INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **28**: 1269–1281 (2008)
Published online 14 November 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1633

RMetS
ROYAL METEOROLOGICAL SOCIETY

Radiosonde temperature trends and their uncertainties over eastern China†

Yanjun Guo,[a,d,e]* Peter. W. Thorne,[b] Mark P. McCarthy,[b] Holly A. Titchner,[b] Binxun Huang,[c] Panmao Zhai[a] and Yihui Ding[a]

[a] *National Climate Center, China Meteorological Administration (CMA), Beijing, China*
[b] *Met Office, Hadley Centre, Exeter, UK*
[c] *Chinese Academy of Meteorological Science, CMA, Beijing, China*
[d] *Laboratory for Climate Studies, CMA, Beijing, China*
[e] *Graduate School of the Chinese Academy of Sciences, Beijing, China*

ABSTRACT: Trends and uncertainty in radiosonde temperature records for six sample stations in eastern China are assessed. Results from a complex approach using metadata and a two-phase regression (M-TPR) to capture known and unknown metadata events respectively are compared with an ensemble of possible solutions generated by the Met Office automated homogenization system (QUARC). Independent satellite records from the Microwave Sounding Unit (MSU) record are used to validate breakpoints over the satellite era.

Differences in the treatment of metadata and the strictness of the statistical breakpoint detection methods used lead to relatively poor agreement in breakpoint identification. Agreement in long-term (1958–2003) trends in the homogenized data was found to result from a fortuitous cancellation of large differences in the pre- and post-satellite era trends between the two approaches.

A consideration of independent MSU satellite data lends some credence to the presence and calculated magnitude of many of the assigned breakpoints that were not associated with recorded metadata events, in the later part of the record. However, it also highlights that neither of the approaches is likely to be perfect at identifying breaks. Improved metadata are likely to prove vital in confirming the presence of these breaks and hence the veracity of the various homogenization approaches to data for eastern China. Copyright © 2007 Royal Meteorological Society

KEY WORDS radiosonde temperature; uncertainty; China

Received 13 December 2006; Revised 8 August 2007; Accepted 6 September 2007

1. Introduction

Radiosonde temperature time series contain valuable information for climate change research because they provide the longest record of upper air measurements. Homogenization of the data is required to account for the numerous artificial shifts in the data record that relate to changes in instrumentation and operating procedures (CCSP, 2006). Several homogenized radiosonde temperature datasets now exist (Angell, 2003), LKS/RATPAC (Lanzante *et al*., 2003a,b; Free *et al*., 2005), HadRT (Parker *et al*., 1997); HadAT (Thorne *et al*., 2005a), and RAOBCORE (Haimberger, 2007). However, the degree and sophistication of the homogenization differs greatly. Several studies have addressed the consistency, or lack thereof, among estimates of atmospheric temperature change derived from different radiosonde, re-analyses, and satellite datasets (Oort and Liu, 1993; Christy, 1995; Nicholls *et al*., 1996; Seidel *et al*., 2004; Free and Seidel, 2005; CCSP, 2006). Inconsistencies between radiosonde datasets are a result of different spatial and temporal station sampling, source data, adjustments for inhomogeneities, and other data processing choices (Free *et al*., 2002; Free and Seidel, 2005). It is therefore clear that structural uncertainty (Thorne *et al*., 2005b; CCSP, 2006) is important and that we should be undertaking many independent homogenization approaches and comparing their results to improve our understanding.

China covers a significant region of the globe and has a dense radiosonde network. Radiosondes have been launched twice daily within China since the 1950s. An assessment of Chinese radiosonde data was performed by Zhai and Eskridge (1996), who applied the two-phase regression (TPR) method (Easterling and Peterson, 1995) to two representative stations, located in eastern and western China and detected two breakpoints, both of which were associated with recorded metadata events. However, homogeneity has still not been given enough

* Correspondence to: Yanjun Guo, National Climate Center, China Meteorological Administration, China Zhong- Guan-Cun-Nan-Da-Jie Hai Dian, Beijing, China, 100081. E-mail: gyj@cma.gov.cn

† The contributions of Peter W. Thorne, Mark P. McCarthy and Holly A. Titchner of Met Office, Exeter, were prepared as part of their official duties as employees of the UK Government. It is published with the permission of the Controller of Her Majesty's Stationery Office and the Queen's Printer for Scotland.

attention in a number of climatic change analyses that still use unhomogenized time series (e.g. Wang and Ren, 2005).

We conducted the homogenization of radiosonde temperature time series on six stations located in eastern China (Figure 1) using metadata and the two-phase regression method denoted by M-TPR. These results were compared to adjustments applied to the same data in 100 experiments using QUARC, an automated homogenization scheme based upon the HadAT (Thorne *et al*., 2005a) methodology developed by the Met Office, and described in McCarthy *et al*. (2007). Microwave Sounding Unit (MSU) satellite products (Christy *et al*., 2003; Mears *et al*., 2003; Mears and Wentz, 2005) were used as an independent reference series to address differences in the two approaches since 1979. Trends and uncertainty in the radiosonde temperature data have been assessed by investigating the uncertainties in breakpoint identification and exploring the differing temperature trends obtained by applying different homogenization procedures to the time series.

2. Data and methods

2.1. Source data

The source data were from the unadjusted RAOBCORE (Haimberger, 2007) dataset, which combines the radiosonde ingest to the ERA-40 re-analysis system (Uppala *et al*., 2005) with the quality-controlled Integrated Global Radiosonde Archive (IGRA; Durre *et al*., 2006). During the merging preference is given to ERA-40 ingest data, Only those stations for which a 30-year climatology for 1966–1995 could be calculated using the approach detailed in Thorne *et al*. (2005a) were retained to create a global set of stations. Data were initially prepared as seasonal anomalies of twice-daily observations at 00 UTC and 1200 UTC, and were then combined to produce a merged (day and night) dataset for the final homogenization procedure. Merged series were assigned as missing if both the 00 UTC and the 12 UTC series were missing. We considered six Chinese stations (Figure 1) that represented good geographical distribution and had metadata records. Temperature data were available at 850, 700, 500, 400, 300, 200, and 100 hPa for these stations. Metadata used in M-TPR were from the China Meteorological Administration (CMA), and metadata used in QUARC were from IGRA (Gaffen, 1996 and subsequent updates). Differences in metadata exist between these two sources. In both sources, the metadata relates primarily to changes in the radiosonde model used and radiation correction methods, but other additional influences (e.g. change in ground equipment) may have occurred, which might not have been documented as thoroughly.

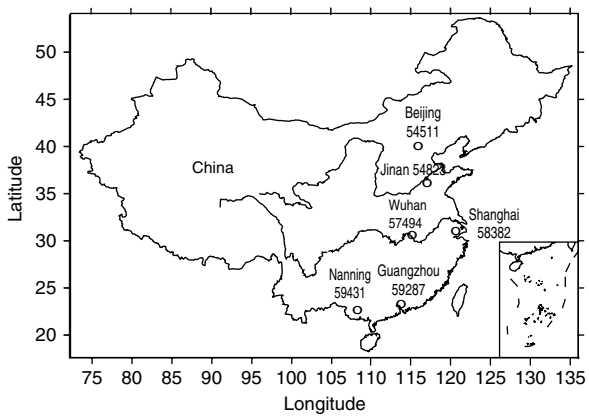

Figure 1. Location of the six stations considered in this study.

