Revisiting radiosonde upper air temperatures from 1958 to 2002

Peter W. Thorne,[1] David E. Parker,[1] Simon F. B. Tett,[2] Phil D. Jones,[3]
Mark McCarthy,[1] Holly Coleman,[1] and Philip Brohan[1]

Received 30 December 2004; revised 3 May 2005; accepted 28 June 2005; published 30 September 2005.

[1] HadAT is a new analysis of the global upper air temperature record from 1958 to 2002 based upon radiosonde data alone. This analysis makes use of a greater number of stations than previous radiosonde analyses, combining a number of digital data sources. Neighbor buddy checks are applied to ensure that both spatial and temporal consistency are maintained. A framework of previously quality controlled stations is used to define the initial station network to minimize the effects of any pervasive biases in the raw data upon the adjustments. The analysis is subsequently expanded to consider all remaining available long-term records. The final data set consists of 676 radiosonde stations, with a bias toward continental Northern Hemisphere midlatitudes. Temperature anomaly time series are provided on 9 mandatory reporting pressure levels from 850 to 30 hPa. The effects of sampling and adjustment uncertainty are calculated at all scales from the station series to the global mean and from seasonal to multidecadal. These estimates are solely parametric uncertainty, given our methodological choices, and not structural uncertainty which relates to sensitivity to choice of approach. An initial analysis of HadAT does not fundamentally alter our understanding of long-term changes in upper air temperature changes.

Citation: Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan (2005), Revisiting radiosonde upper air temperatures from 1958 to 2002, *J. Geophys. Res.*, *110*, D18105, doi:10.1029/2004JD005753.

1. Introduction

[2] Differences in temperature trends over the satellite era between the surface and the (lower) troposphere have been the cause of much controversy [e.g., *National Research Council*, 2000; *Santer et al.*, 2003; *Seidel et al.*, 2004; *Fu et al.*, 2004; *Tett and Thorne*, 2004]. Most, but not all, tropospheric temperature data sets exhibit less warming in the global mean than that reported at the surface. The discrepancy arises primarily in the tropics and Southern Hemisphere [*Brown et al.*, 2000; *Gaffen et al.*, 2000; *Hegerl and Wallace*, 2002]. It may be real, in which case it is important that we understand the underlying mechanisms. Equally, some or all of it may arise because of unresolved residual data set errors [*Seidel et al.*, 2004] or sampling issues [*Fu et al.*, 2004; *Free and Seidel*, 2005].

[3] Confidence in the veracity of upper air temperature trends is relatively low [*Seidel et al.*, 2004], particularly where the observational radiosonde network is sparse. However, recent efforts by a number of research centers, particularly the NOAA National Climatic Data Center (NCDC) in its role as a World Data Center, have recovered a wealth of additional radiosonde data. Radiosondes have been used to monitor "global" changes since the International Geophysical Year (IGY) in 1958, 20 years before the Microwave Sounding Unit (MSU) satellite records began. Hence they are useful to evaluate observed satellite period changes in the context of longer-term change and variability.

[4] There already exists a range of radiosonde (*Parker et al.* [1997], *Sterin* [1999], GUAN [*McCarthy*, 2000], *Angell* [2003], *Lanzante et al.* [2003a, 2003b] (LKS henceforth)), MSU [*Christy et al.*, 2003; *Mears et al.*, 2003; *Grody et al.*, 2004], and reanalysis [*Kalnay et al.*, 1996; *Uppala et al.*, 2005] tropospheric and stratospheric temperature products. These exhibit a marked spread in their long-term trends [*Seidel et al.*, 2004]. There is no historical transfer standard allowing unambiguous quantification of the effects of known and suspected nonclimatic effects in the observations. Hence there will remain uncertainty in climate records arising through seemingly sensible choices made during their construction [*Thorne et al.*, 2005]. It is vitally important to develop multiple climate data sets using distinct approaches to see what the range of "plausible" trends is [e.g., *Seidel et al.*, 2004, *Thorne et al.*, 2005]. HadAT has been constructed with this requirement in mind.

[5] Two of the radiosonde data sets [*Sterin*, 1999; *Angell*, 2003] make no attempt to account for nonclimatic influences. Three are small subsets of the global radiosonde network [*McCarthy*, 2000; LKS; *Angell*, 2003]. *McCarthy* [2000] and LKS, although adjusted, have made no reference to background fields so there is no guarantee that large-scale spatiotemporal consistency will be retained. The true climate system displays marked covariance such that

[1]Hadley Centre for Climate Prediction and Research, Met Office, Exeter, UK.
[2]Hadley Centre, Reading Unit, Reading University, Reading, UK.
[3]Climatic Research Unit, University of East Anglia, Norwich, UK.

Published in 2005 by the American Geophysical Union.

Table 1. Raw Data Set Sources Used in HadAT[a]

Data Set	Number of Stations in Data Set	Number of HadAT1 Stations Chosen From the Data Set	Number of Extra HadAT2 Stations Chosen From the Data Set	Launch Times, UTC	Adjustments Undertaken Following Receipt by Data Centers
CLIMAT TEMP	737	40	28	00+12 mix	real-time quality control post-1995
MONADS 12	2129	45	3	12	CARDS
MONADS 00	2129	84	29	00	CARDS
MONADS 00+12	2129	215	111	00+12 mix	CARDS
GUAN	152	44	20	00+12 mix	median fit of available records
LKS 12	65	3	0	12	expert review of CARDS data
LKS 00	75	5	1	00	expert review of CARDS data
LKS 00+12	87	41	7	00+12 mix	expert review of CARDS data

[a]A station is counted if it contains at least one month's data for the given database. A launch time "00+12 mix" is a combination of all available data, but at any given time there may be solely 00 or 12 UTC data.

variations in one location tend to be associated with changes over much broader regions. Adjustments applied without reference to a background field may yield a physically implausible solution. HadRT [*Parker et al.*, 1997] has the coverage and adjustments approach to make it suitable for global process analyses. However, adjustments are only applied from 1979 when MSU data started and require supporting metadata, which are known to be incomplete (*Gaffen*, 1996 and subsequent updates). Subsequent study has shown that spatiotemporal consistency is unlikely to have been retained [*Thorne et al.*, 2002, 2003]. Most importantly it is no longer independent of at least one version of the MSU record [*Christy et al.*, 1998]. HadAT has been constructed as a truly global, spatiotemporally consistent radiosonde product to fill this perceived gap.

[6] The rest of this paper describes HadAT, available online at http://www.hadobs.org/. Available digital radiosonde records and the derivation of a climatically useable subset are detailed in section 2. The methodology for the development of neighbor composites is outlined in section 3. Section 4 describes the Quality Control (QC) procedure which was undertaken to yield a homogeneous station set, and provides a range of case studies. Section 5 briefly describes the gridding methodology and the changes in data availability. Section 6 outlines the quantification of the errors in the resulting data set. A brief initial analysis of HadAT is given in section 7, while section 8 concludes.

2. Data Set Sources and Selection of Climatically Useful Stations

[7] All available digital station-level data were collated (Table 1), yielding many radiosonde station records (Figure 1). Data were extracted for the nine standard WMO reporting pressure levels (850, 700, 500, 300, 200, 150, 100, 50 and 30 hPa) which are common to all sources.

[8] CLIMAT TEMP data are monthly averages taken directly from the Global Telecommunication System. Limited quality control has been undertaken [*Parker and Cox*, 1995]. They are available only for a "mix" of launch times. Following the IGY most stations have had a two/day launch schedule at 00 and 12Z (UTC). Some, however, have launched once daily, or at nonstandard times, and a very limited number made 4 or more launches daily.

[9] MONADS (MONthly Aerological Data Set) data (available from NCDC) are monthly summaries of the launch resolution CARDS [*Eskridge et al.*, 1995] database. Data are available at launch hour–specific times. Three composites were created for each station: 00Z, 12Z, and mix (a simple average of available 00Z and 12Z monthly mean data). MONADS is in the process of being superseded by IGRA (Integrated Global Radiosonde Archive) [*Durre et al.*, 2005a], which was not available at the time of this analysis. An initial comparison of MONADS and IGRA yields some random differences resulting from the different QC algorithms but no pervasive systematic differences.

[10] The Global Climate Observing System (GCOS) Upper Air Network (GUAN) consisted (when this analysis was developed) of a baseline global network of 152 stations. The Hadley Centre in its role as GUAN analysis center has retrieved all versions of digital records for these stations. A median fit selection procedure and limited QC has been performed to gain a best estimate of the true station record [*McCarthy*, 2000]. The resulting database is available only for a mix of launch times.

[11] LKS have intensively QCed a set of 87 well-spaced long radiosonde records from the CARDS database. Numerous indicators were used to ensure that real breakpoints were identified and adjusted for. Adjustments were made on a station-by-station basis. The homogenized time series were in closer agreement with the MSU satellite record [*Christy et al.*, 1998]. LKS is available as at least one of 00Z, 12Z and mix for 87 stations up until 1997. LKS produced several versions of their data set, of which their recommended LIBCON (LIBeral and CONservative adjustments applied) version is used here. 62 of the 87 LKS stations are also GUAN stations.

[12] A subset of sufficiently long and complete station records was extracted from these data sources. Calculation of a monthly value required at least 12 ascents. For any three-month season to be counted required at least two months of data, and for an annual value at least three seasons had to report. These criteria were applied on a level-by-level basis. Stations (and levels) without at least five years of annual data in each of the three decades used to create normals were excluded. The optimal normals period was assessed by passing these criteria over a moving 30-year climatology window from 1961–1990 to 1971–2000 and counting the number of stations for which at least one level had a climatology. There were significantly more stations (order 100+ (15%)) for 1966–1995 than either of the "standard" WMO climatology periods. Given the sparsity of stations this nonstandard climatology period is used. The resulting coverage is heavily skewed toward Northern Hemisphere continental locations (Figure 1). A full listing of all the stations is

