighbour checks is given in Table 2.3. It has previously been noted elsewhere (Parker et al. 1997 and others), that Indian radiosonde stations are of dubious quality. It is therefore of interest to start this investigation by examining this region. From Figure 2.3 it can be seen that there are a number of errors being detected in South East Asian grid boxes. The Indian grid boxes themselves do not, however, fail the simple tests employed here, both of which take into account the intra-gridbox variance. The Indian series are seen to have contemporaneous break points, as cautioned by Gaffen et al. (2000a). To extend the understanding of processes leading to the grid boxes in this region being flagged, consideration is given to the individual grid-box plots shown in Figures 2.5 and 2.6. These show one of the flagged South-East Asian series and a neighbouring Indian series respectively. It can be seen in Figure 2.5 that in the case of the flagged box to the east of India (17.5°N,85°E), the observations appear to behave consistently, and it is actually the near-neighbour composite which is potentially in error. Examination of Figure 2.6, which is typical of all four of the Indian grid boxes surrounding that of Figure 2.5, confirms this. Two distinct break-points can be seen in most of the Indian grid boxes: one in 1969 coincident with a change in sonde type from Indian Fan Type to Indian Automodulation; and the other in 1990 when there was a change in computation method (Gaffen, 1996). All Indian grid boxes are removed following this analysis. This Indian example illustrates the importance of examining in detail those areas exhibiting erroneous test results rather than using a fully automated quality control procedure. If such a procedure had been used in this particular region then good data would have been discarded and dubious data retained.

Analysing errors on individual pressure levels reveals that the errors in South Africa and southern North America arise at the 850hPa level alone. Considering the standard deviation field at the 850hPa level shown in Figure 2.4, there are two grid boxes, one in each of the regions identified, which have highly anomalous values. Analysis by individual temperature series of both these grid-box locations shows periods of a large positive temperature bias at solely the 850hPa level; the South African grid-box temperature series is shown as an example in Figure 2.7. This grid-box (27.5°S,25°E) only has stations above 1000m altitude, so it may not be well correlated with neighbouring grid box series as it is measuring boundary layer temperatures at 850hPa. This does not, however, explain the consistent positive temperature departure exhibited. Investigating further, the only plausible explanation is that one of the stations contributing to the grid-box series (Bloemfontein, WMO number 68442) changed its radiosonde type from Vaisala (generic) to Vaisala 13 and Vaisala 12 in 1964 and then to Vaisala RS21 in 1973, consistent time steps with the observed break-points in temperatures. Other stations in the grid-box also changed over to Vaisala sonde type RS21 in 1973, but there is no record of a similar change in 1964. Whether one station as part of an average can have such an effect is dubious. Given the synchronicity of other changes within the grid-box station series, it may have been a consistent change in radiosonde type across all stations, solely recorded at one station in the network. As stated in Table 2.3, this grid box has been edited so as to exclude only 850hPa level temperatures. There is no compelling evidence to discard other levels for this location, which behave similarly to neighbouring grid box values.

Errors are not limited to the very lowest levels in the atmosphere. Considering the likely mechanisms which could cause dubious values to arise (see Parker and Cox, 1995, and Gaffen, 1994, for a detailed discussion), expectations are that errors will increase with altitude. This will be complicated in the current analysis by the corrections applied post-1979, to the HadRT2.1s dataset with reference to the MSUc record in the stratosphere (Parker et al., 1997). This analysis has found very few large biases in the stratosphere post-1979, giving increased confidence in the methods used by Parker et al. (1997), at least within this poorly sampled portion of the atmosphere in which relatively few near-neighbour comparisons were possible. As an example of large upper air discrepancies, the New Caledonian grid box at (165°E,22.5°S) is considered (Figure 2.8). It can be seen that a noticeable warm bias, especially at altitude (200 to 100 hPa), exists pre-1976 in this record. Available metadata suggest that the 1976 breakpoint correlates with a change in calculation method from manual to automatic at the station in question (Noumea, WMO number 91592). Such a change in observing practice could easily lead to the systematic bias exhibited in this record. For this grid-box all data at all levels are discarded, as evidence for a breakpoint remains even as low as the 850hPa level. As the bias occurs at least in part during the normalisation period, the use of any of this data is deemed to be dubious as the error will adversely impact the entire grid-box series.

Only those grid box series not used in near-neighbour comparisons were considered in the analysis using AMFDS and S.D indicators. The use of these measures yields a large number of potentially dubious grid-box series. This is not surprising since all of the errors found in the previous analysis were in a very few specific areas, predominantly Northern Hemisphere mid-latitude land areas, and the free atmosphere temperatures may behave differently at other locations. The prior belief in this analysis is that the radiosonde data are good and, therefore, compelling evidence is required to show that the data are dubious before taking action. Visual inspection of the temperature series for those grid boxes flagged in the more data-sparse regions was undertaken. Given that there is no near-neighbour reference series in these plots, identifying spurious breakpoints is both difficult and subjective. In some cases a limited number of neighbouring grid boxes have values, in which case these were used to make more informed decisions as to whether any given grid box was to be considered further. This visual analysis and comparison reduced the number of grid-box series to be considered.

A number of high latitude Northern Hemisphere grid boxes were flagged by the analysis, and do, indeed, show large excursions. These tend to be contemporaneous over large regions and show opposite sign tendencies in the troposphere and stratosphere. The possible break-points are seen to be short term and do not appear to affect the long-term trends in the dataset. Further consideration of seasonal timeseries for these plots suggests that the signal is predominantly from winter and spring. This could therefore be a manifestation of the AO / NAO system in the vertical temperature profile (Wallace and Thompson, 2000). Further work would be required to confirm this. In the current study the available data are sufficient to reject the identification of definite large-scale artificial jump-points in temperatures at these high northern latitudes. These break-points also fail to correlate with known metadata, giving increased confidence that they are a true physical manifestation of the climate system.

For a number of the other grid-box series identified there are either no, or very little useful, metadata. Rather than reject these grid boxes solely on a subjective basis they are retained in the current dataset. This leaves a rather small dataset for further consideration with reference to the metadata. In the final analysis these were cross-referenced and only 3 grid box series were found to be definitely dubious. The grid boxes, error types and possible reasons are summarised in Table 2.4. The errors are not discussed on a grid-box basis here, as they are similar in nature and magnitude to many of those identified in the near-neighbour comparisons.

2.4 Sensitivity of results to methodological assumptions

The results detailed above justify the running of quality control criteria in an attempt to isolate and remove gross residual spatial errors in the gridded HadRT datasets. However, careful consideration should be made as to the suitability of any test and the criteria used in discarding or retaining data. It is important that any test be as objective as possible. The statistical approaches used here are just a few of many potential indicators of inhomogeneities in a given observational series. The aim of this exercise was to check the suitability of using the HadRT record versions considered for fully three-dimensional climate change detection studies and, if necessary, make modifications based upon that analysis.

There is an argument that larger zones should be used because national and regional systems could be and, indeed in many cases, have been subjected to synchronous changes in observational technique (Gaffen, 1996, Gaffen et al. 2000a). If such a region were large enough, and not surrounded by data from other regions, then none of the grid-box values in the region would be picked out by near-neighbour comparisons despite all exhibiting erroneous values. This is confirmed by the analysis using near-neighbour comparison techniques for the Indian sub-continent, where it is only neighbouring grid boxes that exhibit errors according to the statistical analysis. However, there are problems in performing such regional comparisons as the dataset is already gridded, with each grid box potentially containing a number of stations from different nations / regions and therefore with different errors. To undertake such an analysis would require individual station records rather than a gridded product such as HadRT.