2.2. M-TPR methods

In M-TPR, we employ an approach consisting of using metadata and a two-phase regression for homogenization. The former is necessary to capture documented breakpoints and the latter to capture undocumented breakpoints. This homogenization approach was applied to the six Chinese stations only.

(1) Adjustment of breakpoints documented using metadata

Firstly, all CMA-recorded metadata events within the merged (day + night) radiosonde temperature time series were assigned as breakpoints. In order that any real climate trend was not removed, we detrended the series before and after each metadata event. The adjustment was then simply the difference in the mean detrended temperature anomaly before and after the breakpoint, such that data preceding the break were made consistent with those after it. For each station, this process was iterated back in time from the present to the earliest break. Adjustments were estimated and applied in this way for all levels for each breakpoint.

(2) Adjustment of undocumented breakpoints using the TPR method

Radiosonde station histories are known to be incomplete and, in many cases, inaccurate. Therefore adjustment of radiosonde temperature series must also consider the potential for undocumented breakpoints existing within the dataset. We have adopted the TPR method (Easterling and Peterson, 1995) to detect undocumented breakpoints in the (day + night) merged raw time series at each level. The TPR method (a brief description is given in Appendix A), has been widely applied in the homogenization of climate data (Zhai and Eskridge, 1996; Vincent, 1998; Lund and Reeves, 2002; Wang, 2003). Here, this procedure was applied after the correction for known metadata events had been carried out. The TPR tests the significance of a two-phase fit to each point in a series of differences from a reference series using (1) a likelihood ratio statistic, using the two residual sums of squares and (2) the difference in the means of the difference series before and after the potential discontinuity as evaluated using a t-test (assessed at the 5% C.I). Each of the discontinuities that had been thus identified was further tested using a multi-response permutation procedure. The adjustment is calculated as the difference in the means of a reference series for all data before and

DOI: 10.1002/joc

after the discontinuity, and was applied to all data points preceding the breakpoint.

TPR therefore requires the use of a suitable reference series, which should closely resemble the true climate evolution at the candidate site. Zhai and Eskridge (1996), Sherwood *et al*. (2005), and Randel and Wu (2006) have shown that night-time radiosonde data are potentially more suitable than daytime data for this purpose. We first corrected the night-time series only at CMA-recorded metadata events using the approach discussed above, and then used this series as a reference for the merged (day + night) series. If, as seems plausible, the night-time data contain breaks not associated with metadata events, then these will not be accounted for. Therefore we cannot be certain that the reference series is truly homogeneous. This is a problem encountered in all radiosonde (and other climate datasets) homogenization efforts that involve recourse to a poorly characterized reference series (McCarthy *et al*., 2007). Time series before and after applying M-TPR adjustments are shown in Figure 2 for an example station.

2.3. Hundred random experiments using QUARC

The QUARC system uses a neighbour-based iterative approach to detect and adjust breakpoints (Appendix B provides a brief overview; see McCarthy *et al*., 2007 for more detail). The system relies on a set of tunable parameters that control various aspects of the breakpoint identification and adjustment procedures. Therefore, by running the system with different sets of parameter choices we have the capability to generate multiple versions of radiosonde records that explore uncertainty relating to methodological choices (McCarthy *et al*., 2007). In McCarthy *et al*. (2007) the uncertainty in radiosonde records has been investigated for global and tropical regions.

To compare M-TPR and QUARC, we constructed 100 random QUARC experiments in a similar manner to those by McCarthy *et al*. (2007), the main difference being the input dataset and the ensemble size. Homogenization was performed on the full global set of stations (Section 2.1), although we only considered the results from the six Chinese stations within this study. Each experiment was based on different random settings for the 14 adjustable system parameters of the QUARC parameterization considered (McCarthy *et al*., 2007, see also www.hadobs.org QUARC page). These parameters (such as the number of iterations performed) affected the breakpoint identification and the adjustment calculations during the homogenization, akin to making different methodological decisions. For each experiment the parameters were randomly set to within reasonable bounds (Appendix A in McCarthy *et al*., 2007).

3. An inter-comparison of QUARC and M-TPR results

3.1. Uncertainties in breakpoint identification

Breakpoints identified by M-TPR and QUARC are shown in Figure 3. It is clear that the QUARC ensemble consistently identifies certain breakpoints. The QUARC break

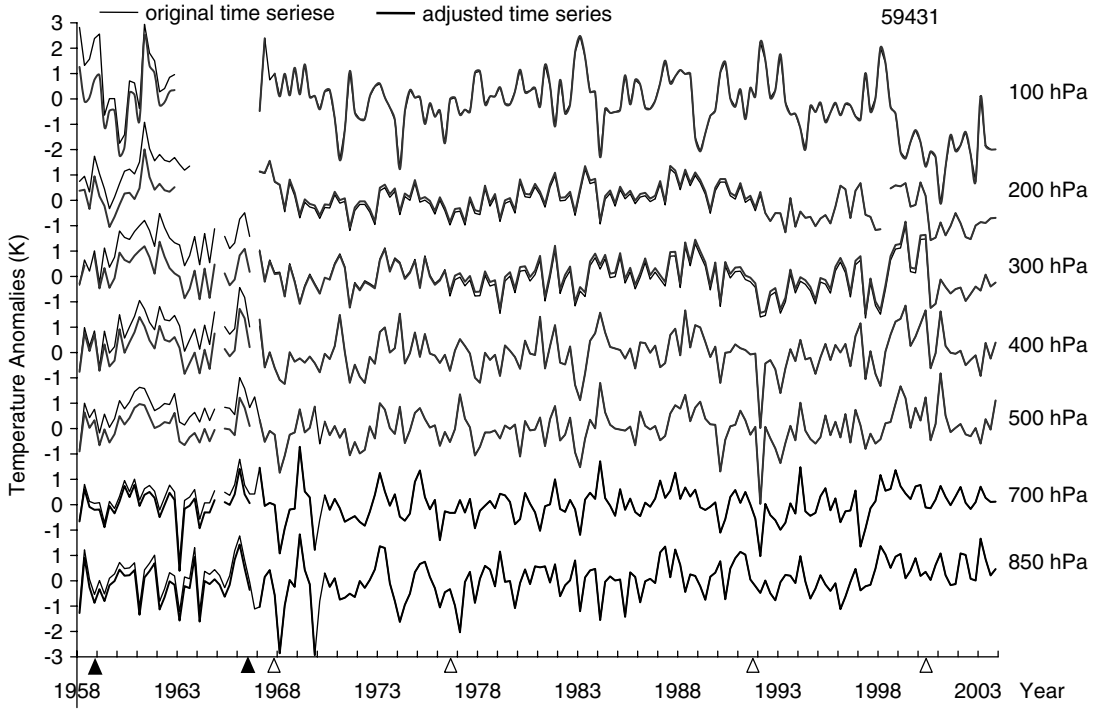

Figure 2. Seasonal temperature anomalies at station 59431 from 850 to 100 hPa. The original merged series (thin solid line), and the M-TPR-adjusted merged series (thick solid line) are shown. Breakpoints identified where metadata events (1958 and 1966 for all levels) exist are denoted by solid triangles and breakpoints detected using the TPR method (200 hPa during 1967–1991, 300 hPa during 1976–2000, and 400 hPa during 1967–1991) are denoted by the open triangles.