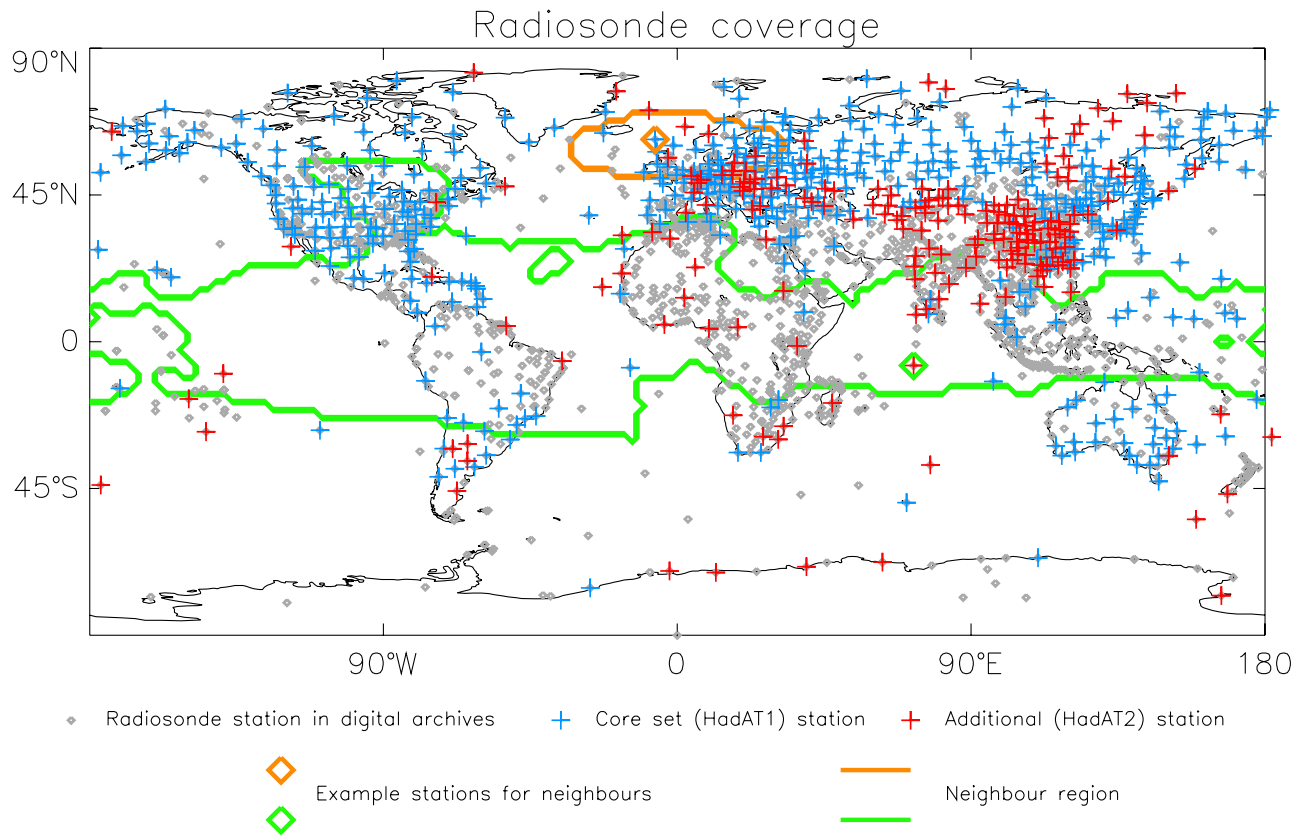

Figure 1. Locations of all available digital radiosonde records from the databases listed in Table 1, stations identified as our core set (for HadAT1), and additional stations (HadAT2). Stations were not used in HadAT unless a 1966–1995 climatology could be calculated. Stations with tropospheric thickness anomalies considered sufficiently similar to our LKS/GUAN network were included in HadAT1. Also shown are two example neighbor regions at 500 hPa for the Northern Hemisphere summer (JJA).

available in Supplementary Table 1[1]. For each station up to eight versions of the climatologies and anomaly time series were calculated (GUAN, CLIMAT TEMP, LKS (00, 12, and mix), and MONADS (00, 12, mix)).

3. Construction of Neighbor Composites

[13] The rationale behind constructing HadAT is to create a consistent depiction of changes in upper air temperatures. A (quasi-) consistent independent background field is required for each station to enable adjustments to be calculated to compensate for any biases.

[14] There are three obvious candidates: a neighbor composite, a reanalysis data set, or a satellite record. It is important to retain independence from satellites to enable truly independent checks on satellite records, so their use was rejected. Reanalyses are constrained by both the data presented, which include satellites and other surface and radiosonde data (all of which contain nonclimatic influences), and the model physics. As a result over the long time periods of interest here they contain systematic large-scale time-varying biases [*Simmons et al.*, 2004; *Bengtsson et al.*, 2004; *Sterl*, 2004]. Attempts have begun be made to use background fields and to remove these systematic effects [*Haimberger*, 2005]. At the time of the HadAT analysis such methods were unavailable.

[15] Therefore a sonde-based neighbor composite series was chosen. It is necessary to minimize the chances of introducing systematic biases into the neighbor series. The method used to do this makes two basic assumptions. The first is that at least a subset of the station series is free of gross inhomogeneities. The second is that any remaining inhomogeneities in this station subset are effectively randomly distributed in sign, magnitude, and timing such that when a number of neighbors are averaged together these will inflate the neighbor-based composite time series variance rather than add significant bias to the neighbor estimate. It is not possible to entirely objectively test either of these assumptions. Therefore the subset was chosen as LKS and GUAN [*McCarthy*, 2000] as they are the least likely to retain gross inhomogeneities. For those stations where data exist for both, LKS were used as their data had been much more rigorously investigated.

3.1. Identification of a Core Set of Station Records

[16] LKS and GUAN are too sparse to form neighbor estimates on individual levels for all the "climatically useable" stations (Figure 1) at the seasonal resolution deemed necessary to accurately quantify adjustments. Therefore this station set was expanded to include grossly

[1]Auxiliary material is available at ftp://ftp.agu.org/apend/jd/2004JD005753.

Table 2. Adjustments Applied to HadAT1 Stations by Iteration of Our QC Procedure[a]

Iteration	Breakpoints	Adjustments	Mean, K	Median, K	Mean Absolute, K	Median Absolute, K	Standard Deviation of Absolutes, K
1	1451	10032	−0.013	−0.104	0.570	0.452	0.457
2	1224	7808	−0.001	0.070	0.469	0.387	0.387
3	690	3653	0.001	−0.081	0.402	0.325	0.325
4	191	929	−0.020	−0.114	0.341	0.280	0.280
5	37	177	0.033	0.141	0.350	0.272	0.272

[a]Breakpoints are identified as unique points in a station time series when an adjustment was required, whereas number of adjustments indicates the number of levels upon which adjustments were applied. The adjustment factor was applied to all points before the breakpoint within a time series.

similar stations. This has the advantage of reducing noise in the neighbor series, and hence the uncertainty in the adjustments required.

[17] Monthly temperature anomalies (relative to 1966–1995) were used to derive annual layer thickness anomaly series for the troposphere (700 to 300 hPa) and lower stratosphere (300 to 100 hPa; in the tropics this is mainly in the upper troposphere), for all stations. It is assumed that temperature anomaly errors are coherent within these layers. The effects of using actual rather than virtual temperatures will be minimal [*Elliott et al.*, 1994]. For each station, series were calculated from each individual available source including LKS and GUAN. So for any station there were up to eight versions of the two thickness anomaly series.

[18] Weighting coefficients for the calculation of a neighbor composite thickness series were derived from NCEP reanalysis data [*Kalnay et al.*, 1996] for 1979–1998. For those reanalysis grid boxes containing the climatically useful radiosonde stations, the contiguous surrounding region with an annual correlation greater than 1/e for each of the two deep layer thicknesses was identified. As atmospheric spatial structure is being considered and this is a period of relatively stable data input, the reanalyses should be adequate. Unlike *Wallis* [1998] no a priori assumptions were made as to the likely shape of the regions. For each station any GUAN/LKS neighbor series which fell within the region were identified. If one or more of these records existed then a neighbor composite was created for the target station from the annual thickness anomaly series of these neighbors. The component neighbor thickness series were averaged, weighting each neighbor by the expected correlation. A number of GUAN/LKS series were deemed to be highly dubious when compared to their neighbor composite estimates, including a qualitative time series analysis. Either a portion of these or their entire records were omitted (Supplementary Table 2) in defining the neighbors used to assess the network of adequate stations.

[19] The neighbor average layer thickness series were compared with each available version of the target station thickness series. Peter Thorne (PWT) decided whether any version was sufficiently similar to the neighbor composite; and if so which version to use in subsequent analysis. Two statistical indicators were employed to aid the decisions – the correlation between the series and the average z score, providing a standardized mean departure from the neighbor expectation defined as:

$$Z - score = \left| \frac{station - neighbours}{\sigma station} \right| \quad (1)$$

The average of the weights used in the creation of the neighbor composite provided an expectation of the correlation. The target station series was rejected outright if the actual correlation was more than 0.1 lower than this. PWT also considered the number of years for which an annual average could be calculated. Generally, the time series with best agreement was chosen. In those cases where the degree of agreement according to the simple indicators was deemed equivalent either LKS/GUAN (if available) or otherwise the most complete record was chosen.

[20] There was often a large range in the indicators for those target stations containing data from different sources. For GUAN and LKS records these differences are likely to result from postprocessing which has been deliberately applied. However, for many stations there were also differences between MONADS and CLIMAT TEMP which must relate to sampling and processing effects before and/or during digitization. In many cases the differences were as large as those between the deliberately adjusted GUAN and LKS records. It is important to retain both the raw data and a full audit trail so that any differences can subsequently be reconciled and understood [*Durre et al.*, 2005b].

[21] There were more target stations with adequate tropospheric than stratospheric layer thickness series agreement. To avoid artificially degrading coverage in the troposphere all stations for which tropospheric thickness series were deemed adequate were retained. Decisions to reject stations disproportionately affected coverage in certain regions (Figure 1). Nearly all data from southern Asia and tropical Africa as well as a large strip of data from eastern Europe were rejected. The 477 retained stations are concentrated in Northern Hemisphere continental regions. However, there are still some from both tropical and Southern Hemisphere midlatitude regions (Figure 1).

[22] Chosen station series span all data sets (Table 1). When both ascent times were available separately often a single ascent time was chosen, as the statistical match to neighbors was much worse for both mixed ascent times series and particularly the other ascent time series. This is consistent with the finding of LKS that 00–12Z differences were a powerful breakpoint indicator. Choice of launch time was most important around 45–135°E and 45–135°W. This implies poorly resolved or implemented radiation corrections as these are sensitive to the low solar elevation angles at launch time at these longitudes.

[23] This "raw" station data set is HadAT0. In reality all these station series have had some form of postprocessing applied at retrieval time, in retrospect, or both, so HadAT0 is not the actual raw observed time series.

3.2. Creating Neighbor Estimates on Individual Pressure Levels

[24] Neighbor averages for the QC procedure were created in a very similar way. NCEP reanalysis data for 1979–

1998 were used to calculate estimates of the seasonal (DJF, MAM, JJA, SON) temperature correlation fields on all 9 pressure levels for grid boxes that contain HadAT0 stations. For each station a seasonal neighbor composite series was created for each level using stations in the contiguous region with correlation greater than 1/e. Figure 1 includes a couple of example neighbor regions at the 500 hPa level for DJF. These regions varied on both a seasonal and a level-by-level basis. Neighbor station series were weighted by the expected correlation to produce the neighbor average composite. Within the stratosphere (above 300 hPa) only those HadAT0 stations which were deemed adequately similar to LKS/GUAN in the stratospheric layer thickness analysis were used in the neighbor composites.

4. Quality Control Procedure

4.1. Correcting the HadAT0 Stations

[25] A seasonal mean difference series for each station series at each level was calculated: station time series minus neighbor time series. If the target station series is a realistic representation of the true climate evolution, and the neighbor series is similarly free of systematic biases, this difference series will be indistinguishable from white noise with a zero mean. This is the basic assumption of all climate anomaly homogeneity approaches [*Conrad and Pollak*, 1962]. The main interest is in long-term trends so the primary aim was to identify and adjust for systematic changes. A nonparametric Kolomogorov-Smirnov test [*Press et al.*, 1992] (KS-test) was passed through the difference series to identify suspected breakpoints. This test can be interpreted as returning the probability that two populations arise from the same distribution. The KS-test was applied to each time series with a 15 season window either side of the current point. Cases at the 10% level or lower were highlighted as suspected breakpoints (Figures 2 and 3). Note that a nonparametric test is weaker (will yield fewer suspected breakpoints) than a parametric test, e.g., a student's t-test.

[26] For each station, including LKS and GUAN stations, a plot similar to that in Figures 2 (left) and 3 (left) was produced. Figures 2 and 3 are for two stations randomly chosen to illustrate our procedures. On the basis of these plots PWT identified times where the KS-test identified a vertically coherent jump point in the difference series. The station series and neighbor series helped in deciding whether a break point resulted from problems in the station or the neighbors. Only in a handful of cases were the neighbors deemed to be the most likely cause. Having identified suspected breakpoints, recourse was made to available metadata (*Gaffen* [1996] and updates) to try to determine an exact date. This was limited to static metadata change point events, i.e., those given a definite timing. If PWT decided there was sufficient evidence for a break in the station time series then a breakpoint was assigned and adjustments implemented as well as noting the metadata event, if any. Inevitably this step required subjective judgment. As it is informed by quantitative measures and knowledge of metadata events (where available) and factors which might impact the difference series (ENSO, explosive volcanic eruptions, etc.), it need not add any significant overall bias.