Results using Z-Scores calculated using different sigma terms are essentially similar, although important differences exist. Those analyses using σ_y as the denominator yield a greater number of critical Z-Scores. Given that it is expected that σ_y will be consistently the smallest of the three σ terms in almost all the grid-box series, this is an encouraging result. However, the problem with using σ_y is that it is likely to be biased by the lower variance of the near-neighbour composite series in cases where there are no errors (confirmed by simple F-Tests, which, in the majority of cases yield significant differences between x and y), and therefore potentially lead to false positive test results. Examining σ_{x-y} as the denominator in the Z-Score diagnostic, results are in line with expectations that the Z-scores are generally normally distributed with only ~1% of individual values being greater than 3 standard deviations from the mean. However, using σ_{x-y} has disadvantages in that it will fail to detect series in error, although being useful in detecting single anomalous points. On average it can be seen that, for any given grid box, the σ terms are ranked in the following order: $\sigma_x>\sigma_y>\sigma_{x-y}$. However, in the cases of anomalous grid-box series it can be simply proven that the term σ_{x-y} will be inflated to such an extent that using an average Z-Score diagnostic will give a null result (effectively yielding $\sigma_{x-y}>\sigma_x>\sigma_y$ in the case of dubious grid boxes). Z-Scores using σ_x as the denominator show intermediate results between the two extremes in terms of diagnostics considering the series average, with the expectation that they will be slightly conservative, as σ_x will be inflated in any dubious grid boxes. Given such considerations, results using σ_x or σ_y should be implemented here, as in this procedure there is a greater interest in finding series exhibiting, on average, a low degree of agreement rather than individually erroneous points. From a purely theoretical viewpoint it would be safer to use σ_y, as it is independent of the series being tested and therefore should detect all potential errors. However, results using these terms are very similar and those points picked out using the σ_y term, above and beyond those picked out by σ_x, have very small deviations between the observed and near-neighbour series. This means that, when cross-referenced with available metadata, they fail to be proven dubious with high enough confidence to be modified in the final dataset.

Errors which were found in the near-neighbour comparisons tended to be due to break points and/or high series variance. Therefore, two simple statistical diagnostics were used to look for these features in more data sparse regions, with critical values derived from those locations found to be in error in the near-neighbour analysis.

The most subjective part of the entire procedure for all grid-box series occurs when those grid boxes flagged for further consideration (and, in the near-neighbour comparisons, near-neighbour grid boxes) are examined by eye. This examination considers the general pattern of the series and attempts to define specific outliers and/or breakpoints. In those regions with near-neighbour (y) series, this is aided by having a reference series with which to make comparisons. It is recognised that the isolation of outliers in series will depend upon the individual(s) undertaking the analysis and their particular approach and judgement. There is no obvious way to reduce this potential error, although it should be ameliorated by the fact that objective tests were used as pointers towards consideration of the individual series. This can only be a conservative error however, as it will only fail to identify actual errors. This is because suspected errors are then compared to station histories through the (most likely incomplete) metadata dataset of Gaffen (1996), and data rejected only if a physically plausible reason can be found, unless they are extremely obviously in error.

2.5 Analysis of the treated HadRT2.1s temperature dataset

To check whether the spatial quality control has been successful, the near-neighbour comparison procedure was re-run for seasonal and annual diagnostics for both HadRT 2.1 and 2.1s using V2 of the treated dataset. The results of this analysis (not shown here) illustrate that the major spurious points have been removed. A number of residual errors remain in South East Asia on a seasonal basis, but on an annual basis no significant points of error remain in this region. The use of HadRT 2.1 also acts as a sensitivity study; in both the original and subsequent analysis most of the same points were picked out as in major error regardless of which HadRT version was considered. This gives increased confidence that the analysis is finding real spatial errors rather than those due to dataset treatment.

Having completed a set of quality control criteria and removed those locations for which there is a high degree of confidence that they are in error, the impact this has both upon univariate diagnostics such as global mean temperature, and more complex diagnostics should be considered. Comparing the global and hemispheric mean temperature series for the corrected and uncorrected series shows that the majority of changes are of order 0.1°C (although larger (~ 0.5-1°C) in the stratosphere in the less well-sampled Southern Hemisphere); the 500hPa level plot is shown in Figure 2.9 as an example. The great majority of these differences in global and hemispheric means between the series occur early in the analysis period. There is a tendency for a warmer stratosphere and cooler troposphere early in the corrected series, increasing the observed trends. This is contrary to the findings of Gaffen et al. (2000a), who tend to see a reduction rather than an increase in the observed trends for individual stations. It should be noted that the processing of the data used here was designed to identify and remove gross spatial inhomogeneities, whereas the Gaffen et al. (2000a) method attempted to identify and remove all spurious trends, therefore these results need not necessarily be contradictory.

Given that previous climate change detection studies using upper atmosphere temperature data have focused on zonal mean temperatures (Santer et al., 1996, Tett et al., 1996, Allen and Tett, 1999, for example), it is of interest to see how modifications made to the HadRT dataset impact zonal mean temperatures. Figure

2.10 depicts the zonal mean decadally-averaged temperatures for both the unedited and edited HadRT2.1s V2 records. Relatively small differences occur between the two series, with the maximum difference recorded being of the order 0.5 to 1°C in tropical Northern latitudes, where the Indian stations have been edited out, and early in the record. It should be noted that the changes have been implemented in discrete bands, and that confidence in the zonal mean pattern is greatest in Northern Hemisphere mid-latitudes which are well sampled and where proportionately more near-neighbour grid-box checks could take place. The more complex patterns of zonal mean temperatures in the tropical regions are potentially an artefact of residual errors within the sparse sampling network. Whether such small changes as those following the editing made here are significant is hard to quantify, but there are certainly some effects whereby the patterns in the zonal analysis appear smoother and more coherent in the edited version. It would seem pertinent to re-run the zonal mean detection algorithms (Allen and Tett, 1999, Tett et al., 2001) on this edited dataset to ensure that the relatively small changes observed do not have any significant effect on the results (see chapter 5).

To consider more complex patterns of change, decadally averaged temperature on pressure levels are used as a way of reducing the noise due to natural variability inherent in the series. Decadal means are only calculated if a minimum of five years of data are present and there are no more than two years of consecutive missing data for any given decade. The results from this analysis show, at least qualitatively, that the temperature patterns on pressure levels are far more spatially consistent and coherent in the treated dataset than in the unedited dataset. As an example, the 850hPa level decadal chunks for the V2 of the dataset are shown in Figure 2.11. It can be seen that in the raw dataset the dubious gridboxes in South Africa, Mexico, and Turkey are highly anomalous. In the treated dataset these gridboxes have been removed (boxed areas in Figure 2.11 have been removed) and so the field as a whole varies more coherently.

Finally, the question of the adequacy and accuracy of using a single dataset must be addressed. Santer et al. (1999) conclude that the use of a single upper air temperature dataset in climate change studies is likely to yield ambiguous results. They advocate the use of multiple datasets, preferably from different data sources (radiosondes, satellite retrievals and re-analyses) as a way to increase the confidence in any conclusions reached. Both the MSU satellite retrieval system and available first generation of reanalysis datasets have residual uncertainties, which reduce confidence in their suitability for use in climate change studies (NRC, 2000). To try to reduce the uncertainty in any results of future climate studies that use only HadRT data, an approach using all three versions of HadRT2.1 and 2.1s is recommended. This will provide an estimate as to how data treatment affects the results.

2.6 Conclusions

In this chapter consideration has been given as to the suitability of the HadRT radiosonde upper air temperature dataset for use in fully three-dimensional spatial climate change detection and attribution studies. A number of versions of the dataset have been constructed, with spatial quality control checks performed on one of these datasets. The quality control procedure identified a number of clearly dubious gridbox series within the record. Reasons for these dubious series were identified with reference to available station metadata, and the individual series deleted. The resulting dataset is reduced in coverage, but can be seen to be more spatio-temporally consistent and coherent. However, no claim is made that this series is correct. There is a high probability that a number of remaining (although small-scale) inhomogeneities exist within the reduced series. The next generation of radiosonde datasets (the CARDS dataset (CARDS website)) should hopefully be able to identify and correct for these remaining smaller scale errors in the radiosonde temperature records (Eskridge et al., 1995, Gaffen et al., 2000a).

Following the quality control procedure, a discussion of the methods used concluded that they were adequate and it was noted that, given the nature of the gridded dataset, other, more complex techniques (for example those used by Gaffen et al. (2000a)), were not viable. Examination of the treated dataset determined that it is likely to be suitable for use in future climate change detection and attribution studies, although a number of important caveats apply. There is no guarantee that the reduced dataset is free from residual (mainly small amplitude) errors and, therefore, care should be taken in any interpretation of results. As stated by Santer et al. (1999), the use of a single record in detection and attribution studies in the free atmosphere is highly unsatisfactory and could yield ambiguous results. Unfortunately there is no obvious gridded dataset, other than HadRT, that is long enough in duration and well enough constrained to be used in this work. In an attempt to ameliorate this situation the use of both HadRT2.1 and 2.1s is advocated for future detection studies, using all three versions developed for each.