DOI: 10.1002/joc

years located close together, for example in the early 1960s for station 57494, may reflect uncertainty of the breakpoint timing in the homogenization and does not necessarily suggest the existence of multiple breakpoints at this time. A summary of the consistency of the two methods is given in Table I. The best agreement was for station 59431 (Nanning), with all M-TPR breakpoints identified within 1 year by QUARC. The worst agreement was for station 58362 (Shanghai) where QUARC did not detect breakpoints in the 1960s at this station and identified a number of additional breakpoints later in the series. Most Chinese stations changed radiosonde model from RZ 049 to GZZ-2 during the 1960s resulting in step-like declines in temperature, at least for those stations considered by Lanzante *et al.* (2003a). If a coincident break occurs at neighbouring stations we might expect QUARC to be inefficient at detecting this change, but at this time we cannot discount the possibility that at least some of the documented metadata events had no discernible impact on the temperature series. Conversely, from 1970 to 2003 several breakpoints were consistently detected by QUARC and not M-TPR. Over this period there are relatively few recorded metadata events.

3.2. Uncertainties in adjustments

Uncertainty in the estimated adjustments can be of order several K between QUARC experiments and M-TPR. An example for a breakpoint in June 1960 is shown in Figure 4. Such an uncertainty in adjustment estimates has a significant impact on the uncertainty in long-term trend estimates. Trends are on the order of a few tenths of a Kelvin per decade and therefore a small number of systematic biases close to this magnitude could significantly shift the trend estimate away from its true value.

Taking the results from the M-TPR as the standard for comparison, the differences in adjustments at seven atmospheric levels between QUARC and the M-TPR were examined using the standard deviations of the differences between the QUARC adjusted and M-TPR-adjusted data (Figure 5). Stations 54511 (Beijing) and 59287 (Guangzhou) showed the best agreement between the two adjustment methods, with maximum deviations < 1 K. The differences increased with height, as found in previous analyses (e.g. CCSP, 2006).

3.3. Uncertainties in station trends

The temperature trend is a key metric in communicating climate change. Unfortunately, the long-term trend is precisely the metric upon which adjustment uncertainties will project most strongly as each adjustment uncertainty adds (unintentional) red-noise into the series. The vertical profile of temperature trends was averaged over the six stations (Figure 6). Trends were derived for the raw data, the adjusted M-TPR data, and the 100 QUARC experiments for the periods 1958–2003, 1958–1979, and 1979–2003. For the full period (1958–2003), the trend profile from M-TPR matched the median QUARC trend profile well except at 100 hPa. Both M-TPR and QUARC

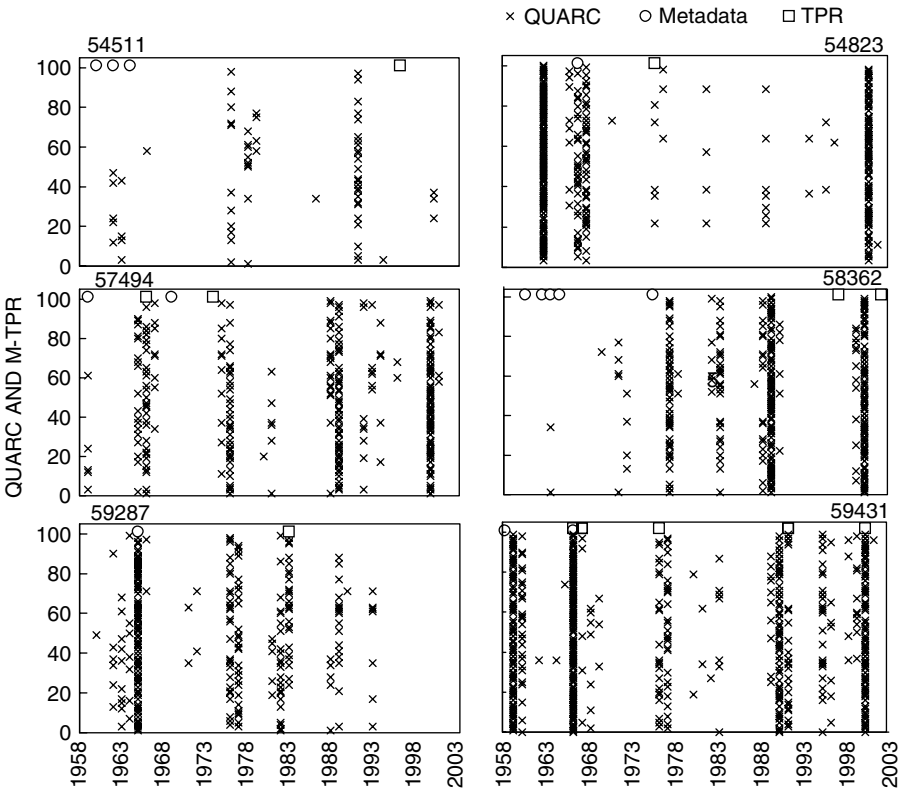

Figure 3. Temporal distribution of breakpoints detected by QUARC (crosses, one for each experiment) and by M-TPR using metadata (circles) and the TPR method (squares) at the six stations considered.

DOI: 10.1002/joc

Table I. Comparison of the timing of breakpoints detected using M-TPR and using QUARC.

Breakpoint		Station (WMO code)					
		54511	54823	57494	58362	59287	59431
Consistency	M-TPR breaks identified by at least one QUARC experiment within 1 year	1962	1966	1959	None	1965	1958
		1964	1975	1967	–	1983	1966
		–	–	1974	–	–	1968
		–	–	–	–	–	1976
		–	–	–	–	–	1991
		–	–	–	–	–	2000
	M-TPR breaks identified by at least one QUARC experiment within 2 years	1960	None	1969	1975	None	None
		–	–	–	1998	–	–
		–	–	–	2002	–	–
Inconsistency	QUARC break not within 2 years of an M-TPR break (year/number of QUARC experiments).	1976/11	1962/92	1981/6	1971/5	1976/28	1983/7
		1978/9	1981/4	1988/18	1972/4	1977/27	1995/25
		1979/4	1988/7	1989/49	1982/6	1989/17	1996/7
		1991/25	2000/73	1992/8	1983/28	1993/7	–
		2000/3	–	1993/6	1988/19	–	–
		–	–	1994/5	1989/73	–	–
		–	–	2000/61	1990/6	–	–
		–	–	2001/4	–	–	–
	M-TPR break not within 2 years of a QUARC break.	1996	None	None	1960	None	None
		–	–	–	1962	–	–
	–	–	–	–	1963	–	–
	–	–	–	–	1964	–	–