[27] A bootstrap type approach was used to estimate the required adjustment factor at each breakpoint. Adjustments at each level were defined as the change in the mean of the difference series between the ten years before and after, or a shortened period so as not to overlap with the next breakpoint. To verify this adjustment factor 1000 additional estimates were created. A random number generator was used to define what proportion, up to 40%, of values to omit from the neighbor difference series. This proportion was calculated independently either side of the break point, e.g., 5% could be dropped from one side and 25% from the other for a given estimate. A second random number generator provided an index of times to be dropped. These subsampled series were used to create an estimate of the required adjustment factor. By randomly dropping values, bimodal or multimodal distributions result if there are dubious value(s) present as these bias the solutions only when they are included.

[28] A number of checks were performed on the population of adjustment factor estimates:

[29] 1. The first check was to ensure that the adjustment factor is significantly nonzero: Are the 5th and 95th percentiles of the adjustment estimates distribution of the same sign?

[30] 2. The second check was to test whether the population of estimates is normally distributed: (1) Is the 1st (99th) percentile within 1.5 ± 0.4 times the 5th (95th) percentile distance from the median? (2) Are the fifth and ninety-fifth percentiles approximately equidistant from the median value? (3) Are the initial estimate and the median of the population within 0.03 K or 25% of the absolute value of the median adjustment?

[31] 3. The third check was to check for grossly erroneous values: Are all absolute seasonal difference values <4 K?

[32] If all three tests passed then the median value was used as the best guess adjustment factor.

[33] If any of the tests failed then any values deemed by PWT to be obviously dubious in the context of the rest of the difference series were deleted. If values were deleted then the adjustment calculation procedure was repeated. In total order 1–2% of seasonal values were deleted. Soviet data until the mid-1960s were found to be highly suspect in the winter season at all heights, but particularly in the stratosphere (Figure 4, left). The absolute differences to the neighbor composite series were often >10 K (the time series shown are temporally smoothed), whereas subsequently they were generally within the range ±2 K. A number of stations from developing countries were also particularly poor. Conversely, relatively few deletions were made for U.S., Canadian, Australian, Japanese, and NW European series.

[34] Only significantly nonzero change points were adjusted. Implementing small and insignificant adjustments could artificially redden the spectrum by adding spurious step changes to the time series. Adjustments were applied as seasonally invariant changes to all points in a station time series before the break point.

[35] Once decisions for all HadAT0 stations regarding adjustments/deletions had been made, they were implemented and the seasonal climatologies recalculated. The adjusted series were then used to create new neighbor

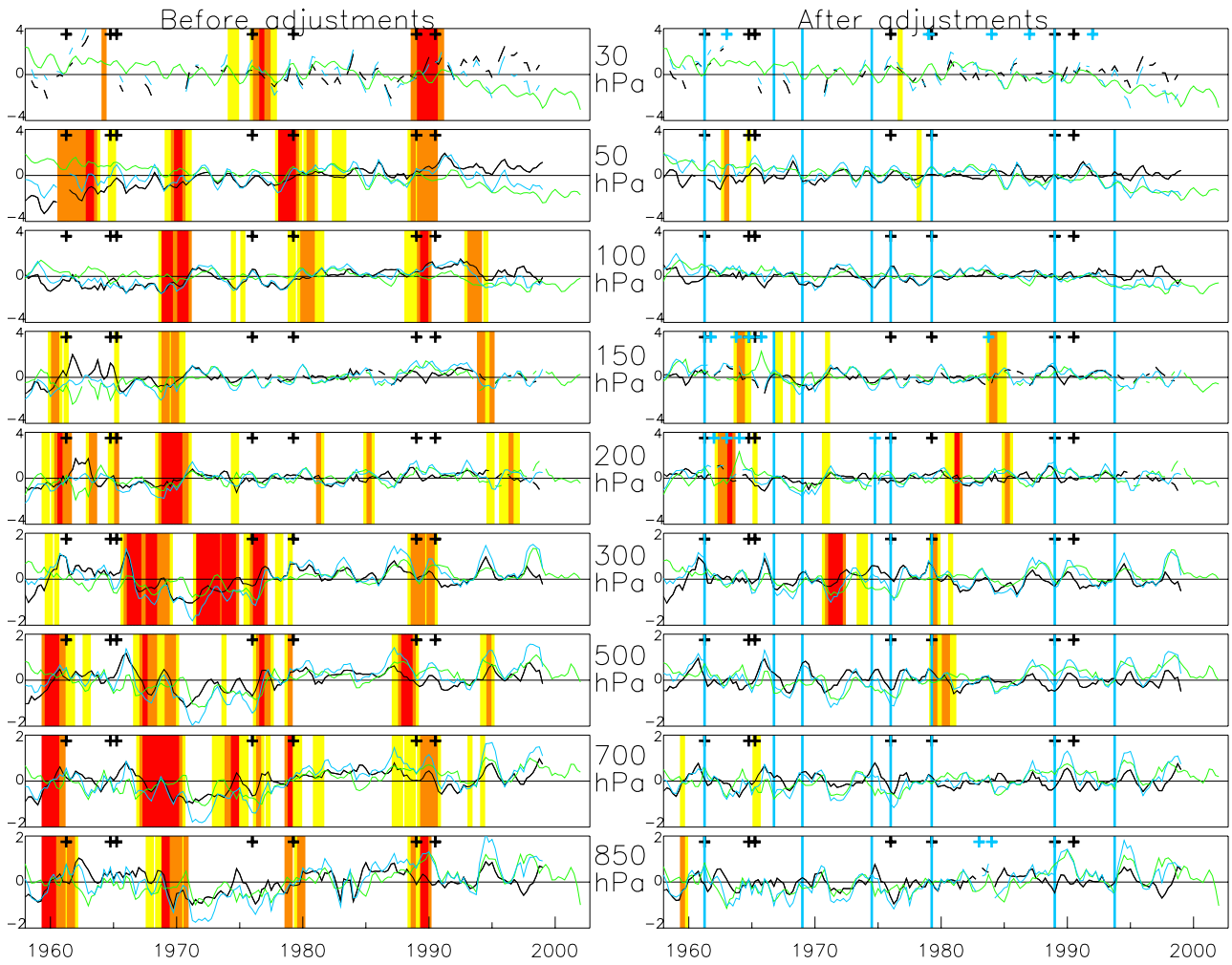

Figure 2. Time series plots for station 8495 (Gibraltar) (left) before and (right) following the QC procedure. Plots are for 9 levels (30 hPa to 850 hPa). Each plot shows station time series (blue), neighbor series (green), and difference series (black). All time series have had a simple seven-point filter applied. For levels above 300 hPa the y axis range is -4 to 4 K, and below is -2 to 2 K. Superimposed on each plot are static metadata events (black crosses). The KS-test statistic results are denoted by vertical bars for differing probabilities below 0.1 (<0.01 red, >0.01 and <0.05 orange, >0.05 and <0.1 yellow). The metadata and KS-test indicators taken together with the time series characteristics were used to guide expert judgment as to the locations of breakpoints. Figure 2 (right) additionally shows blue crosses where deletions were implemented and vertical blue bars where adjustments were applied. Note that several iterations of the procedure were performed and at these intermediate steps additional breakpoints may have been identified as the station and neighbors series were made more homogeneous.

composite and difference series and the quality control procedure repeated. On the first iteration, only breakpoints which PWT assessed as very definite breaks in the station data were adjusted, to minimize the chances of aliasing spurious neighbor series trends into the adjustments. In subsequent iterations all suspected breakpoints were considered and, where significant, adjusted. Once a station had no adjustment or deletion applied on a given iteration of the procedure it was considered homogeneous and no longer a candidate for future adjustment. This prevented the procedure from forcing each station series to become identical by iterating indefinitely. The entire QC procedure was carried out a total of five times, after which PWT decided that convergence had been attained. We caution that another expert or group of experts (e.g., the LKS approach) may have reached different decisions in performing this QC so there are questions as to repeatability.

[36] Following QC a final check for outliers was performed removing all values greater than 3.5σ in the homogenized difference series from the target station series. This led to the further removal of 0.05% of points. Some of these values might be real extreme events. However, the primary interest is in characterizing the long-term behavior of upper air temperatures. Hence it is more important to remove erroneously large anomalies which could have a disproportionate influence. The approach may artificially reduce the interannual variability.

[37] For HadAT0 stations at all pressure levels the final difference series is closer to random noise around zero than the initial version (e.g., Figures 2, 3, and 4). The

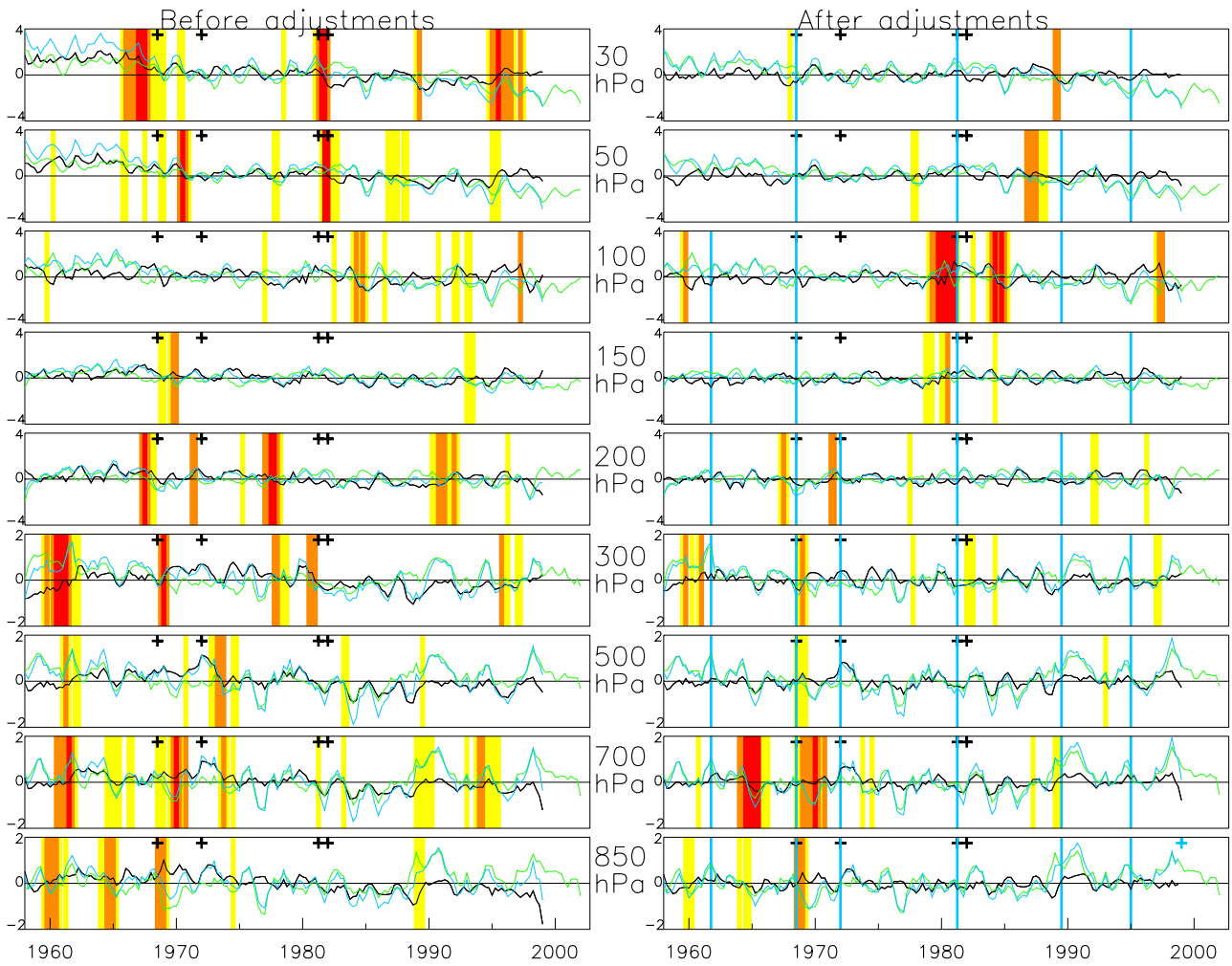

Figure 3. As Figure 2 but for station 47412 (Sapporo, Japan).

homogenized time series yield KS-test results that are approximately normally distributed, whereas the raw data KS-test results are highly negatively skewed implying the presence of discontinuities in these data (Figure 5).