Locations flagged for further consideration by near neighbour checks

Grid box location (°N,°E)	DJF	MAM	JJA	SON	*Annual*
(67.5,75)			X	X	
(62.5,5)		X			
(62.5,45)		X			
(62.5,65)		X		X	
(57.5,15)					X
(57.5,45)	X				
(57.5,-125)			X		X
(52.5,45)					X
(52.5,55)		X			
(52.5, -115)				X	
(47.5,25)			X	X	X
(47.5,55)				X	
(42.5,5)	X		X	X	X
(42.5,25)				X	
(42.5,35)		X			X
(42.5,45)	X	X		X	X
(42.5,65)	X	X	X	X	
(42.5,135)				X	
(42.5,-115)			X		
(42.5,-65)			X	X	
(42.5,-5)			X		
(37.5,5)				X	
(37.5,15)				X	
(37.5,25)	X	X		X	
(37.5,35)			X		X
(37.5,45)	X	X	X	X	X
(37.5,55)		X	X	X	X
(37.5,-5)		X	X	X	X
(32.5,35)		X			
(32.5,-115)					X
(32.5,-105)					X
(27.5,85)	X				
(27.5,135)		X			
(27.5,-105)			X	X	
(27.5,-95)					X
(22.5,75)			X	X	X
(22.5,105)		X			
(22.5,115)			X		
(22.5,-105)				X	
(17.5,75)			X	X	
(17.5,85)		X			
(17.5,95)	X	X	X	X	X
(17.5,105)	X		X	X	X
(12.5,95)	X	X	X	X	X
(12.5,105)			X	X	X
(12.5,-65)	X	X			

(7.5,105)			X	X	
(7.5,115)			X	X	
(-22.5,25)				X	
(-22.5,165)	X	X	X	X	X
(-27.5,25)	X	X	X	X	X
(-27.5,125)	X			X	
(-27.5,145)		X			
(-27.5,155)				X	
(-27.5,-45)			X		
(-32.5,25)	X	X		X	X
(-32.5,145)	X	X			
(-37.5,-65)				X	

Table 2.1 Location of potentially dubious grid box values, by season and annually, from near neighbour comparisons using version 2 of HadRT2.1s. Those locations exhibiting solely individual Z-Score errors are denoted by X, all others are denoted by X.

Consistency of temporal sampling characteristics between flagged and unflagged grid boxes for near neighbour comparisons

	DJF	MAM	JJA	SON	Annual
Average number of values for unflagged grid boxes	31.4	31.7	32.0	32.5	31.1
Average number of values for flagged grid boxes	28.5	30.7	29.5	30.1	30.6
F-test significance					95%
T-test significance			90%	90%	

Table 2.2 Table defining the temporal sampling characteristics of flagged and unflagged grid boxes resulting from near-neighbour comparisons on both an annual and a seasonal basis. The tests used here are the student's t-test and standard F-test. Values of significance are quoted only when 90% or greater.

Errors in near neighbour comparisons on an annual basis

Gridbox location	Period in error	Level(s) in error	Error type	Potential reason(s)	Action taken
(52.5°N 35°E)	Entire period	300hPa and 500 hPa	High variance	Numerous sonde changes	300 and 500hPa levels deleted
(42.5°N 5°E)	1968 to 1972	All levels	Positive bias	Sonde change	Delete period for all levels
(42.5°N 35°E)	1970 to 1986	All levels	Positive bias and high variance	Sonde changes and changes to corrections	Entire gridbox deleted
(37.5°N 35°E)	1986 to 1998	850hPa	Systematic 2 degree Celsius cold bias	Cessation of radiation corrections	Entire 850hPa level deleted
(37.5°N 45°E)	1976 to 1987	all levels	Positive bias and high variance	Sonde changes and correction changes.	Entire gridbox deleted
(32.5°N 345°E)	1958 to 1970	all levels	Highly dubious oscillations in temperature	Unsure	Delete record until 1970 at all levels
(27.5°N 75°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(27.5°N 85°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(27.5°N 95°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(27.5°N 255°E)	1966 to 1977	850hPa and 700hPa	Large systematic positive bias	Changed use of ground equipment	Delete entire 850hPa and 700hPa levels
(22.5°N 75°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(22.5°N 85°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(17.5°N 75°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(17.5°N 85°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(12.5°N 75°E)	Entire record	all levels	Breakpoint in1990	Calculation change	Entire gridbox deleted

(12.5°N 85°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(12.5°N 95°E)	Entire record	all levels	Breakpoints in 1969 and 1990	Sonde change and calculation change	Entire gridbox deleted
(-22.5°N 165°E)	Before 1976	all levels	Systematic warm bias	Change to automatic calculations in 1976	Entire gridbox deleted
(-27.5°N 25°E)	1963 to 1974	850hPa level	Large (6 degree) warm bias	Sonde changes	Entire 850hPa record deleted

Table 2.3 Listing of grid boxes found to be anomalous by near neighbour comparisons, detailing potential reasons from available metadata. For more details on the status and potential deficiencies in the metadata see Gaffen (1996).

Errors found in sparse network analysis on an annual basis

Gridbox location	**Period in error**	**Level(s) in error**	**Error type**	**Potential reasons**	**Action taken**
(17.5°N 35°E)	Pre-1974	all levels	Cold bias	Station introduced to the series in 1974	Entire record deleted
(17.5°N 335°E)	Pre-1970	all levels	high variance	Change in sonde and tracking method	Entire record deleted pre-1970
(7.5°N 75°E)	Entire record	all levels	two break-points	Change in sonde type and calculation methods	Entire record deleted

Table 2.4 Listing of grid boxes found to be anomalous in data sparse regions, detailing potential reasons from available metadata. For more details on the status and potential deficiencies in the metadata see Gaffen (1996).

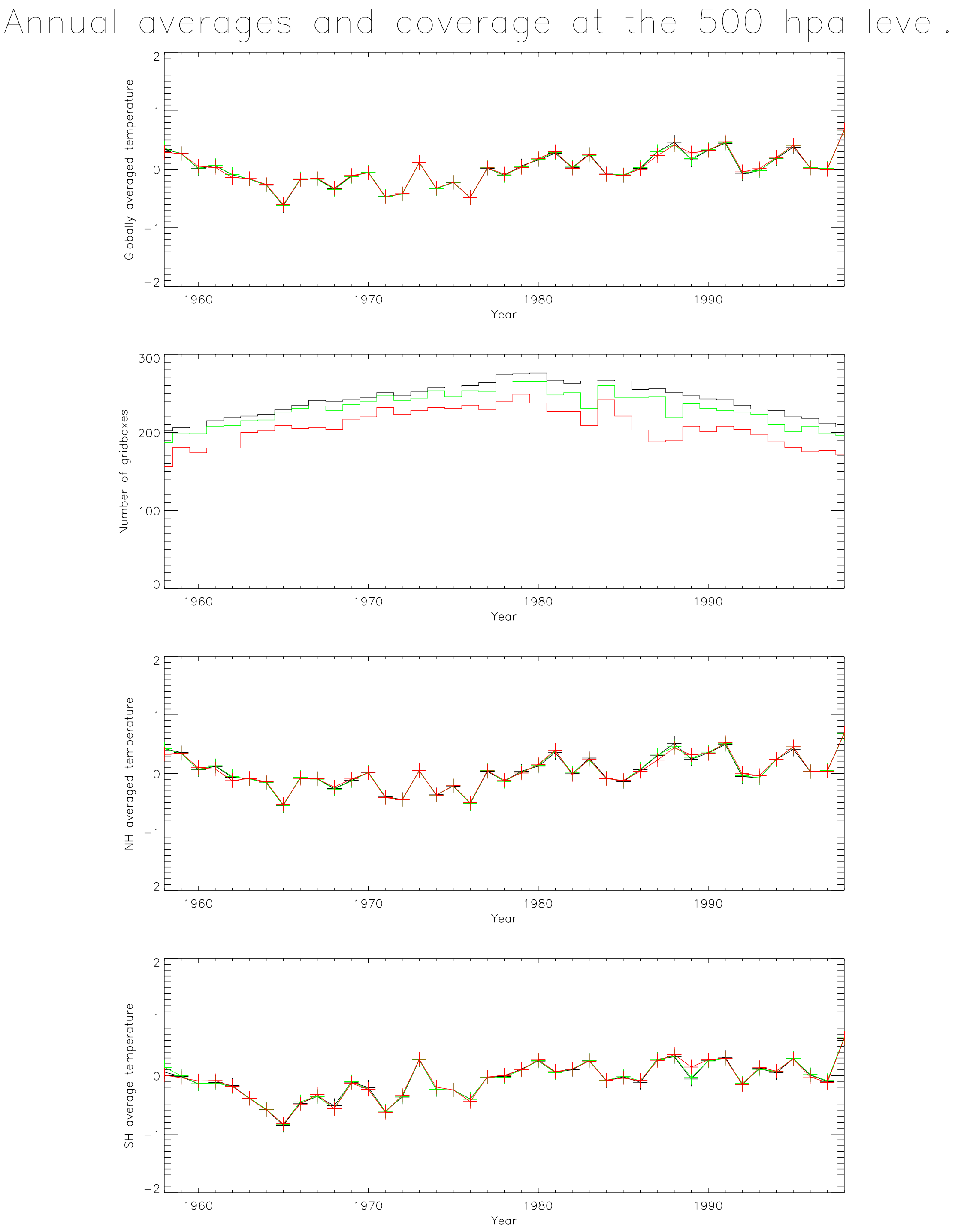

Figure 2.1 Global and hemispheric annual average temperatures for the HadRT dataset at the 500hPa level using the three inclusion criteria. V1 is denoted by the black lines, V2 (the default version) green lines, and V3 red lines. Where values are identical they are overlain. A global coverage statistic is also shown to denote the effects on the spatial completeness of each record. Grid boxes are only included in this analysis if they contain primary observational data.