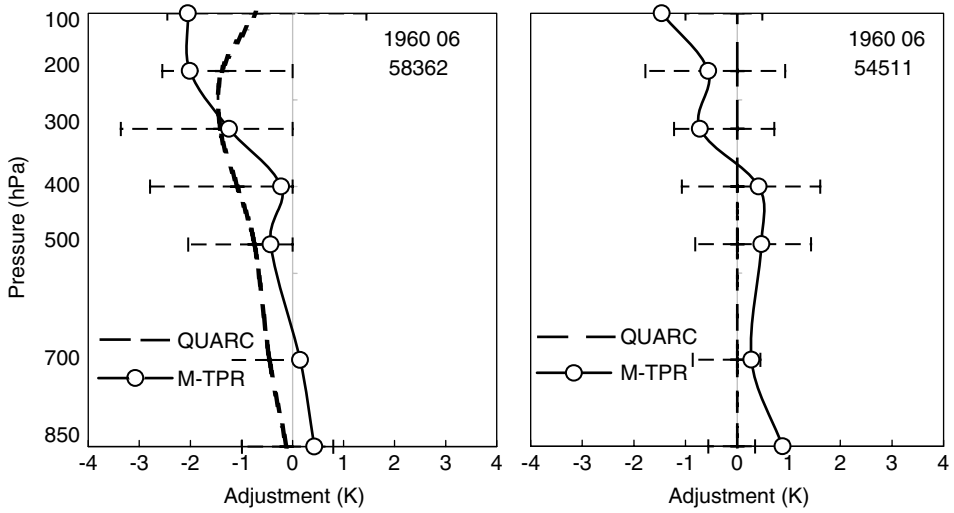

Figure 4. Breakpoint profiles for stations 58 362 and 54 511 for June 1960. Adjustments for M-TPR data (solid with circles) and QUARC experiments (dashed line is the median of 100 experiments; error bar gives the maximum and minimum) are shown.

During the pre-satellite era (1958–1978), the raw data imply cooling of the entire atmospheric column. QUARC weakens this cooling trend at all levels. The profile from M-TPR indicates small warming trends below 500 hPa. Trend profiles from the two approaches agreed well only within the upper troposphere (400–200 hPa). The trend from M-TPR was out of range of the QUARC results at 850, 700, and 100 hPa. The QUARC profile is the closer to the raw data. Despite the methods showing reasonable agreement with respect to the existence of breaks over this period (Table I), the resultant adjusted trends differ substantially. This relates to the different adjustment methodologies. Finding a break represents only part of the challenge – making an adequate adjustment is also required. Over the satellite era (1979–2003), the M-TPR derived trend almost overlaps that of the original data because few metadata events are recorded after the 1970s, and the M-TPR approach applies very few breaks in the absence of metadata (Table I). Therefore, M-TPR makes essentially no changes to the raw data over this latter period. Conversely, the QUARC data suggest greater warming than the raw data throughout the atmospheric column. Although tropospheric warming was observed using both M-TPR and QUARC, the close agreement between these two approaches over the full period (1958–2003) is a fortuitous cancellation of very substantial differences in reconstructed trends prior to and after 1979.

Differences in long-term trends, therefore, relate to both identification of breakpoints and their adjustment. To illustrate this Figure 7 gives time series of adjustments from the two approaches for station 54511 (Beijing) from 850 to 100 hPa during 1958–2003. M-TPR made a downward adjustment early in the record corresponding to an identified metadata event, which resulted in a relative warming trend. In contrast QUARC found most breakpoints in the latter part of the record. Moreover, the adjustment from QUARC increases with altitude, whereas the applied M-TPR adjustments are more constant with height. These differences lead to significant differences in trends during 1958–1978 between M-TPR and QUARC at lower levels (Figure 6).

4. Can independent datasets shed light on causes of differences between the approaches?

Breakpoints that occurred after 1979 and were detected by either the M-TPR procedure or by at least 25 QUARC experiments (largely QUARC) were examined further by comparing MSU-equivalent layer temperatures to collocated MSU data from UAH v 5.2 (Christy *et al*., 2003; Christy and Norris, 2006) and RSS v.2 (Mears *et al*., 2003; Mears and Wentz, 2005). Radiosonde data were converted to equivalent weighted MSU soundings using static weighting functions provided by the University of Alabama Huntville (see http://hadobs.metoffice.com/hadat/msu_equivalents.html for details). For QUARC breaks, the distribution of MSU-equivalent adjustments were compared with those expected from geographically coincident MSU time series for periods of 2–5 years about each breakpoint. This is equivalent to using MSU as the reference series against which to adjust the radiosonde records; a method previously employed by Parker *et al*. (1997). Agreement between the various estimates would imply that the target station exhibited a similar step change when compared with neighbouring radiosonde stations and with co-located MSU data, thereby supporting the QUARC results. If the MSU-derived adjustments and QUARC were in disagreement, this would imply that QUARC was likely placing spurious breaks into the Chinese radiosonde data, which would be highly undesirable or that the MSU datasets also contained spurious breaks.

Two example station series are shown for channel 4 (stratosphere) and 2LT (lower troposphere) (Figure 8). The difference series between the station and its neighbours and the station and collocated MSUs exhibit similar high frequency behaviour. However, the running