[38] As the iterations proceeded fewer breakpoints were identified (and slightly fewer levels were adjusted per breakpoint) and the magnitudes of the adjustments decreased (Table 2). The distribution of absolute adjustment factors is highly positively skewed – there were a large number of relatively small adjustments and a small number of very large adjustments, especially in the initial iteration. There is little indication of a systematic sign of the adjustments – given the methodological approach this is not surprising. Although there are large variations between stations it is striking how invariant the average number of breakpoints identified per station by WMO region is (Table 3). For any station PWT identified on average 6 breakpoints over the 45 year period (1.3 per decade, although many stations are incomplete). The mean and median absolute adjustment factors applied are also similar except for North America and the Pacific region where they are lower, reflecting the traditionally higher quality stewardship by U.S. operators. The frequency of breakpoints identified is reduced at the ends of the record as a result of the reduced power of the KS-test and more conservative breakpoint identification undertaken by PWT when less than 15 seasons are available before or after the time step. Over the rest of the record there is little variation by decade. Station practices have been forecasting- rather than climate-driven and numerous changes to procedure have been and continue to be made.

[39] Particularly outside of developed nations there were few metadata, so most breakpoints identified (c. 70%) had no accompanying metadata (Table 4). This is a major impediment to the unambiguous identification and removal of nonclimatic influences. A subset of stations with seemingly complete metadata (the exception rather than the rule) from a range of countries yields an average of 13 metadata events per station over the HadAT period. So the average number of adjustments applied here per station may be an underestimate of the pervasiveness of nonclimatic influences and the series may retain heterogeneities. Alternatively, many metadata events may lead to no discernible influence on long-term continuity of the station records. Most metadata associated with the breakpoints were documented as either a change to the basic sonde model (or one or more of its components) or a change in the calculation methods, primarily how radiation effects were removed. The resulting data set is HadAT1.

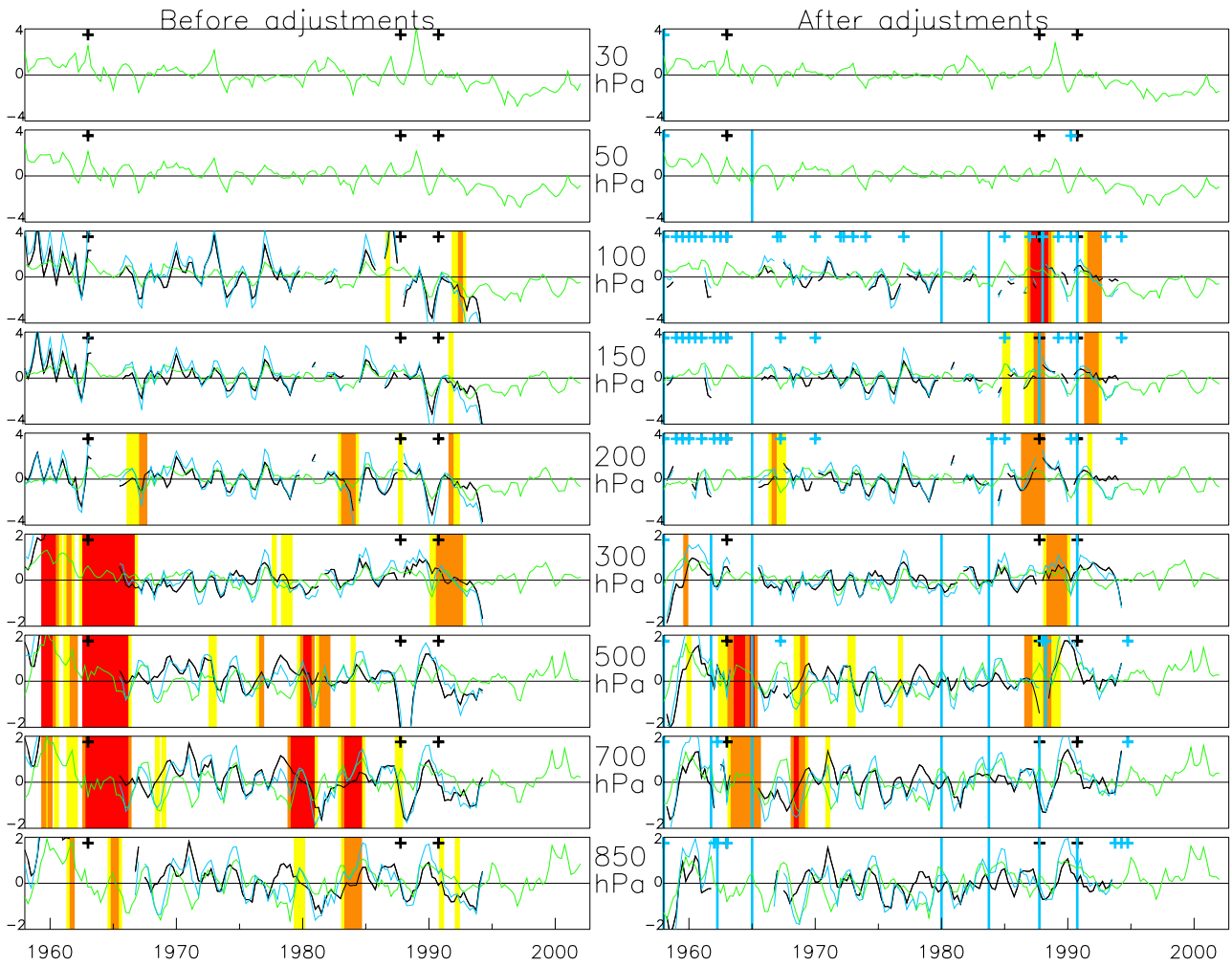

Figure 4. As Figure 2 but for station 20107 (Barencburg, Russia).

4.2. Expanding the Station Network

[40] Having homogenized HadAT0 stations to form HadAT1, those stations which were initially deemed to be insufficiently similar to the LKS/GUAN network were reconsidered. Adjusted HadAT1 stations were used to create neighbor composites for these stations, relaxing the stratospheric requirement so that all HadAT1 stations, which were now homogenized, contributed. Hence the neighbor series were not updated upon the completion of each iteration of the QC procedure. In all other respects the methodology was identical to that employed for the HadAT1 stations.

[41] Of the remaining stations for which it was possible to calculate a climatology, 199 were adjusted. The rest were either deemed by PWT to be too heterogeneous, without sufficient neighbors, or contained limited data for two levels at most. A total of four iterations were required to homogenize these series. The homogenized series pooled with the HadAT1 station series produce HadAT2.

[42] Previous investigations by LKS and *Parker et al.* [1997] concluded that Indian station data are highly dubious. However 15 Indian stations qualified for HadAT2 (Figure 1), These series did indeed exhibit large heterogeneities, having on a national average the largest discrepancies vis-à-vis the neighbor composites. However, it was relatively simple to identify breakpoints, many of which correlated with the available metadata. We see no compelling reason why the adjusted Indian data should not reflect the true long-term behavior, so long as the HadAT approach is sufficiently powerful and unbiased. Figure 6 gives temperature time series before and after adjustments for an example Indian station (cf. Figures 2, 3, and 4) showing that the most pervasive breakpoints have seemingly been removed.

5. Gridding Methodology and Network Reporting Performance

[43] Having completed the QC procedure, HadAT2 and the intermediate products HadAT0 and HadAT1 were gridded. These gridded products are available on the data set website along with the station records. For consistency with the HadRT data set the station data were gridded onto a 10° longitude by 5° latitude grid. The larger correlation scales in the free atmosphere justify this grid box scale which is larger than that of the HadCRUT2v surface time series which are available on a 5° by 5° grid [*Jones and Moberg*, 2003]. Where more than one station contributed to a grid box the grid box time series was taken as the simple average of the available station values.

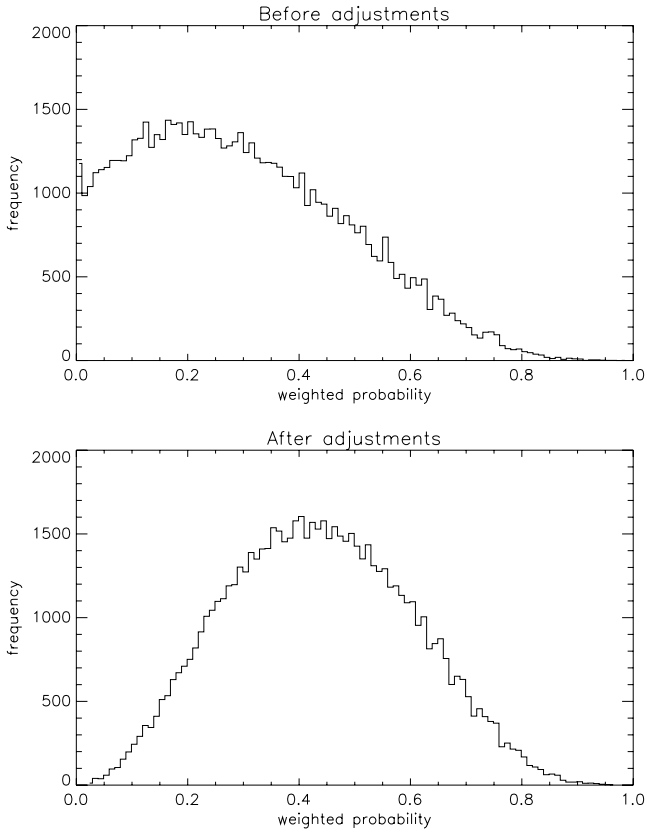

Figure 5. Summary of KS-test results before the first iteration and following completion of HadAT1. For each station at each time step the results have been multiplied together and then renormalized by taking the power 1/n where n is the number of pressure levels with a KS-test result. Probabilities are of the truth of the null hypothesis of no breakpoint, so that low probabilities suggest a discontinuity. The statistic is constrained to lie between 0 and 1 and would have a mean of 0.5 for simple white noise. Taking the geometric mean reduces this expectation to c.0.4 if there are nine levels with data upon which the test is performed.

[44] The HadAT1 and HadAT2 gridded products consist of many grid boxes containing no data, many containing one or two stations and a few with up to 8 contributing stations (Figure 7). Both HadAT1 and HadAT2 are primarily Northern Hemisphere continental data sets. In incorporating the additional HadAT2 stations this bias has been ameliorated, improving in particular coverage over Africa, Southern Asia, and the Southern Pacific Ocean. There are also more stations in HadAT2 than in HadAT1 in some grid boxes where both have data.

[45] Not all stations contribute to a grid box value for a given time and many only for a subset of levels. Especially those grid boxes consisting of one or two stations may contain significant periods of missing data, and remaining grid boxes exhibit variations in grid box sampling density. Station attendance by WMO region and by level (Figure 8) varies greatly over time. Coverage drops off significantly above 100 hPa, with a large drop in Northern Hemisphere sampling in winter (relating to balloon burst in extreme cold so that the climatology criteria were not met) leading to a pronounced seasonal cycle in the HadAT coverage at these altitudes. Up to 100 hPa the coverage is relatively seasonally and vertically invariant. The analysis here has made no attempt to account for the time-varying sampling seen in Figure 8. Recent drops in coverage in part relate to data rescue efforts taking part with a significant lag: real-time updates to HadAT will rectify this (H. Coleman et al., manuscript in preparation, 2005). However, there have also been significant drops in

Table 4. Summary of Metadata Events Associated With Breakpoints Adjusted in the HadAT1 Station Set[a]

Metadata Event Type Associated With Breakpoint	Number of Breakpoints	Percentage of Total
No known event	2088	69.95
Radiosonde model change	522	17.49
Humidity sensor change	116	3.89
Computational/calculation method	115	3.85
Ground equipment replacement	53	1.78
Radiation corrections applied changed	38	1.27
Cutoffs for data changed	23	0.77
Cord length change (Japan only)	16	0.54
Observations time change	6	0.20
Wind speed measurements	5	0.17
Wind measurements	1	0.03
Duct change	1	0.03
Station operator change	1	0.03

[a]Only static metadata events (known timing) were considered. A consideration of suspected metadata events (unknown or highly uncertain timing) would have associated more events with breakpoints, but with reduced confidence.