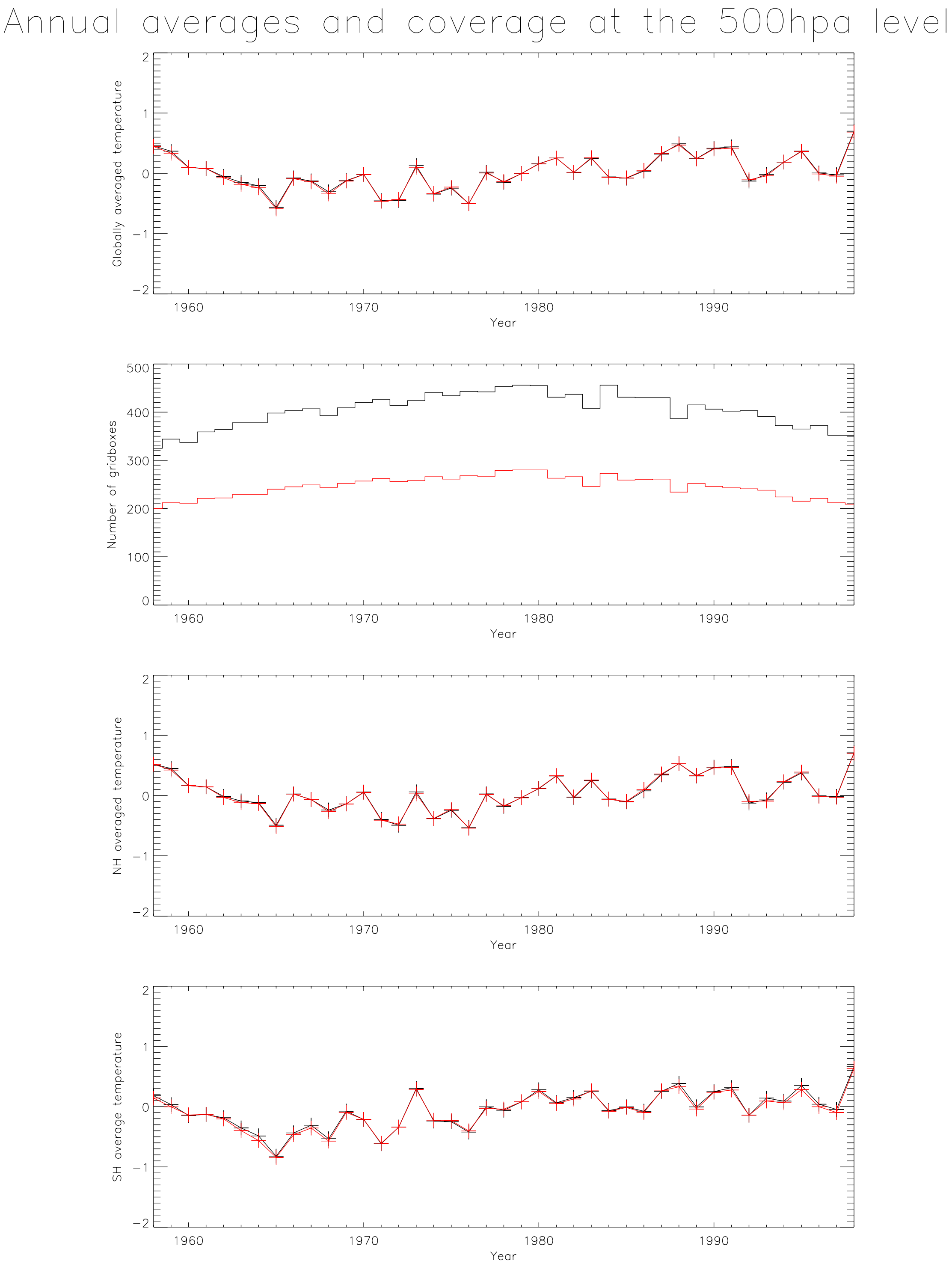

Figure 2.2 The effects of removing infilled grid box values from V2 of the HadRT2.1s data. Coverage is greatly reduced, but there is little difference in the global and hemispheric means. For comparison purposes the green line in Figure 2.1 is equivalent to (but not directly the same as) the red line in this Figure.

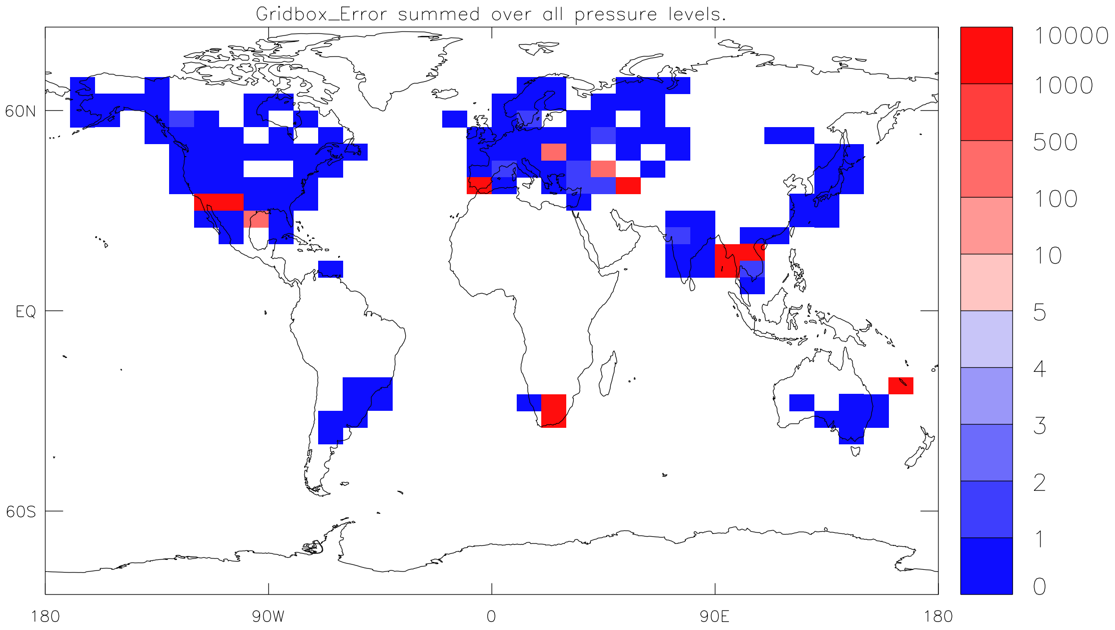

Figure 2.3 Grid box errors summed over all pressure levels, different error types have been given different values, based upon their perceived importance. Individual Z-Scores have a value of 1, average Z-Scores 100, and correlations 1000. These were chosen solely so different error types would be clear in this graphical representation.