DOI: 10.1002/joc

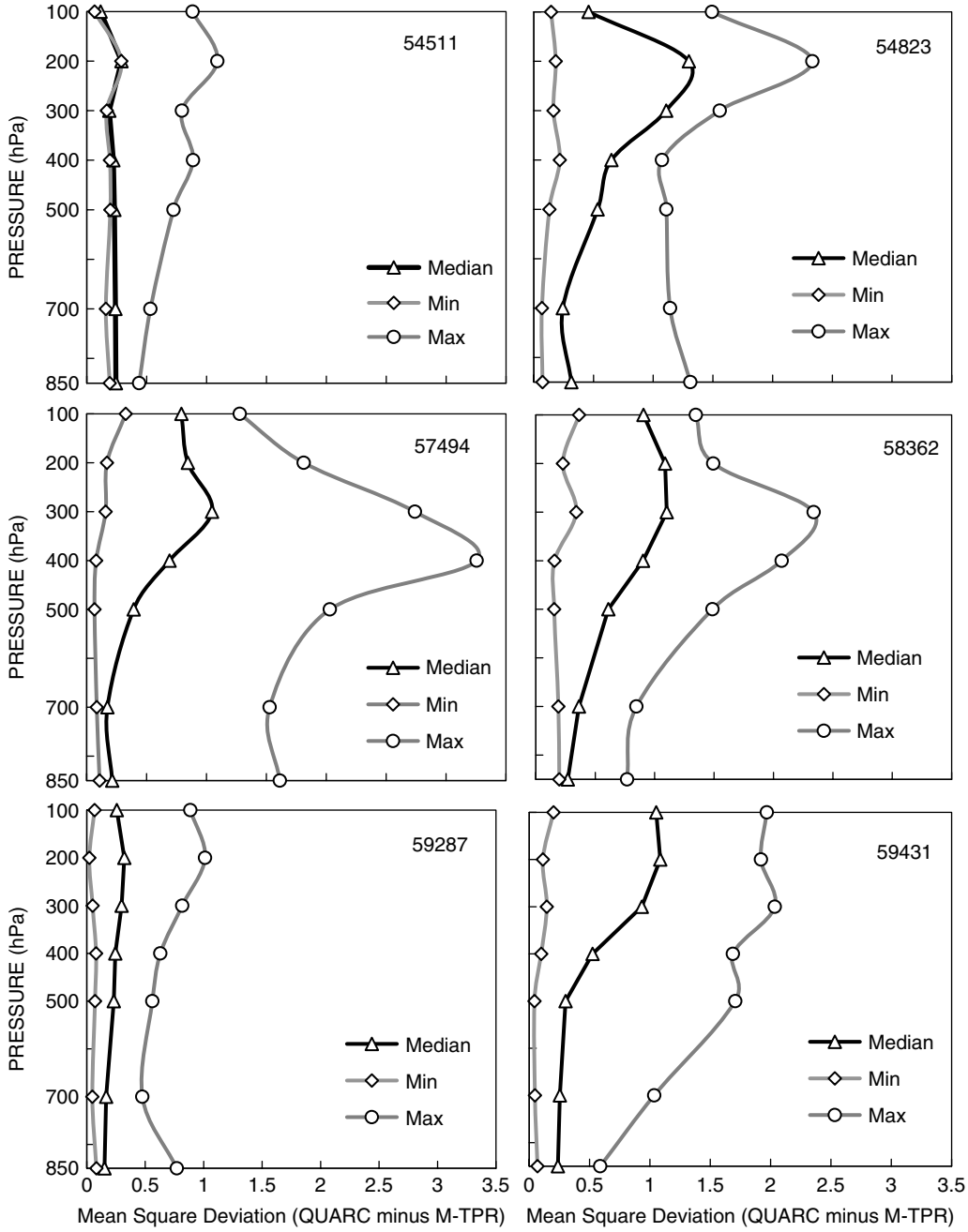

Figure 5. Vertical profiles of mean square deviation between results from M-TPR and 100 QUARC experiments.

mean adjustment estimates clearly highlight differences between the two different MSU series (e.g. for 54511 in 2LT in the early 1990s), as well as with the neighbour series (e.g. for 54511 in 2LT in the mid to late 1990s). The estimate of the required adjustment also varies depending upon whether stations from the same country or with similar metadata timing are allowed to contribute to the neighbours or not, particularly in the stratosphere.

For station 54 511 neither system has identified breakpoints particularly well if the MSU series represent the truth. M-TPR finds a break at approximately the time of maximum implied tropospheric adjustment from the station minus neighbour series, but only at one tropospheric level (850 hPa). A worse phenomenon is that the MSU series imply that the adjustment, if required, should be of the opposite sign. QUARC finds the smaller break around 1993 in the station minus neighbour tropospheric series, but here the two MSU series disagree as to the magnitude of the implied break (RSS implies no break at all); so confidence is low in the reality of this break. The large stratospheric break around 2000, implicit in both MSU series and the difference series when country/metadata are set, is missed by both systems. So, results for this station are far from encouraging, but this is found to be the worst example.

Station 58 362 shows much better agreement between MSU and neighbour-based adjustment series. Furthermore, QUARC has assigned breaks at approximately the

DOI: 10.1002/joc

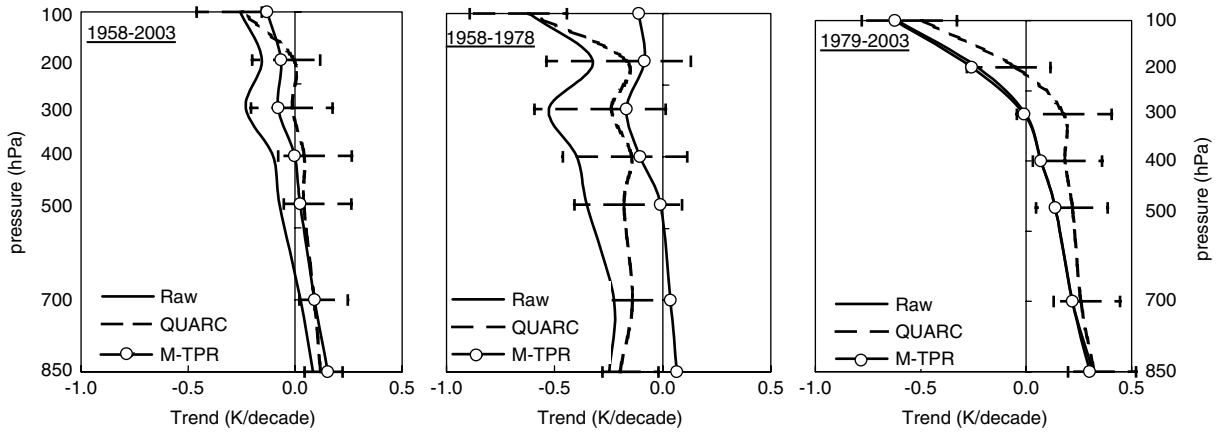

Figure 6. Vertical profiles of temperature trends (K/decade) as a function of pressure computed for the raw data (solid), the adjusted M-TPR data (solid with circles), and the QUARC ensemble of experiments (dashed line with horizontal bars denoting the range).

maxima and minima of these adjustment series and hence the range of breaks agrees well between the QUARC adjustments and those estimated from MSUs as a result.

Results for all QUARC-identified breaks across the full set of stations exhibit reasonable agreement with regard to break sign and magnitude both in the troposphere and in the stratosphere (Figure 9). For most, but not all, estimates overlap substantially such that they are likely to be consistent. So, where the QUARC system identifies breaks it makes sensible adjustments if MSU series can be considered an adequate reference truth. However, this does not imply that QUARC necessarily finds the right breaks, as is evident for station 54 511 (Figure 8) and in some other cases (not shown). Interestingly, results are slightly more consistent when using RSS as the MSU reference series than when using UAH, contrary to published inter-comparisons of raw sonde data with the two satellite products in the tropics (Christy *et al*., 2007). However, the sample is insufficient to make meaningful inferences.

5. Discussion and conclusions

An in-depth study of the two homogenization approaches: the M-TPR procedure developed at CMA; and an automated neighbour-based procedure QUARC developed at the Met Office Hadley Centre (McCarthy *et al*., 2007) was conducted for six radiosonde temperature time series in eastern China. An initial encouraging agreement between the approaches in terms of long-term trends was found to arise from a fortuitous cancellation of substantial differences in pre- and post-1979 trends. These differences arose from a combination of differences in both breakpoint identification and adjustment calculation approaches, as has been found across a broader inter-comparison of previous datasets by Free *et al*. (2002). However, both our analysis and previous efforts have been unable to determine an optimal approach as we lack a definitive 'ground truth' against which to make such an assessment.

QUARC and other neighbour-based approaches may have limitations in areas such as China where changes tend to be contemporaneous across large regions. The QUARC approach recognizes this by allowing neighbours from the same country or with similar metadata to be removed from a neighbour composite as one of a number of processing choices. Equally, trying to identify and adjust for breakpoints in station data as done in M-TPR is likely to be more difficult as the time series are inherently noisier. The differences in breakpoint identification relate both to the treatment of metadata and to the approach to testing for breakpoints. M-TPR assigns breaks at all levels at every metadata event and finds a relatively few breaks at individual levels elsewhere by the TPR method. Because there are more metadata events early in the record it made more significant adjustments in the pre-satellite era.

Metadata are often incomplete, missing, or sometimes actually erroneous (Peterson *et al*., 1998). There is substantial evidence that, globally, metadata for radiosondes are grossly inadequate with many apparent breakpoints not associated with metadata (Thorne *et al*., 2005a; Haimberger, 2007). The QUARC results here suggest this may also be the case for these stations. The metadata used in M-TPR relate to all changes in radiosonde model and radiation corrections and were collected officially by CMA and checked by consulting station operators. However, we are still not able to definitively conclude that all metadata events that may lead to a discontinuity have been recorded. Some other additional influences (e.g. change in ground equipment) may have occurred.