Table 3. Summary Statistics From Our QC of HadAT1 Stations[a]

WMO Region	Number of Stations in HadAT1	Number of Breakpoints Identified by Time Period (Average Per Station)						Mean of All Absolute Adjustment Factors, K	Median of All Absolute Adjustment Factors, K	Standard Deviation of All Absolute Adjustment Factors, K
		1958–1967	1968–1977	1978–1987	1988–1997	1997–2002	Full Period			
Europe (01–19)	70	86 (1.2)	106 (1.5)	125 (1.8)	100 (1.4)	15 (0.2)	432 (6.2)	0.553	0.444	0.430
Russia (20–39)	142	164 (1.2)	249 (1.8)	269 (1.9)	204 (1.4)	22 (0.2)	908 (6.4)	0.555	0.468	0.368
Asia (40–49)	67	73 (1.1)	111 (1.7)	101 (1.5)	106 (1.6)	26 (0.4)	417 (6.2)	0.510	0.371	0.468
Africa	14	6 (0.4)	29 (2.1)	26 (1.9)	22 (1.6)	1 (0.1)	84 (6.0)	0.575	0.421	0.464
North America	121	162 (1.3)	199 (1.6)	191 (1.6)	176 (1.5)	25 (0.2)	753 (6.2)	0.396	0.319	0.390
South America	21	14 (0.7)	43 (2.0)	37 (1.8)	25 (1.2)	1 (0.0)	120 (5.7)	0.510	0.400	0.392
Pacific area	42	45 (1.1)	81 (1.9)	79 (1.9)	59 (1.4)	7 (0.2)	271 (6.5)	0.446	0.352	0.347

[a]Results are summarized by WMO reporting region.

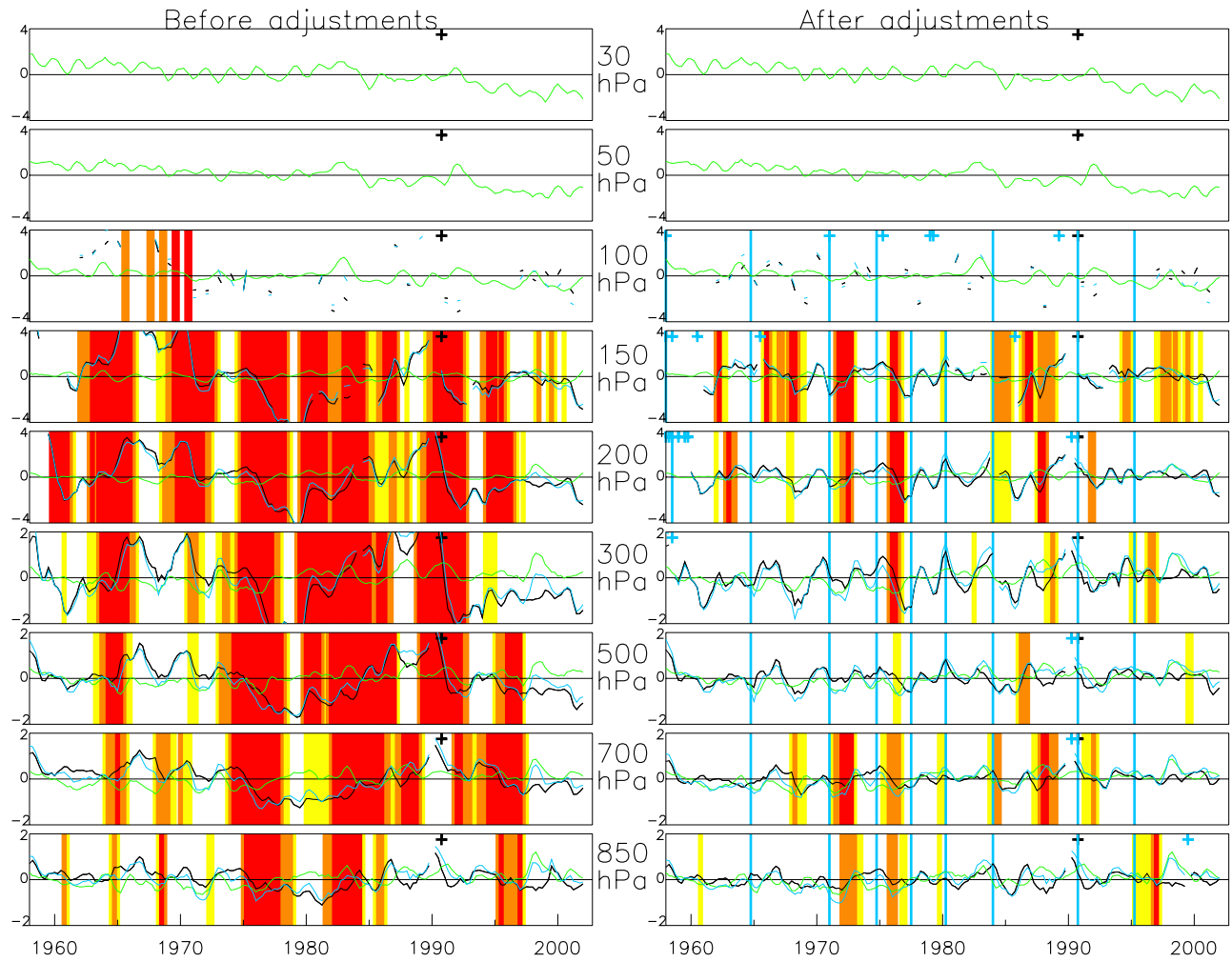

Figure 6. As Figure 2 but for station 43333 (Port Blair, India).

station sampling density, particularly in the former USSR, and partly as a result of a shift toward greater use of satellites for operational forecast input leading to a reduction in the radiosonde network.

6. Assigning Uncertainty Estimates

[46] The QC approach permits the assignment of uncertainty estimates. These uncertainty estimates are parametric (alternatively internal or value) uncertainty estimates rather than structural uncertainty estimates [*Thorne et al.*, 2005]. We cannot currently explicitly calculate the structural uncertainty that would result from a different station set choice, an additional/different break point identification approach or expert(s) assigning the breakpoints, adjusting stations in isolation, or the effect of any other methodological choices. To do this robustly would require many repetitions of the QC under different approaches. The structural uncertainty can begin to be quantified by comparing HadAT and its uncertainty estimates to the (very limited) number of alternative upper air temperature data sets which represent different choices in these respects. The uncertainty estimates also do not account for the subglobal coverage of the data set or the changes in this coverage with time.

6.1. Deriving Station Series Uncertainty Estimates

[47] The adjusted station series value for any point can be decomposed as follows:

$$A_{obs} = T_{obs} - (T_C + \varepsilon_C) + \varepsilon_{obs} + Adj + \varepsilon_{Adj} \quad (2)$$

where A_{obs} is the adjusted anomaly value, T_{obs} the observed value, T_C the true climatological mean value, ε_C the uncertainty in the climatology, ε_{obs} the uncertainty in the observation, Adj the total adjustment factor applied, and ε_{Adj} the uncertainty in this adjustment. It is assumed that the uncertainty terms are independent.

6.1.1. Uncertainty in the Calculated Climatology

[48] The climatology error estimate is restricted to the effects of incomplete temporal sampling over the climatology period. This will affect the absolute accuracy of the calculated climatology, adding a systematic bias to the anomaly time series. All HadAT1 seasonal resolution station level data which are temporally complete over the 1966–1995 period following the QC procedure were subsampled (1493 levels between the 477 stations). From these were randomly dropped out up to 50% of values. The two-tailed 90% confidence limits on the resulting climatologies increase linearly with the proportion of missing points. The

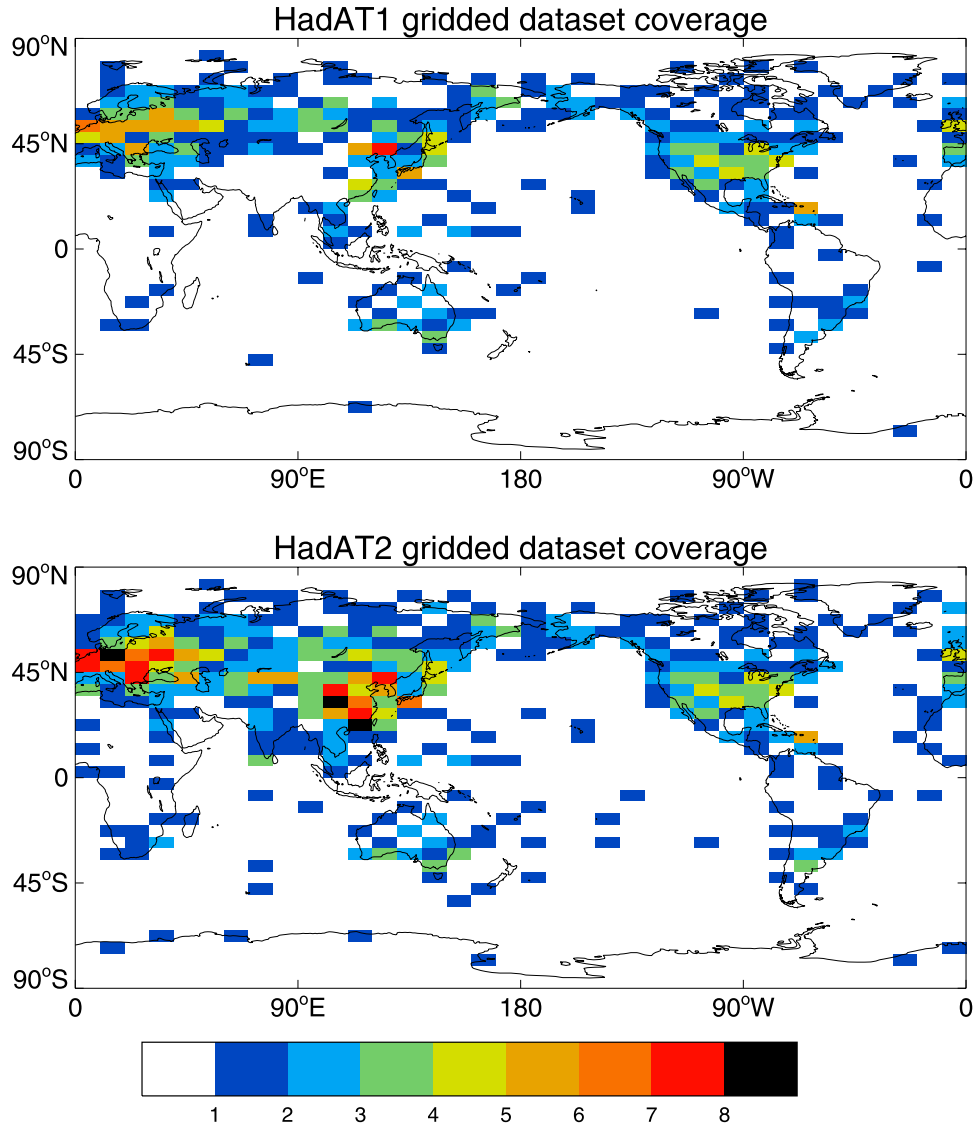

Figure 7. Grid box station coverage for (top) HadAT1 and (bottom) HadAT2 products. This is the maximum number of stations used. Actual data coverage varies over time and with height.

slope is almost entirely a function of the underlying time series interseasonal variability. So, values were scaled by the seasonal time series standard deviation, leading to a tight clustering of results. These were averaged together to form a single best estimate of the effect. From this best estimate the following relationship was ascertained that is applied to all station series:

$$\varepsilon_C = 0.64 \times \sigma \times f \quad (3)$$

where ε_C is the climatology uncertainty (K), σ is the seasonal time series standard deviation over the climatology period (K) and f is the fraction of missing data. The relationship accounts for over 98% of the variance in the best estimate.