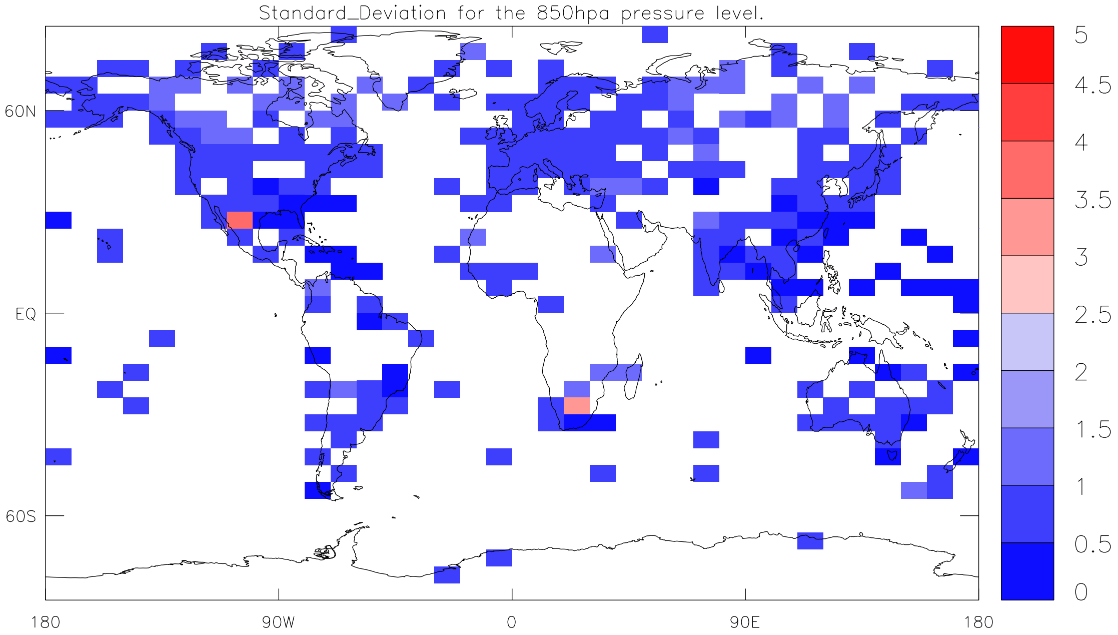

Figure 2.4 Global field of grid box standard deviations at the 850hPa level. The field is seen to behave in a relatively smooth and coherent manner, although two anomalous points stand out, one in South Western North America and one in Southern Africa.

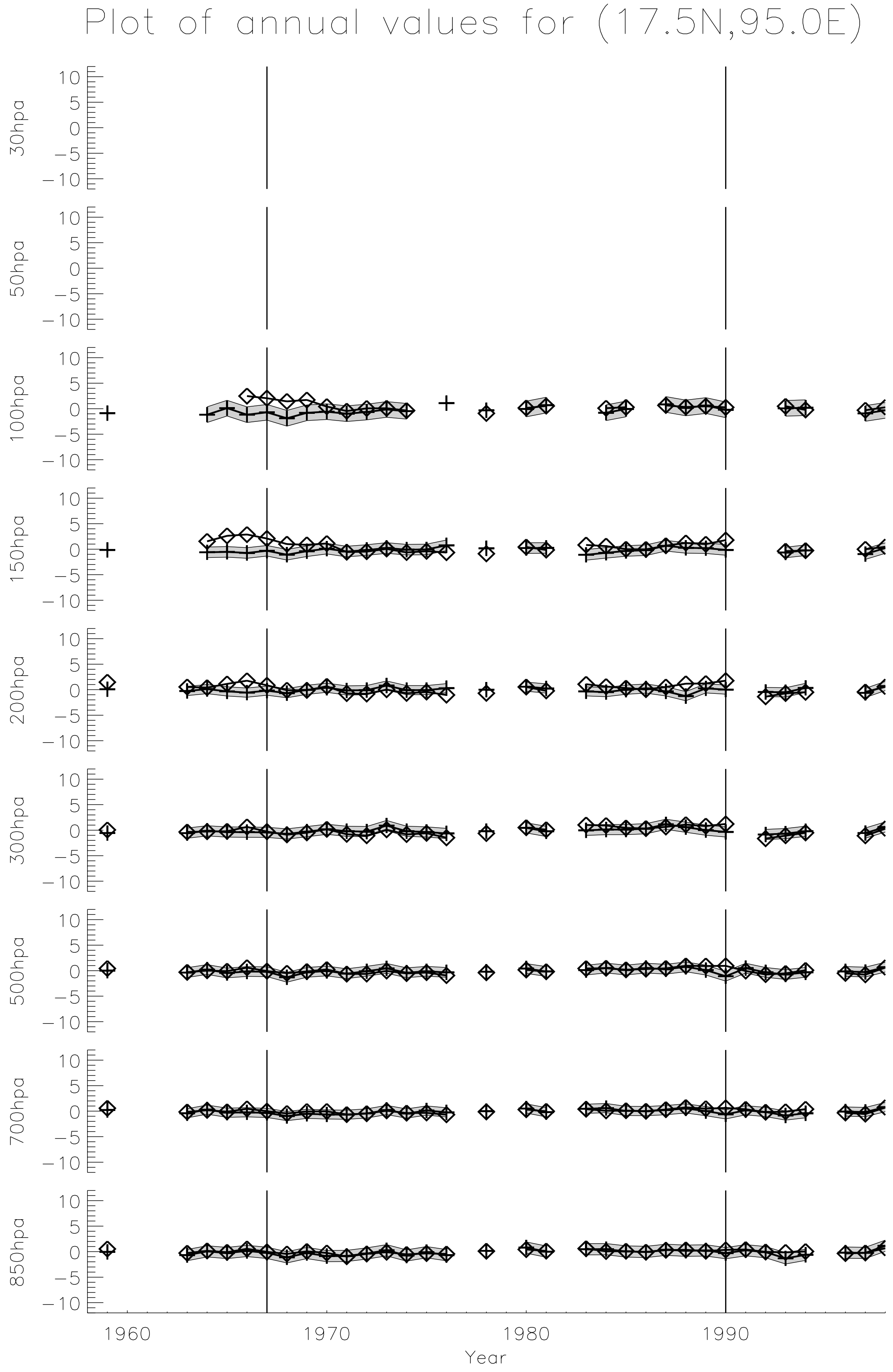

Figure 2.5 Grid box temperature series on levels for the South East Asian grid box. Observations are represented by crosses, with shading to ± 2σ ; near neighbour averages are represented by diamonds. Note that it is the y values that appear dubious, especially at height.

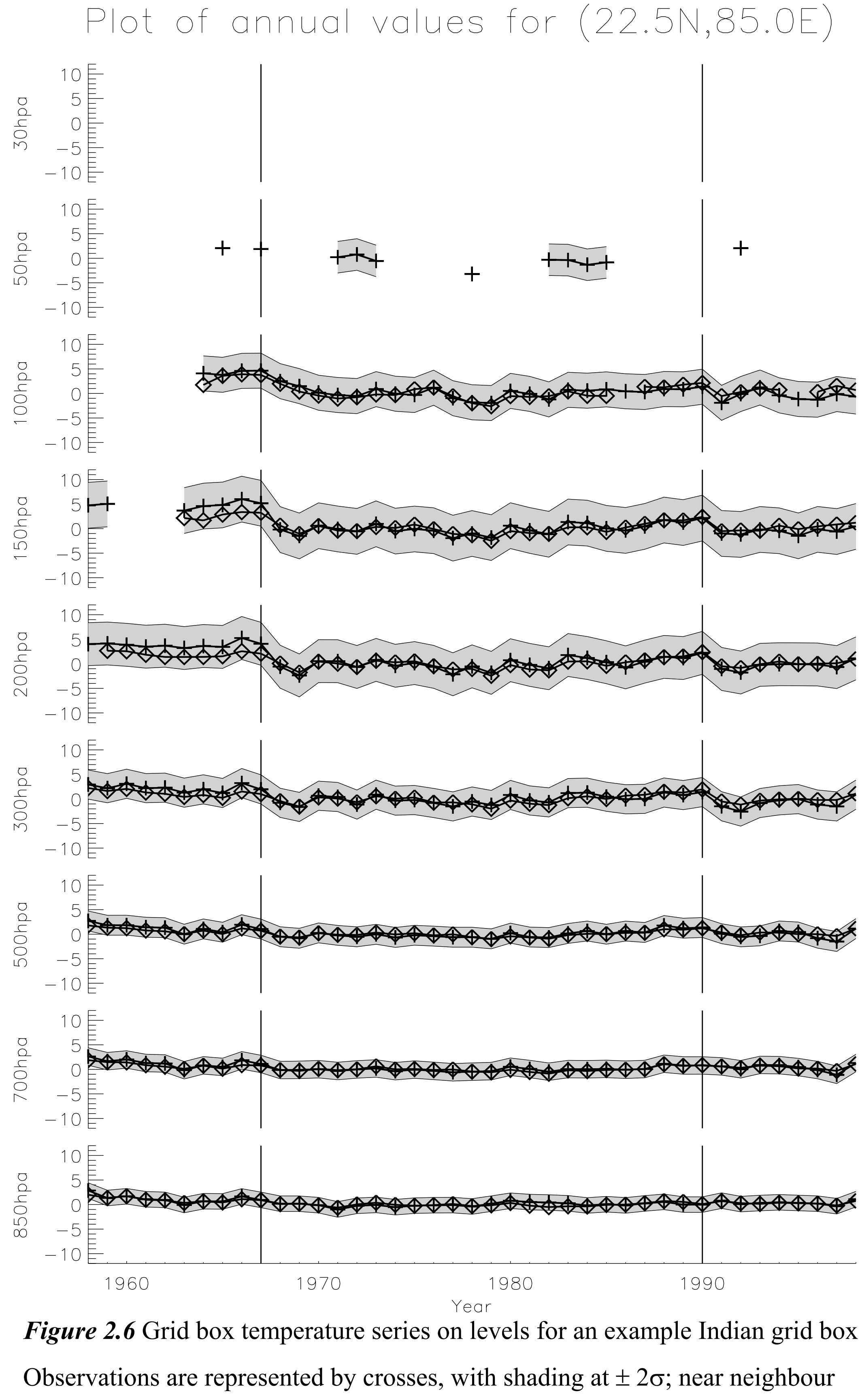

Figure 2.6 Grid box temperature series on levels for an example Indian grid box. Observations are represented by crosses, with shading at ± 2σ; near neighbour averages are represented by diamonds. Note significant breakpoints in 1969 and 1990.