Fortunately, we have other datasets, derived independently (and from different platforms) with which to compare. Over the MSU era both UAH and RSS provide support for the sign and significance (non-zero value) of most QUARC-identified breaks. These independent estimates therefore provide quantitative support for the presence of breaks not associated with current metadata and not picked up in M-TPR. There is large uncertainty in the adjustment estimates from the two MSU series and the neighbour-based radiosonde composites. This

DOI: 10.1002/joc

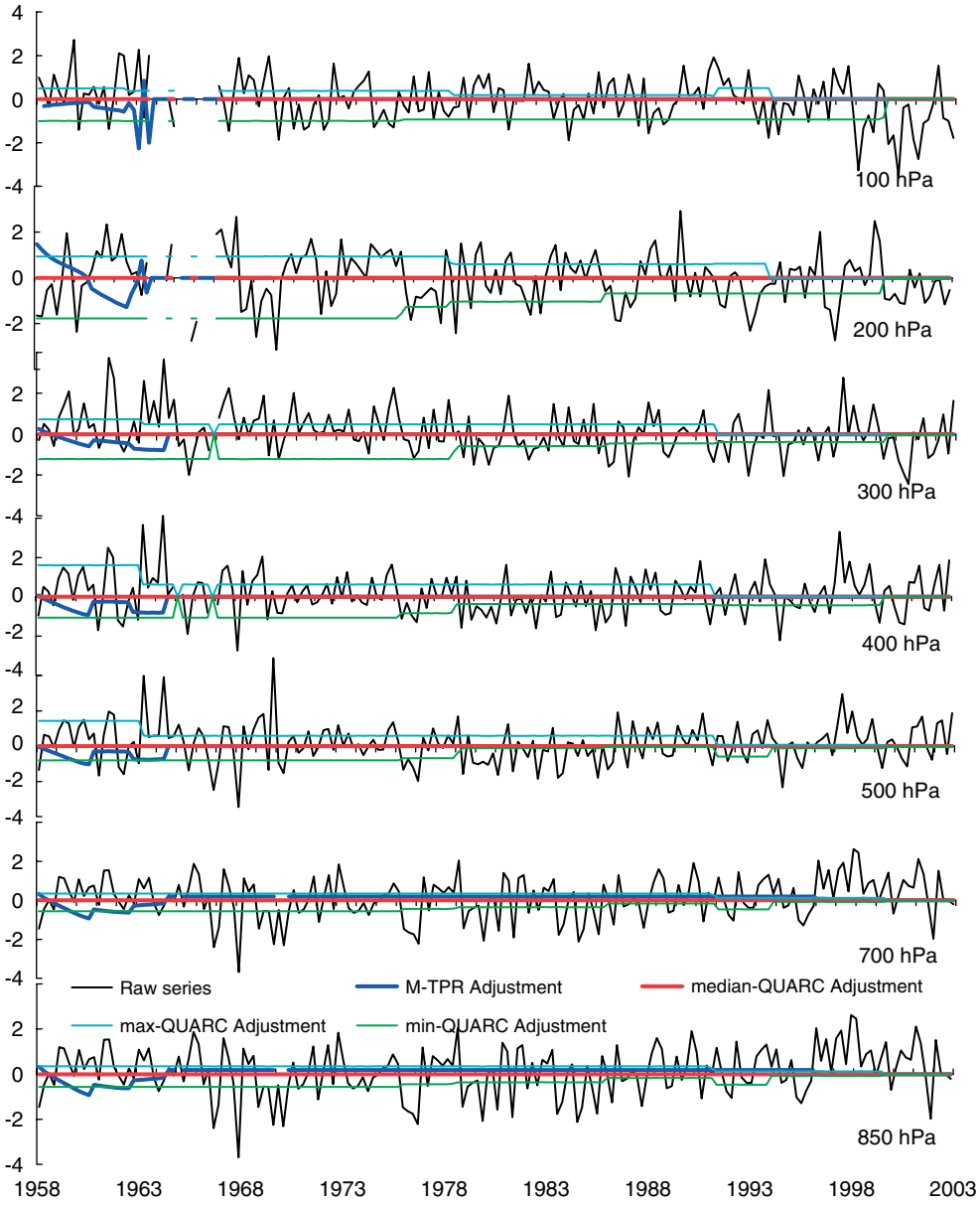

Figure 7. Time series of adjustments from QUARC and M-TPR overlaid on the raw seasonal temperature merged series (black line) at station 54511 from 850 to 100 hPa during 1958–2003. Among them are: the M-TPR-adjusted adjustment (blue solid line), median of QUARC adjustment (red line), maximum of QUARC adjustment (grey line), and minimum of QUARC adjustment (green line).

affects the likely number of breaks, their timing, and their magnitude, particularly in the stratosphere. Our analysis indicates that some breakpoints may have been missed by both approaches, and that there can be significant uncertainty in the timing and magnitude of breakpoints.

So, there is large uncertainty in station time series for the six stations considered in this study. This uncertainty projects most strongly onto trend behaviour, but can be masked if a fortuitous cancellation of errors occurs. There is quantitative evidence that the M-TPR procedure is too conservative in assigning breaks outside of metadata events. However, it is not clear that QUARC is optimal at catching and adjusting for these either. The limiting factors in all this is a lack of both high-quality metadata and a necessary 'ground truth' comparison database as would be provided by, for example, GRUAN (WMO, 2007). For metadata, it is not just radiosonde changes but any changes in station practices, personnel, ground equipment, recording practices, and radiation corrections and so on. Until such data are collected both for these stations and more globally it will be impossible to unambiguously confirm the large number of breaks, typically 70% in global homogenization efforts (Thorne *et al*., 2005a; Haimberger, 2007), that are assigned in the absence of metadata.

Acknowledgements

This paper was based on the RAOBCORE and IGRA datasets, kindly provided by Dr Haimberger of the University of Vienna. We also thank Mr David Parker and