6.1.2. Uncertainties in the Observed Values

[49] Uncertainties in individual ascents arise from, among other factors, individual instrumental biases, launch timing biases, sampling of a single time slice of the chaotic atmospheric system, and coding and transmission errors. Some of these will have systematic characteristics over short time periods. For example, a station might receive and launch a dubious batch of instruments for some time (a few weeks or months) before the problem is noticed and rectified. Such short-term effects will not have been detected or corrected by the QC other than through deletions of very obvious spikes in the seasonal difference time series values.

[50] Given the range of causes it is difficult to gain an unbiased a priori estimate of ε_{obs}. Many manufacturers provide absolute accuracy claims on their sondes. These could in theory be used, dividing by $\sqrt{n}$, where n is the number of ascents, to give uncertainties across timescales. However, for most stations in HadAT, this information is not available in full. Instrument accuracy is also not the sole source of observational uncertainty so any such estimate would be too small. Informed estimates of the remaining sources and their magnitude could be made. In reality these

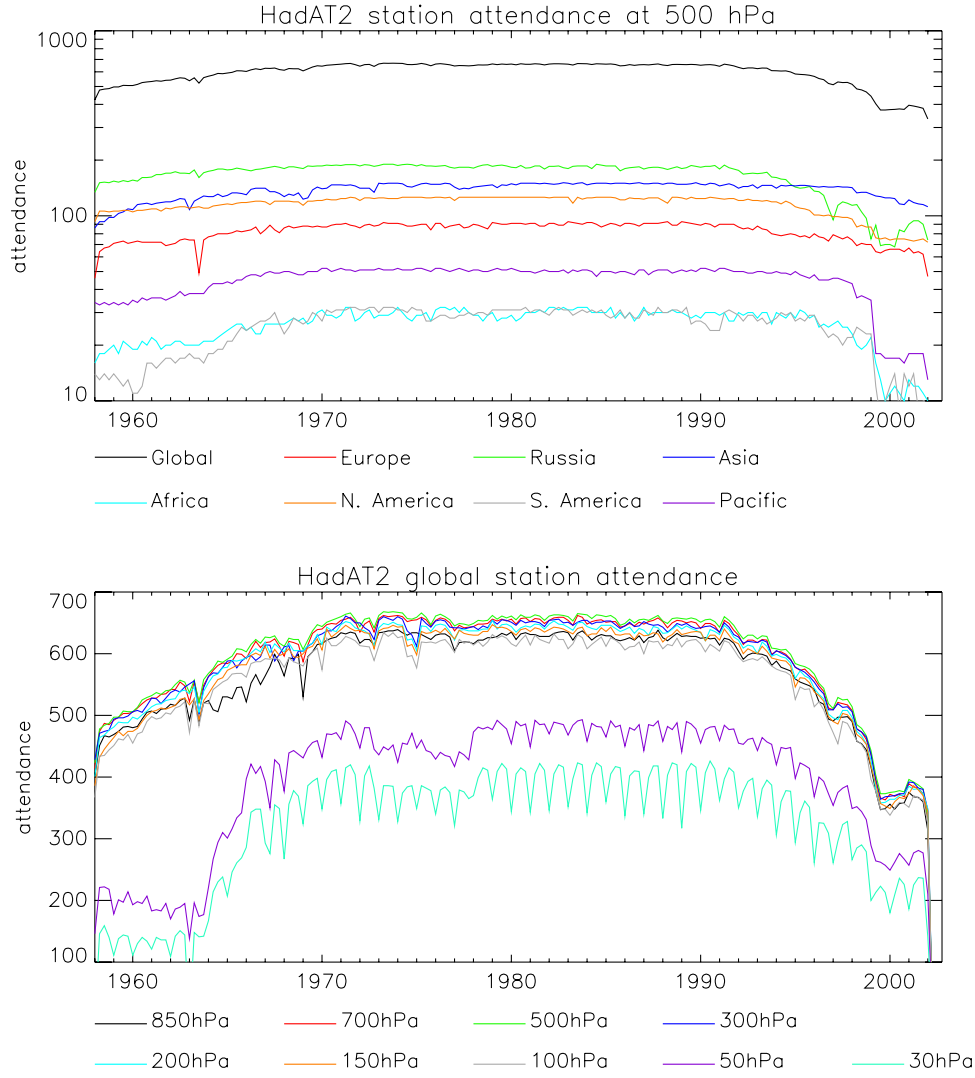

Figure 8. Seasonal station attendance in HadAT2. (top) By WMO region and globally at the 500 hPa level. Note the logarithmic attendance axis. (bottom) Globally by pressure level (linear axis).

will be station-specific resulting from protocols and practices which are both time varying and in most cases unknown. An average expectation would overestimate the sampling error at some stations and underestimate it at others and inevitably be subjective.

[51] Given the problems of gaining an unbiased estimate of ε_{obs} empirically is it possible to gain such an estimate from the available time series? For each station the QC method yields a station series, a neighbor series, and a difference series. The neighbor series clearly contains no information on the station sampling error. The station series contains both real trends and signals arising from natural climate variability in addition to the ε_{obs} term. Following the QC procedure the difference series should be indistinguishable from white noise. This series is used to estimate $\varepsilon_{\mathrm{obs}}$. It can be decomposed as follows (cf. (2)):

$$D_{obs} = \left(T_{obs} - (T_C + \varepsilon_C) + \varepsilon_{obs} + Adj + \varepsilon_{adj}\right) - (T_{neigh} - (T_{C_{neigh}} + \varepsilon_{C_{neigh}}) + \varepsilon_{obs_{neigh}} + Adj_{neigh} + \varepsilon_{Adj_{neigh}}) + Physical \quad (4)$$

where D_{obs} is the observed difference, *Physical* is a real physical discrepancy which would arise in the limit of all the uncertainty terms being zero (and will time average to zero), and all other symbols are as defined in equation (2) with subscript $_{\mathrm{neigh}}$ denoting a neighbor average. It is assumed that the errors for both the station and the neighbor series will be proportional to their overall variance. So the difference series D_{obs} is scaled to estimate ε_{obs}:

$$D_{obs}(scaled) = \left(\frac{D_{obs}}{\sqrt{1. + \mathrm{var}neigh/\mathrm{var}station}}\right) \quad (5)$$

1.64σ of $D_{obs}(scaled)$ yields the estimated 90 percent confidence limits on ε_{obs}.

6.1.3. Adjustment Uncertainty

[52] Following the adjustment factor calculation there are quantifiable estimates of ε_{Adj}. The 5th and 95th percentiles from the adjustment factor distribution are used. Uncertainty relative to the present-day increases back in time as more adjustments are included. The uncertainties are

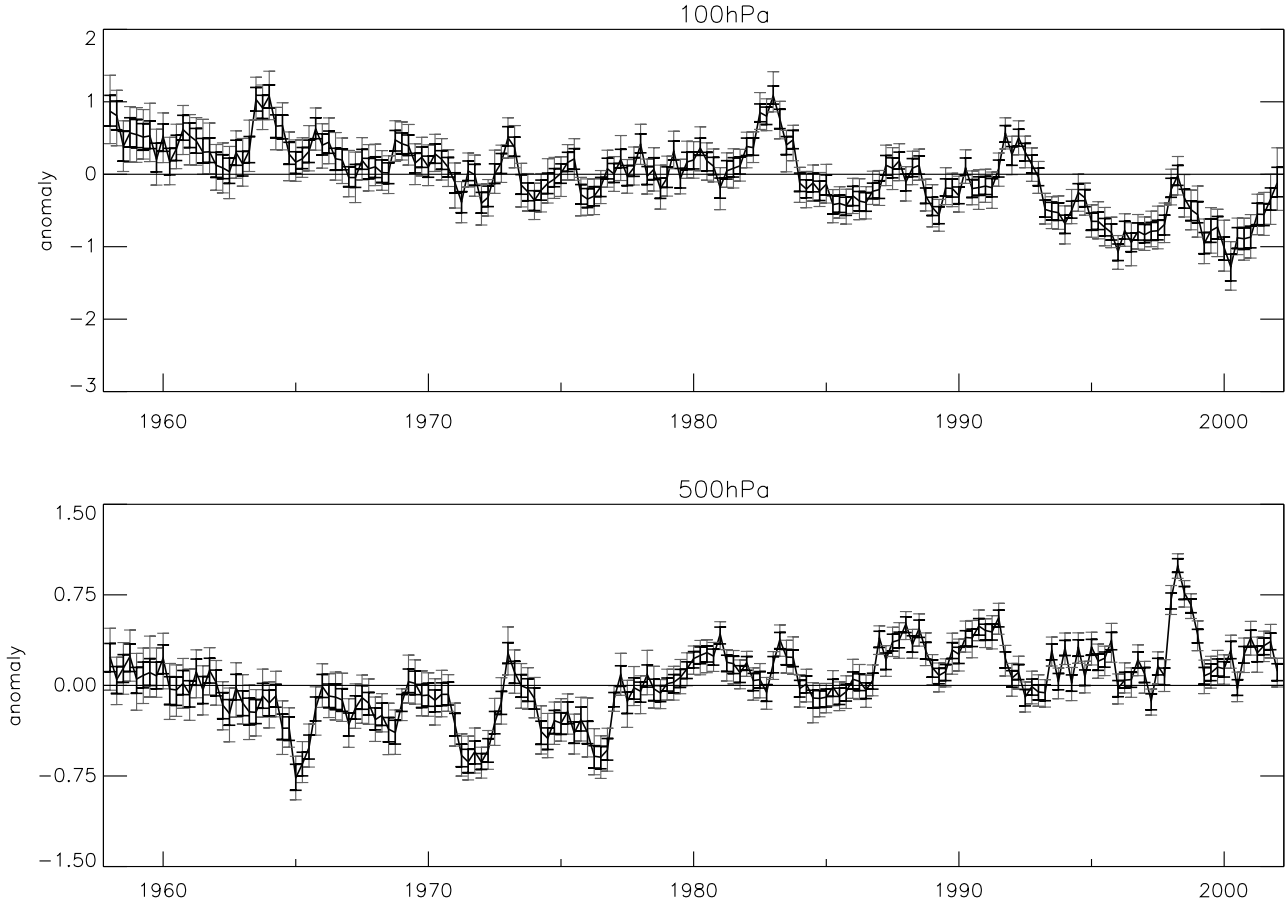

Figure 9. Global mean HadAT2 time series for 100 hPa and 500 hPa on a seasonal basis. Bars denote the absolute (faint) and 5th to 95th percentile (bold) ranges of solutions from our 1000 realizations. Global means have been attained through zonally averaging all the available gridded data and then taking a cos(lat) weighted average. This mitigates the effects of the unequal spatial sampling (Figures 1, 7, and 8), yielding a more truly representative global mean value, but comes at the cost of inflating uncertainty estimates by concentrating large-scale mean diagnostics upon poorly sampled regions.

assumed to be independent of one another within a station series so sum quadratically. In reality there is likely to be some interdependence, so adjustment uncertainty may be underestimated. The degree of interdependence will be specific to each individual adjustment. It will be a function not simply of the station series but also of the neighbor series, and the proximity and relative sign compared to other suspected breakpoints.

6.2. Quantifying Uncertainty in Large-Scale Mean Trends

[53] Uncertainty estimates in large-scale means could be gained directly from the station series and their uncertainty estimates. However, this would require estimates of intra–grid box station correlations and the number of effective degrees of freedom on a range of time and space scales [*Jones et al.*, 1997] and good estimates of variability in the very large areas not sampled in HadAT. To explicitly characterize the uncertainty across the full range of space and timescales instead a range of plausible realizations of the final data set were calculated.