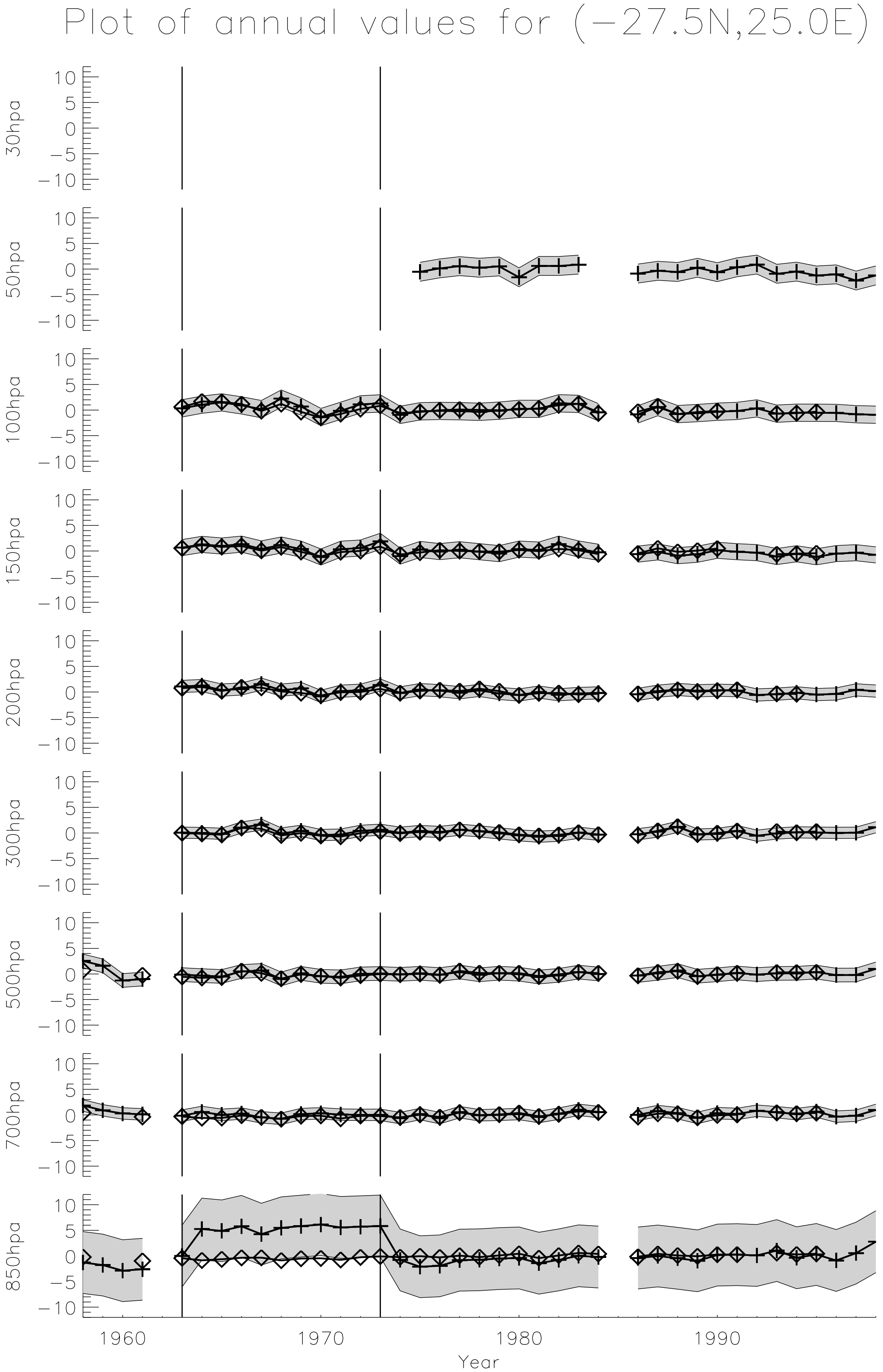

Figure 2.7 Grid box temperature series on levels for the South African grid box. Observations are represented by crosses, with shading to ± 2σ ; near neighbour averages are represented by diamonds. Note significant anomaly at the 850hPa level.

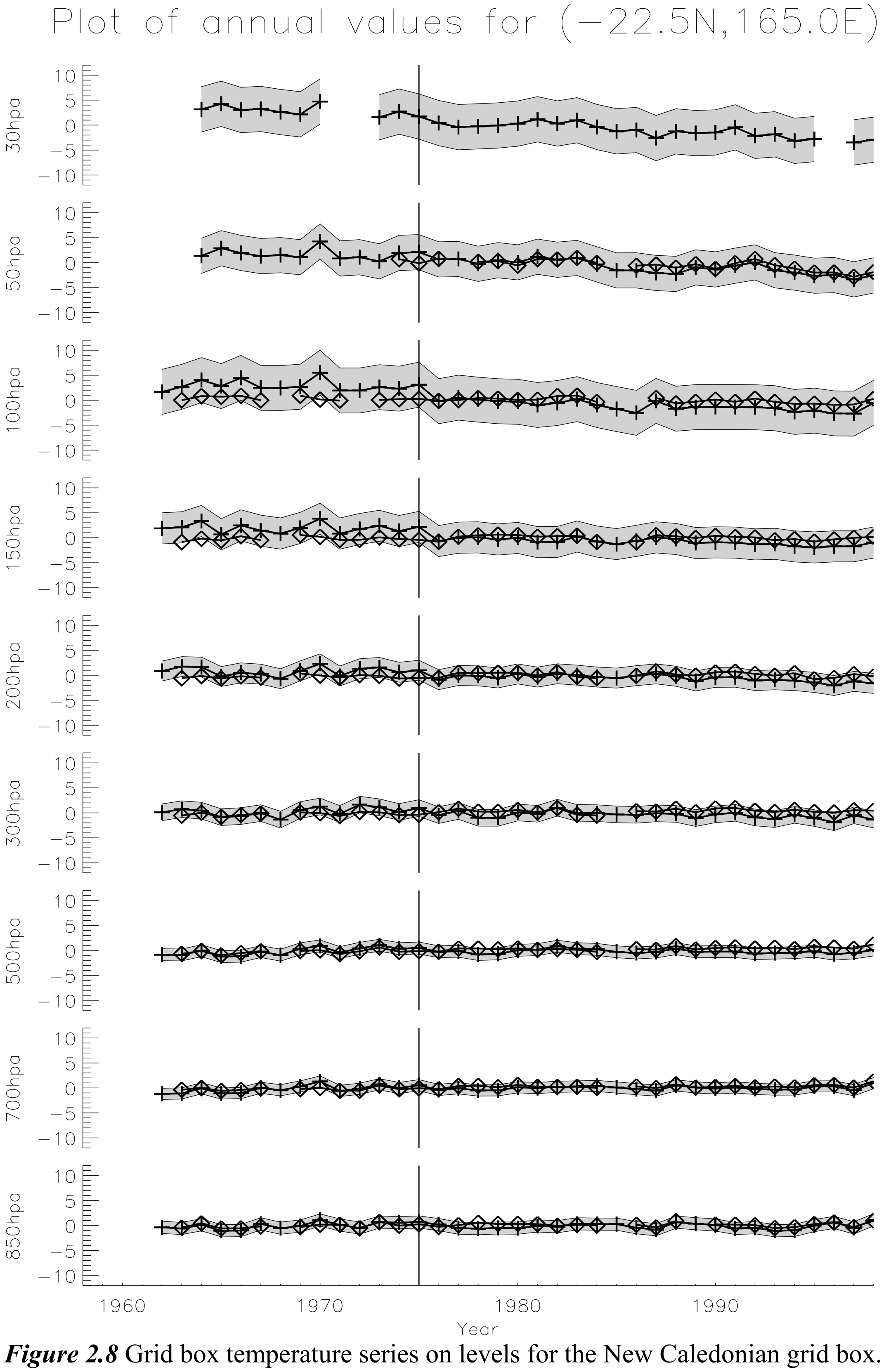

Figure 2.8 Grid box temperature series on levels for the New Caledonian grid box. Observations are represented by crosses, with shading to ± 2σ ; near neighbour averages are represented by diamonds. Note significant warm anomaly at height in the early record.