DOI: 10.1002/joc

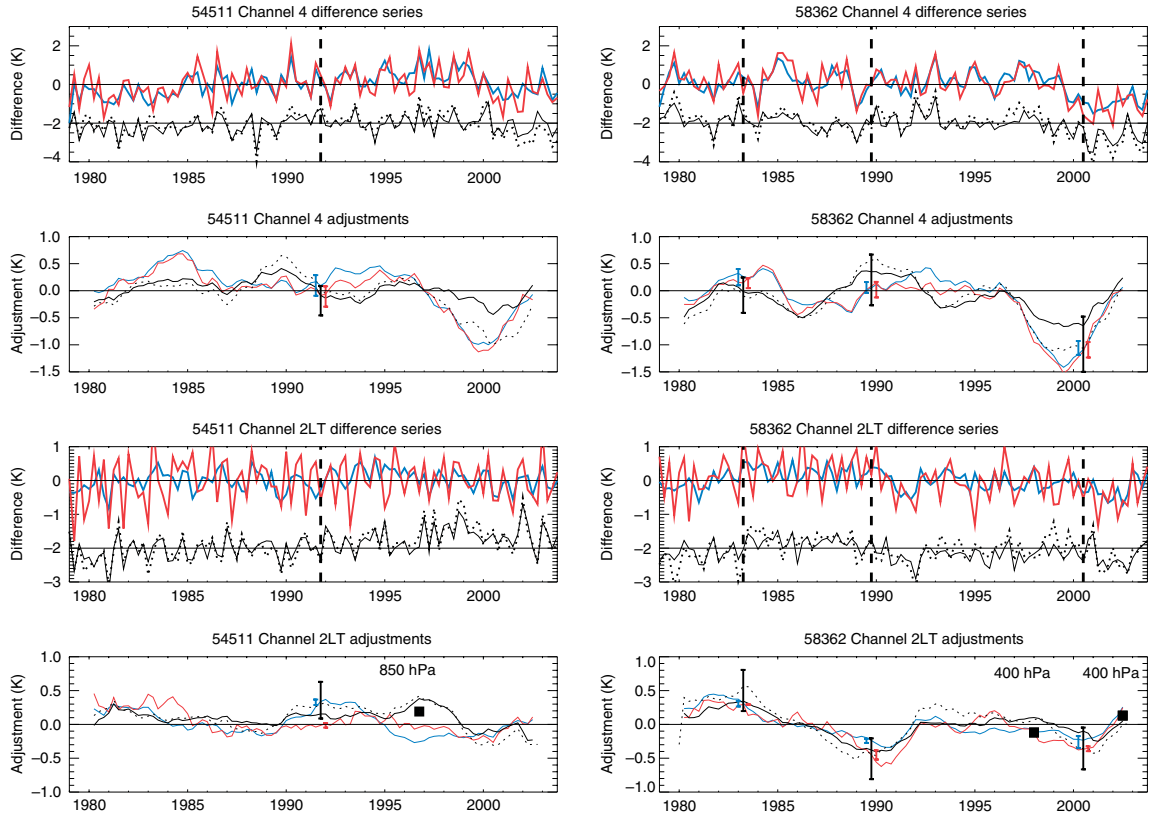

Figure 8. Channel 4 and 2LT series for stations 54511 and 58362. The top panels show the MSU minus station series (red RSS, blue UAH) and, vertically offset by −2 K, the neighbour-based difference series for the raw data (solid: all neighbours, dashed: subset excluding same country and similar metadata stations). QUARC breaks identified by 25 experiments or more are shown by dashed vertical lines. The lower panels show a running average implied adjustment effect from each series based upon an average of 2-, 3-, 4-, and 5-year estimates. At the points of assigned breaks the different adjustments from the QUARC experiments (black vertical bars) and the assigned ranges estimated from the two MSU series are shown. For each T-MPR break the adjustment is shown as a square along with the level at which it was applied.

Dr Thomas C. Peterson for helpful comments and suggestions. We are particularly grateful to Dr Peterson who kindly provided his Fortran code for the TPR method. This work was supported in part by the Climate Change Special Fund of China Meteorological Administration (CCSF2007-7). Met Office authors are funded by the UK Department of the Environment, Food, and Rural Affairs under contract PECD 7/12/37. Through the contribution of the Met Office authors, this paper is a British Crown Copyright.

References

Angell JK. 2003. Effect of exclusion of anomalous tropical stations on temperature trends from a 63-station radiosonde network, and comparison with other analyses. *Journal of Climate* **16**: 2288–2295.

CCSP. 2006. Temperature trends in the lower atmosphere: steps for understanding and reconciling differences. *A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research*, Karl TR, Hassol SJ, Miller CD, Murray WL (eds). National Oceanic and Atmospheric Administration, National Climatic Data Center: Asheville.

Christy JR. 1995. Temperature above the surface layer. *Climatic Change* **31**: 455–474.

Christy JR, Norris WB. 2006. Satellite and VIZ-radiosonde intercomparisons for diagnosis of nonclimatic influences. *Journal of Atmospheric and Oceanic Technology* **23**: 1181–1194.

Christy JR, Norris WB, Braswell WD, Parker DE. 2003. Error estimates of version 5.0 of MSU/AMSU bulk atmospheric temperatures. *Journal of Atmospheric and Oceanic Technology* **20**: 613–629.

Christy JR, Norris WB, Spencer RW, Hnilo JJ. 2007. Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements. *Journal of Geophysical Research* **112**: D06102. DOI:10.1029/2005JD006881.

Durre I, Vose RS, Wuertz DB. 2006. Overview of the integrated global radiosonde archive. *Journal of Climate* **19**: 53–68.

Easterling DR, Peterson TC. 1995. A new method for detecting undocumented discontinuities in climatological time series. *International Journal of Climatology* **15**: 369–377.

Free M, Seidel DJ. 2005. Causes of differing temperature trends in radiosonde upper air data sets. *Journal of Geophysical Research* **110**: D07101, DOI:10.1029/2004JD005481.

Free M, Seidel DJ, Angell JK, Lanzante J, Durre I, Peterson TC. 2005. Radiosonde atmospheric temperature products for assessing climate (RATPAC): A new data set of large-area anomaly time series. *Journal of Geophysical Research* **110**: D22101. DOI:10.1029/2005JD006169.

Free M, Durre I, Aguilar E, Seidel D, Peterson TC, Eskridge RE, Luers JK, Parker D, Gordon M, Lanzante J, Klein S, Christy J, Schroeder S, Soden B, McMillin LM, Weatherhead E. 2002. Creating climate reference datasets-CARDS workshop on adjusting radiosonde temperature data for climate monitoring. *Bulletin of the American Meteorological Society* **83**: 891–899.

Gaffen DJ. 1996. *A Digitized Metadata Set of Global Upper-air Station Histories*. NOAA Technical Memorandum ERL ARL-211.

Haimberger L. 2007. Homogenization of radiosonde temperature time series using innovation statistics. *Journal of Climate* **20**: 1377–1403.

Lanzante JR, Klein SA, Seidel DJ. 2003a. Temporal homogenization of monthly radiosonde temperature data. Part I: methodology. *Journal of Climate* **16**: 224–240.

Lanzante JR, Klein SA, Seidel DJ. 2003b. Temporal homogenization of monthly radiosonde temperature data. Part II: trends, sensitivities, and MSU comparison. *Journal of Climate* **16**: 241–262.