[54] First 100 plausible realizations of each station time series were created. To incorporate sampling uncertainties, ε_{obs}, 100 bootstrap estimates of the scaled difference series (equations (4) and (5)) were added on to the station series. In a few cases (less than 0.1% of all station series values) the station series contained a value but the difference series did not (no neighbor values). In these cases the scaled difference series standard deviation was used to scale a random normal distribution. Values from this were then added on to the original station series at these points.

[55] It could be argued that the actual difference series should first be removed from the station series, before adding on a randomized version and applying a further ad hoc correction of $\sqrt{2}$ to account for the additional variance this two-step process incorporates. However, sometimes, particularly around volcanic and strong ENSO events, HadAT2 fell outside this range of uncertainty estimates. Clearly there is a component of the difference series which is truly physical in origin (section 6.1, equation (5) and discussion). This may relate to these periods being atypical of longer-term variability and hence not well resolved by the correlation-based neighbor series construction approach (section 3). Discrepancies arose almost entirely in the tropics, and were greatest at height, where there may remain unresolved problems both with the radiosonde data and with

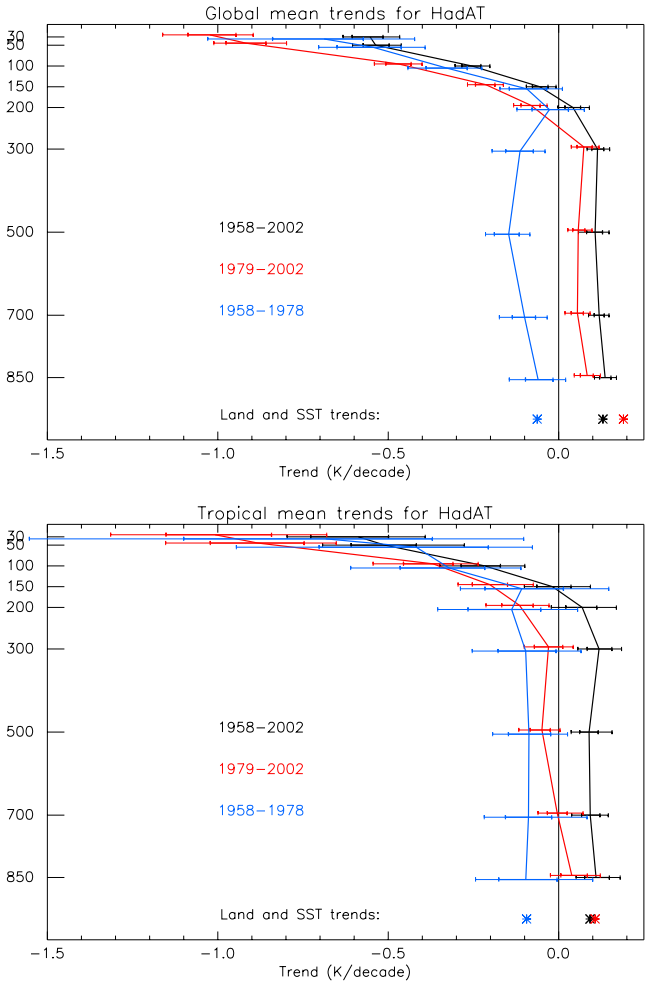

Figure 10. Global and tropical (20N to 20S) mean linear temperature trends (K/decade) derived using Median of Pairwise Slopes fit [*Lanzante*, 1996] for HadAT2 for the full period and pre-MSU and MSU record eras. Bold error bars denote 5th and 95th percentiles, and faint error bars the absolute maximum and minimum from our distribution. Stars denote surface trends from HadCRUT2v [*Jones and Moberg*, 2003] subsampled to 500 hPa HadAT2 radiosonde availability. There are no available uncertainty estimates on the surface data, but they will have some uncertainty associated with them. Uncertainty ranges aloft reflect observational uncertainty alone following section 6.2 and not the goodness-of-fit of the linear trend to the underlying data. The analysis was repeated using a simple ordinary least squares (OLS) estimator to assess sensitivity (not shown). For the satellite era, OLS produced systematically slightly increased tropospheric trends because of the outlier effect of the strong warming associated with the 1997/1998 ENSO, but the impact is within MPS trend uncertainty bounds. Otherwise the two approaches led to essentially indistinguishable results.

the NCEP reanalyses upon which the neighbor coefficients were based [*Simmons et al.*, 2004].

[56] For each adjustment applied, for each station realization, an additional small adjustment factor was calculated to reflect the uncertainty, ε_{Adj}. A large normal distribution scaled based upon the 5th and 95th adjustment percentiles (assumed to be $\pm 1.64\sigma$) was created. Values from this distribution were randomly sampled and added as extra adjustment factors to the synthetic station series.

[57] It was decided not to renormalize the synthetic station series over the climatology period as the uncertainty in the records relative to present really does increase back in time. After renormalizing, the records would be artificially similar over the climatology period of 1966–1995 and increasingly divergent in other periods. Regardless, such an approach does not capture the effects of missing data which were parameterized in section 6.1.

[58] The resulting synthetic station series distribution was compared to that predicted from the station error estimates derived in section 6.1. A count of synthetic values outside of the predicted 90% ranges was performed for all stations. The synthetic series were in good agreement with these estimates, ranging from 5% to 20% with most cases in the range 8% to 12%. Therefore we proceeded to create 1,000 realizations of the HadAT2 product. For each realization one version from the population of 100 time series for each station was randomly picked. These were then combined and gridded.

7. A Brief Initial Analysis of HadAT2

[59] Within the lower stratosphere, at 100 hPa, the three major volcanic eruptions, Agung (1963), El Chichón (1982), and Pinatubo (1991) produce obvious warming spikes with a duration of the order 18 months in the global series (Figure 9, top). The warming spikes are greater at 50 hPa and 30 hPa and reduced at 150 hPa and 200 hPa. There are other interseasonal to interannual variations. These variations are largest in the tropics and mainly reflect changes relating to the Quasi-Biennial Oscillation (QBO). Over the entire period of 1958 to 2002 there is an overall global cooling at 100 hPa. As found quantitatively by *Seidel and Lanzante* [2004] for other data sets, this evolution could as well be described qualitatively by a series of stepwise coolings following volcanic eruptions, as by a linear trend.

[60] The El Niño–Southern Oscillation (ENSO) event of 1997/98 is the most prominent feature in the global series at 500 hPa (Figure 9, bottom). The global mean warming associated with this ENSO event increases with height from ~0.75K at 850 hPa to ~1.25K at 300 hPa. In the tropics the effects reach higher – up to at least 150 hPa. In addition there are other interseasonal to interannual timescale variations, some of which correlate with ENSO and volcanic events. The warmth in the late 1950s/early 1960s is primarily a Northern Hemisphere effect. There is some evidence for a systematic shift to a warmer regime in the mid to late 1970s [*Trenberth*, 1990], but this is complicated by the elevated interseasonal to interannual variability from the mid-1960s until this shift. In the tropics, the evidence for this shift is more pronounced.

[61] Over 1958–2002, the global and tropical troposphere warmed at all levels at rates indistinguishable from those observed at the surface (Figure 10). However, linear trend fits for both pre-MSU and MSU subperiods are more negative aloft than the whole period trend The majority of the net tropospheric change within HadAT is a quasi-step change in the late 1970s (Figure 9), close to the break

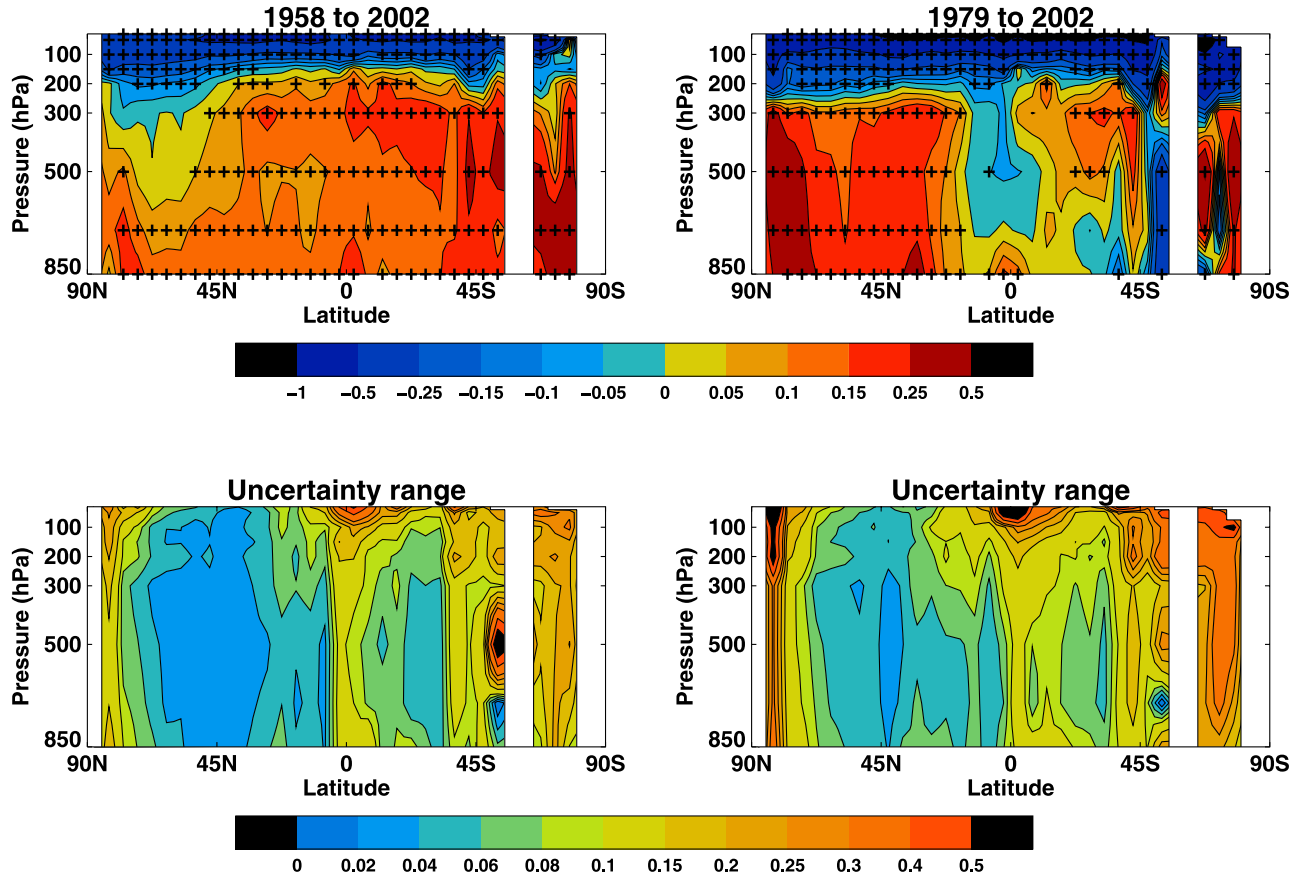

Figure 11. (top) Zonal mean ordinary least squares linear trends for HadAT2 for the full and satellite periods and (bottom) their associated uncertainties in K/decade. In the trend panels significantly nonzero values are denoted by a cross. The uncertainty estimates take no account of either missing data regions or the goodness-of-fit. They solely show the uncertainty as described in section 6.2.

between the two subperiods. By contrast, surface time series are more linear in nature (not shown), though the surface record exhibits less warming than 1958–2002 in the pre-MSU period and more warming than 1958–2002 during the MSU period. The apparent agreement between surface and HadAT tropospheric temperature evolution over 1958–2002 should be interpreted with great caution. The fact that trends agree does not imply a common time series evolution and may be entirely fortuitous.