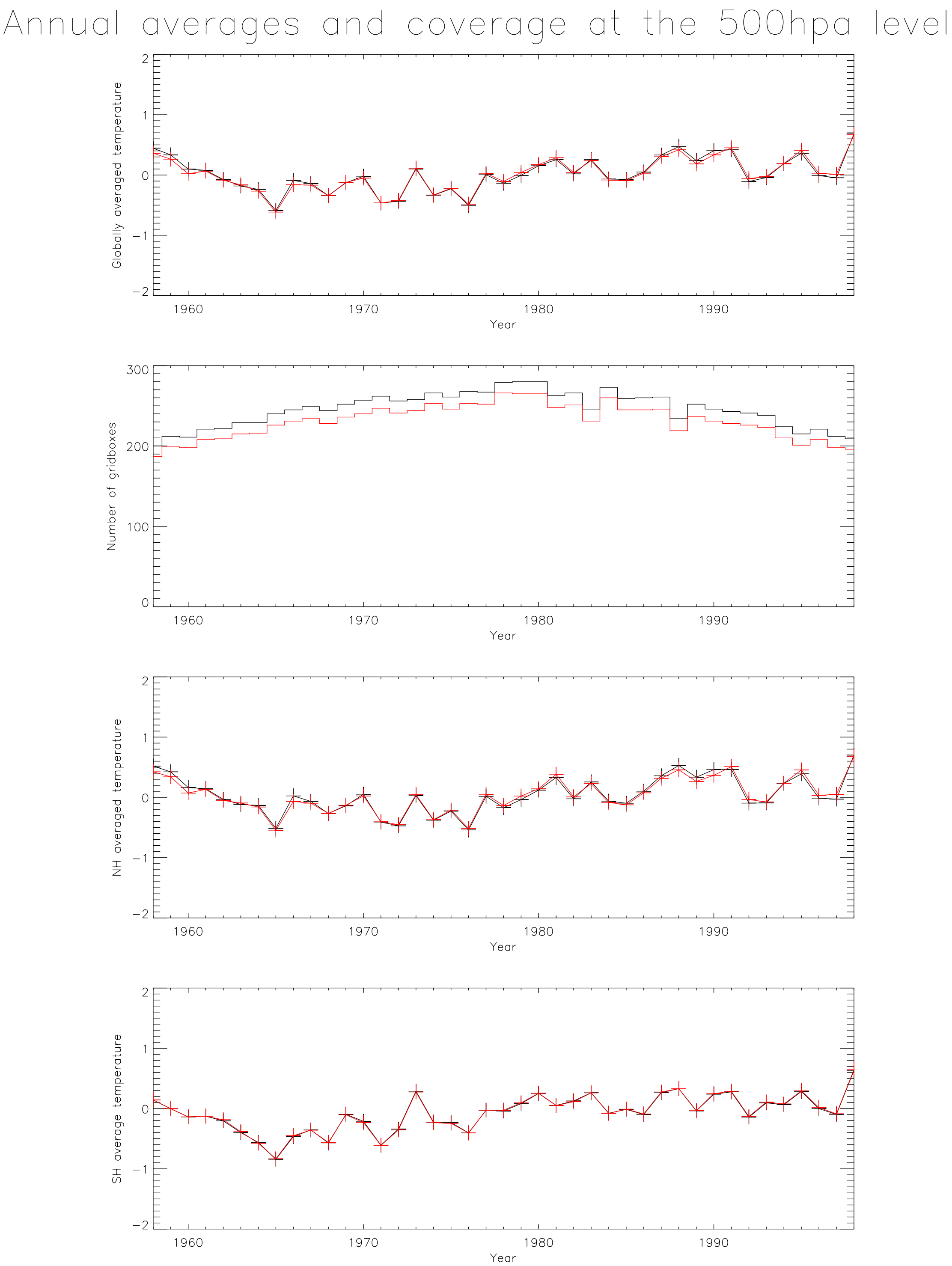

Figure 2.9 Plot showing the effect of editing the HadRT dataset, based upon quality control checks, on global and hemispheric mean temperatures and coverage at the 500hPa level. The unedited version is shown as a black line and is the same as the red line in Figure 2.2; the red line is directly analogous to the green line in Figure 2.1 and shows the effects of editing.

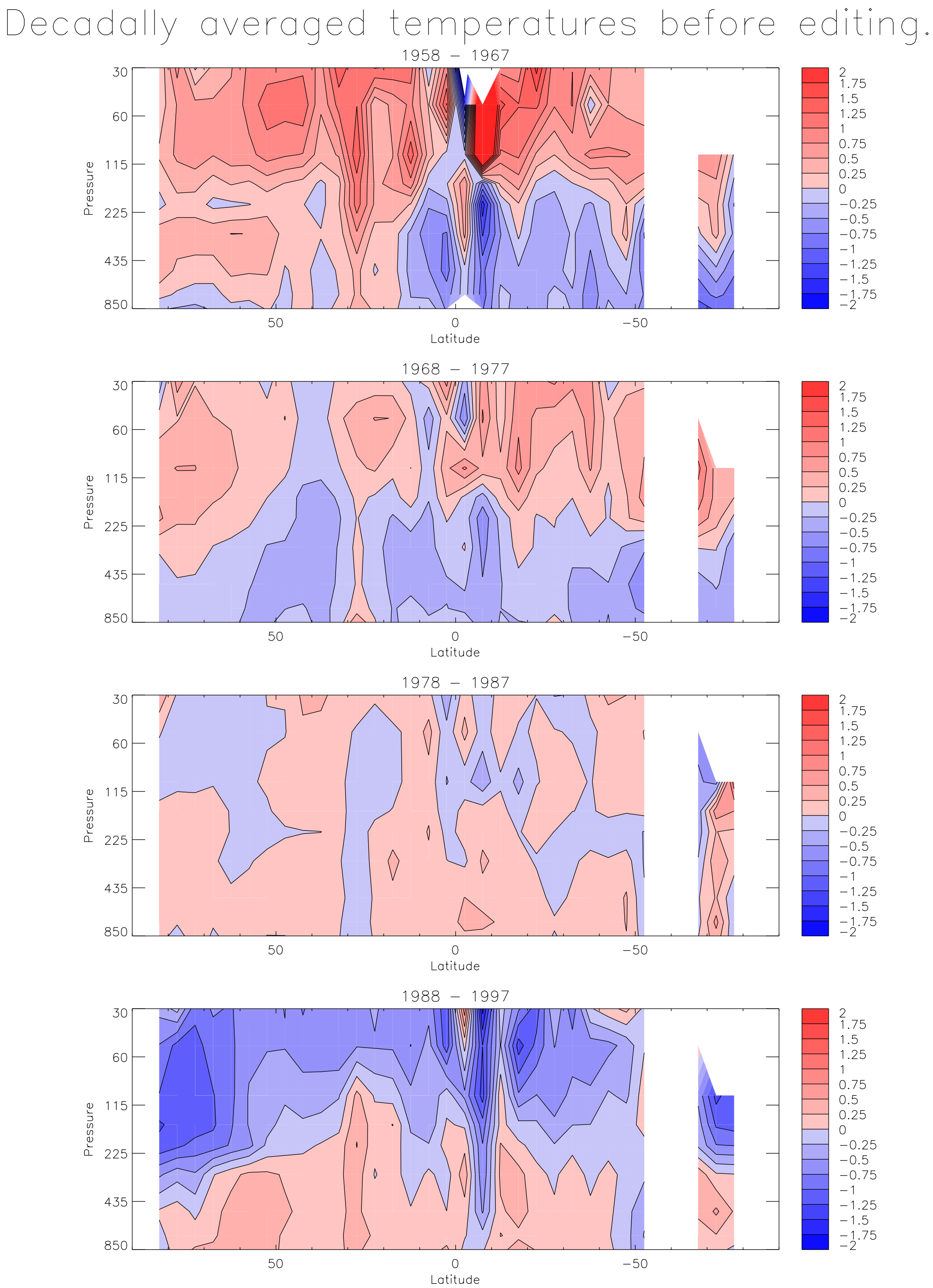

Figure 2.10(i) Decadally averaged zonal mean temperatures from V2 of the HadRT 2.1s data as anomalies from 1971-90 climatology. Contour intervals are at 0.25 degrees Celsius.