[62] Therefore tropospheric temperature evolution, at least in HadAT2, has been too nonlinear to justify the indiscriminate use of a linear trend. Even the sign of the tropospheric temperature linear trend fit can change given a sufficiently careful *a posteriori* choice of start and end dates. Linear trends should only be used cautiously to describe the climate system evolution, and alternative models [e.g., *Seidel and Lanzante*, 2004] and/or metrics should be strongly considered to avoid ambiguity in interpretation. Only with this caveat in mind are zonal mean trends used below to ascertain the geographic origins of the large-scale average trends.

[63] Zonal mean trends over the full period 1958–2002 exhibit warming throughout the troposphere, and cooling in the stratosphere (Figure 11). There is a relative minimum in the warming in the Northern Hemisphere around 60°N. Both tropospheric and stratospheric trends are significant across most of the globe. The observed trends are most certain from approximately 70°N to 30°N and around 30°S, where sampling density is best (Figure 7). Over the satellite period the pattern of trends within the troposphere is more complex. There is strong warming north of 30°N, a cooling in the tropics and slight warming in the Southern Hemisphere midlatitudes. Outside of the strong Northern Hemisphere warming the trends over the satellite period are not generally significantly nonzero. Within the stratosphere there is a strong and significant cooling at all latitudes.

[64] During both the full and satellite periods data south of 45°S yield very heterogeneous zonal trend structures. Given the sparsity of the network here, zonal mean trends are likely to be unreliable and may reflect the inadequacies of a neighbor based homogenization approach in data-sparse high-latitude regions where correlation distances are small. We caution against an overinterpretation of these data.

8. Discussion

[65] HadAT is a new radiosonde temperature data set drawing upon all available digital data sources. It builds on recent intense efforts to homogenize a subset of the global network series (LKS, GUAN [*McCarthy*, 2000]). This homogenized subset was employed as a skeletal

reference network to enable the selection of a larger set of grossly consistent station records. This larger network was then used to construct neighbor-based estimates which were used to identify breakpoints and perform adjustments. Breakpoint adjustments were developed iteratively because adjustments impacted the neighbor series. LKS and GUAN stations were allowed to be adjusted during this procedure. If global or large-region systematic biases pervade the raw data then these will have only been reduced rather than removed, but the choice of LKS and GUAN to define the network minimizes the chances of this.

[66] Uncertainty estimates placed upon the resulting station and gridded time series account for adjustment uncertainty and observational sampling effects, but not for "structural uncertainty" arising from the choice of techniques. Work in progress at the Hadley Centre and elsewhere aims to gain a more quantitative estimate of this uncertainty through creating an ensemble of radiosonde Climate Data Records.

[67] Initial analyses of the resulting data set do not fundamentally alter our understanding of late 20th Century free atmospheric temperature changes. Namely:

[68] 1. Linear trend analysis shows that between 1958 and 2002 the troposphere warmed at a similar rate to the surface both globally and in the tropics, consistent with climate model predictions.

[69] 2. This linear trend agreement is misleading. Almost all of the tropospheric warming is the result of a step-like change in the mid to late 1970s which has been ascribed to a "regime shift", particularly in the tropics.

[70] 3. For the satellite era, evidence for tropospheric warming is weak away from the Northern Hemisphere midlatitudes, and the data do not preclude an absolute cooling in the tropics.

[71] 4. From the 1958 to 2002 the lower stratosphere cooled, punctuated by volcanic warming events.

[72] The gridded data sets, station time series, and a full audit trail are available at http://www.hadobs.org/ for bona fide research purposes. HadAT2 will be made available in near real time as a monthly anomaly temperature product (H. Coleman et al., manuscript in preparation, 2005).

[73] **Acknowledgments.** We thank two anonymous reviewers for their in-depth comments which greatly improved this paper. Met Office authors were funded by the Department for the Environment Food and Rural Affairs and the Government Met Research program. Through their contribution this paper is Crown Copyright. Much of this work was undertaken while P.W.T. was acting as a consultant to the Met Office through the Climatic Research Unit at the University of East Anglia. Phil Jones was supported by the Office of Science (BER), U.S. Department of Energy, grant DE-FG02-98ER62601.

References

Angell, J. K. (2003), Effect of exclusion of anomalous tropical stations on temperature trends from a 63-station radiosonde network, and comparison with other analyses, *J. Clim.*, *16*, 2288–2295.

Bengtsson, L., S. Hagemann, and K. I. Hodges (2004), Can climate trends be calculated from reanalysis data?, *J. Geophys. Res.*, *109*, D11111, doi:10.1029/2004JD004536.

Brown, S. J., D. E. Parker, C. K. Folland, and I. Macadam (2000), Decadal variability in the lower-tropospheric lapse rate, *Geophys. Res. Lett.*, *27*, 997–1000.

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

Christy, J. R., R. W. Spencer, W. B. Norris, W. D. Braswell, and D. E. Parker (2003), Error estimates of version 5.0 of MSU/AMSU bulk atmospheric temperatures, *J. Atmos. Oceanic Technol.*, *20*, 613–629.

Conrad, V., and L. D. Pollak (1962), *Methods in Climatology*, 459 pp., Harvard Univ. Press, Cambridge, Mass.

Durre, I., T. Reale, D. Carlson, J. R. Christy, M. Uddstrom, M. Gelman, and P. W. Thorne (2005a), Report from the Workshop to Improve the Usefulness of Operational Radiosonde Data, Asheville, North Carolina, 11th–13th March 2003, *Bull. Am. Meteorol. Soc.*, *86*, 411–416.

Durre, I., R. S. Vose, and D. B. Wuertz (2005b), Overview of the integrated global radiosonde archive, *J. Clim.*, in press.

Elliott, W. P., D. J. Gaffen, J. D. W. Kahl, and J. K. Angell (1994), The effect of moisture on layer thicknesses used to monitor global temperatures, *J. Clim.*, *7*, 304–308.

Eskridge, R. E., O. A. Alduchov, I. V. Chernykh, Z. Panmao, A. C. Polansky, and S. R. Doty (1995), A comprehensive aerological reference data set (CARDS)—Rough and systematic errors, *Bull. Am. Meteorol. Soc.*, *76*, 1759–1775.

Free, M., and D. J. Seidel (2005), Causes of differing temperature trends in radiosonde upper air data sets, *J. Geophys. Res.*, *110*, D07101, doi:10.1029/2004JD005481.

Fu, Q., C. M. Johanson, S. G. Warren, and D. J. Seidel (2004), Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends, *Nature*, *429*, 55–58.

Gaffen, D. J. (1996), A digitized metadata set of global upper-air station histories, *NOAA Tech. Memo. ERL ARL-211*.

Gaffen, D. J., B. D. Santer, J. S. Boyle, J. R. Christy, N. E. Graham, and R. J. Ross (2000), Multidecadal changes in the vertical temperature structure of the tropical troposphere, *Science*, *287*, 1242–1245.

Grody, N. C., K. Y. Vinnikov, M. D. Goldberg, J. T. Sullivan, and J. D. Tarpley (2004), Calibration of multisatellite observations for climatic studies: Microwave Sounding Unit (MSU), *J. Geophys. Res.*, *109*, D24104, doi:10.1029/2004JD005079.

Haimberger, L. (2005), Homogenization of radiosonde temperature time-series using ERA-40 analysis feedback information, *ERA-40 Rep. Ser. 22*, Eur. Cent. for Med.-Range Weather Forecasts, Reading, U. K.

Hegerl, G. C., and J. M. Wallace (2002), Influence of patterns of climate variability on the difference between satellite and surface temperature trends, *J. Clim.*, *15*, 2412–2428.

Jones, P. D., and A. Moberg (2003), Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001, *J. Clim.*, *16*, 206–223.

Jones, P. D., T. J. Osborn, and K. R. Briffa (1997), Estimating sampling errors in large-scale temperature averages, *J. Clim.*, *10*, 2548–2568.

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

Lanzante, J. R. (1996), Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data, *Int. J. Clim.*, *16*, 1197–1226.

Lanzante, J. R., S. A. Klein, and D. J. Siedel (2003a), Temporal homogenization of monthly radiosonde temperature data. part I: Methodology, *J. Clim.*, *16*, 224–240.

Lanzante, J. R., S. A. Klein, and D. J. Siedel (2003b), Temporal homogenization of monthly radiosonde temperature data. part II: Trends, sensitivities, and MSU comparison, *J. Clim.*, *16*, 241–262.

McCarthy, M. (2000), Improved global upper-air data from the Hadley Centre, paper presented at Royal Meteorological Society Conference, Cambridge, U. K. (Available at http://www.metoffice.com/research/hadleycentre/pubs/posters/McCarthy/index.html)

Mears, C. A., M. C. Schabel, and F. J. Wentz (2003), A reanalysis of the MSU channel 2 tropospheric temperature record, *J. Clim.*, *16*, 3650–3664.

National Research Council (2000), *Reconciling Observations of Global Temperature Change*, 85 pp., Natl Acad. Press, Washington, D. C.

Parker, D. E., and D. I. Cox (1995), Towards a consistent global climatological rawinsonde database, *Int. J. Clim.*, *15*, 473–496.

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

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed., pp. 617–622, Cambridge Univ. Press, New York.

Santer, B. D., et al. (2003), Influence of satellite data uncertainties on the detection of externally forced climate change, *Science*, *300*, 1280–1284.

Seidel, D. J., and J. R. Lanzante (2004), An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes, *J. Geophys. Res.*, *109*, D14108, doi:10.1029/2003JD004414.

Seidel, D. J., et al. (2004), Uncertainty in signals of large-scale climate variations in radiosonde and satellite upper-air temperature datasets, *J. Clim.*, *17*, 2225–2240.

Simmons, A. J., et al. (2004), Comparison of trends and variability in CRU, ERA-40 and NCEP/NCAR analyses of monthly-mean surface air temperature, *J. Geophys. Res.*, *109*, D24115, doi:10.1029/2004JD005306.

Sterin, A. M. (1999), An analysis of linear trends in the free atmosphere temperature series for 1958–1997 (in Russian), *Meteorol. Gidrol.*, *5*, 52–68.

Sterl, A. (2004), On the (in)homogeneity of reanalysis products, *J. Clim.*, *17*, 3866–3873.

Tett, S. F. B., and P. W. Thorne (2004), Comment on tropospheric temperature series from satellites, *Nature*, *432*, doi:10.1083/nature03208 7017.

Thorne, P. W., P. D. Jones, T. J. Osborn, T. D. Davies, S. F. B. Tett, D. E. Parker, P. A. Stott, G. S. Jones, and M. R. Allen (2002), Assessing the robustness of zonal mean climate change detection, *Geophys. Res. Lett.*, *29*(19), 1920, doi:10.1029/2002GL015717.

Thorne, P. W., P. D. Jones, S. F. B. Tett, M. R. Allen, D. E. Parker, P. A. Stott, G. S. Jones, T. J. Osborn, and T. D. Davies (2003), Probable causes of late 20th century tropospheric temperature trends, *Clim. Dyn.*, *21*, 573–591.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears (2005), Uncertainties in climate trends: Lessons from upper-air temperature records, *Bull. Am. Meteorol. Soc.*, in press.

Trenberth, K. E. (1990), Recent observed interdecadal climate changes in the Northern Hemisphere, *Bull. Am. Meteorol. Soc.*, *71*, 988–993.

Uppala, S., et al. (2005), The ERA-40 reanalysis, *Q. J. R. Meteorol. Soc.*, in press.

Wallis, T. W. R. (1998), A subset of core stations from the Comprehensive Aerological Reference Dataset (CARDS), *J. Clim.*, *11*, 272–282.

P. Brohan, H. Coleman, M. McCarthy, D. E. Parker, and P. W. Thorne, Hadley Centre for Climate Prediction and Research, Met Office, Exeter EX1 3PB, UK. (peter.thorne@metoffice.gov.uk)

P. D. Jones, Climatic Research Unit, University of East Anglia, Norwich, NR4 7TJ, UK.

S. F. B. Tett, Hadley Centre, Reading Unit, Reading University, Reading, RG6 6AH, UK.