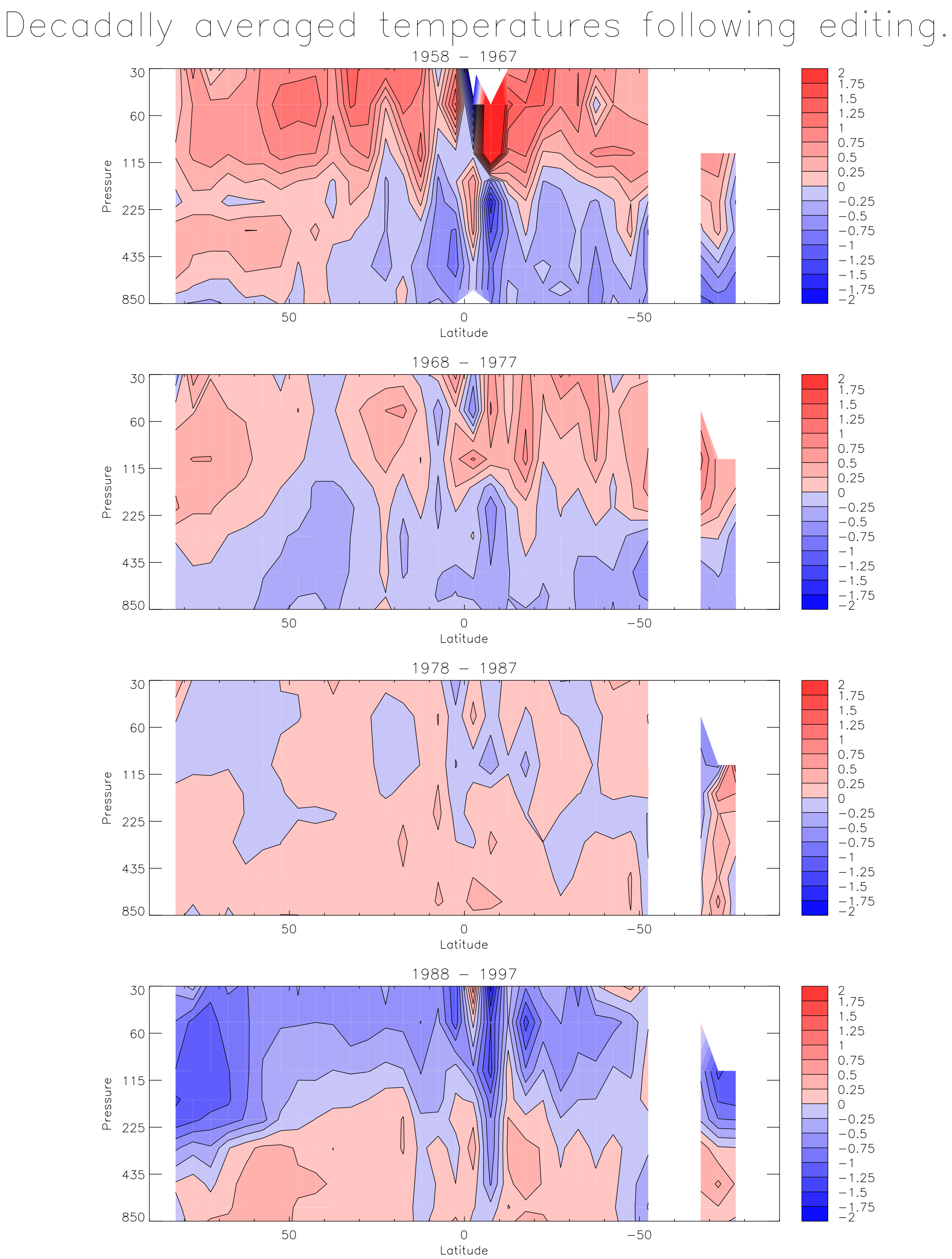

Figure 2.10(ii) Decadally averaged zonal mean temperatures from V2 of the edited HadRT 2.1s data, following quality control analysis, as anomalies from 1971-90 climatology. Contour intervals are at 0.25 degrees Celsius.

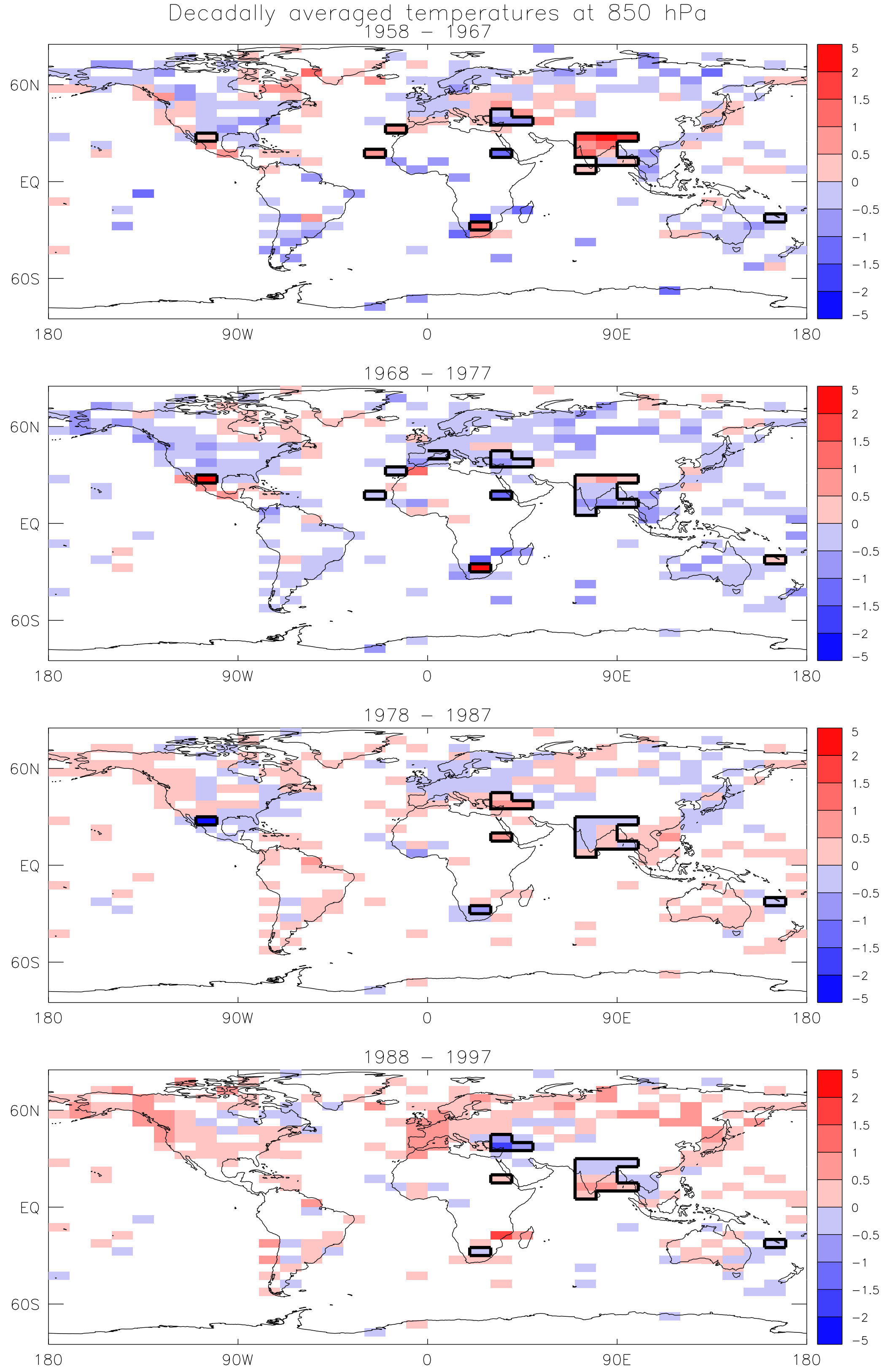

Figure 2.11. Figure showing decadally averaged temperatures at the 850hPa level in the original HadRT version. Regions edited in the final dataset are denoted by boxes.