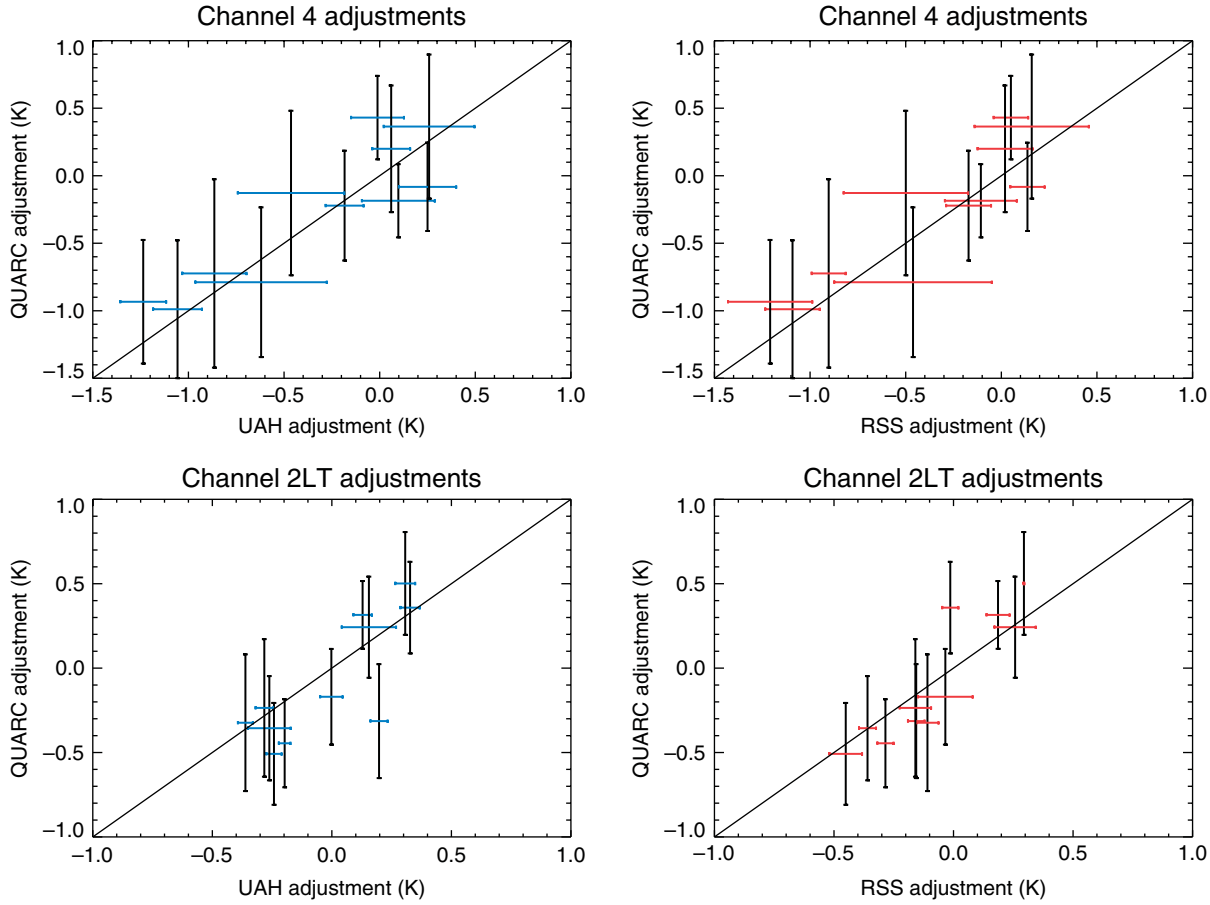

Figure 9. Comparison of MSU-derived estimates to those from QUARC at all QUARC-assigned breaks across the stations over the satellite era. The breaks should encompass the 1 : 1 diagonal if the QUARC procedure is adequate. This analysis cannot account for whether the breaks are correctly assigned (see Figure 8 and main text).

Lund R, Reeves J. 2002. Detection of undocumented changepoints: a revision of the two-phase regression model. *Journal of Climate* **15**: 2547–2554.

McCarthy MP, Titchner HA, Thorne PW, Tett SFB, Haimberger L, Parker DE. 2007. Assessing bias and uncertainty in the HadAT adjusted radiosonde climate record. *Journal of Climate* (in press).

Mears CA, Wentz FW. 2005. The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science* **309**: 1548–1551.

Mears CA, Schabel MC, Wentz FW. 2003. A reanalysis of the MSU channel 2 tropospheric temperature record. *Journal of Climate* **16**: 3650–3664.

Mielke PW. 1991. The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth-Science Reviews* **6**(31): 55–71.

Nicholls N, Tapp R, Burrows K, Richards D. 1996. Historical thermometer exposures in Australia. *International Journal of Climatology* **16**: 705–710.

Oort AH, Liu H. 1993. Upper-air temperature trends over the globe, 1958–1989. *Journal of Climate* **6**: 292–307.

Parker DE, Gordon M, Cullum DPN, Sexton DMH, Folland CK, Rayner N. 1997. A new gridded radiosonde temperature data base and recent temperature trends. *Geophysical Research Letters* **24**: 1499–1502.

Peterson TC, Easterling DR, Karl TR, Groisman P, Nicholls N, Plummer N, Torok S, Auer I, Boehm R, Gullett D, Vincent L, Heino R, Tuomenvirta H, Mestre O, Szentimrey T, Salinger J, Forland E, Hanssen-Bauer I, Alexandersson H, Jones P, Parker D. 1998. Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology* **18**: 1493–1517.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical Recipes in Fortran the Art of Scientific Computing*, 2nd edn. Cambridge University Press: New York; 617–622.

Randel WJ, Wu F. 2006. Biases in stratospheric and tropospheric temperature trends derived from historical radiosonde data. *Journal of Climate* **19**: 2094–2104.

Seidel DJ, Angell JK, Christy J, Free M, Klein SA, Lanzante JR, Mears C, Parker D, Schabel M, Spencer R, Sterin A, Thorne P, Wentz F. 2004. Uncertainty in signals of large-scale climate variations in radiosonde and satellite upper-air temperature datasets. *Journal of Climate* **17**: 2225–2240.

Sherwood S, Lanzante J, Meyer C. 2005. Radiosonde daytime biases and late 20th century warming. *Science* **309**(5740): 1556–1559.

Solow AR. 1987. Testing for climate change: an application of the two-phase regression model. *Journal of Climate and Applied Meteorology* **26**: 1401–1405.

Thorne PW, Parker DE, Tett SFB, Jones PD, McCarthy M, Coleman H, Brohan P. 2005a. Revisiting radiosonde upper air temperatures from 1958 to 2002. *Journal of Geophysical Research* **110**: D18105. DOI: 10.1029/2004JD005753.

Thorne PW, Parker DE, Christy JR, Mears CA. 2005b. Uncertainties in climate trends – Lessons from upper-air temperature records. *Bulletin of the American Meteorological Society* **86**: 1437–1443.

Uppala SM, Kallberg PW, Simmons AJ, Andrae U, da Costa Bechthold V, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, van den Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Holm E, Hoskins BJ, Isaksen L, Janssen PAEM, Jenne R, McNally AP, Mahfouf JF, Morcrette JJ, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth K, Untch A, Vasiljevic D, Viterbo P, Woollen J. 2005. The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society* **131**: 2961–3012.

Vincent LA. 1998. A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of Climate* **11**: 1094–1104.

Wang XL. 2003. Comments on "detection of undocumented changepoints: a revision of the two-phase regression model". *Journal of Climate* **16**: 3383–3385.
Wang Y, Ren GY. 2005. Change in free atmospheric temperature over china during 1961–2004. *Climate and Environment Research* (in Chinese) **10**(4): 780–790.
WMO. 2007. *GCOS Reference Upper-Air Network (GRUAN): Justification, Requirements, Siting and Instrumentation Options*. GCOS report No 112, WMO (WMO/TD No. 1379).
Zhai PM, Eskridge RE. 1996. Analysis of inhomogeneities in radiosonde temperature and humidity time series. *Journal of Climate* **9**: 884–894.

Appendix A – Two-Phase Regression (Tpr) Method

The TPR method, a technique initially described by Solow (1987), was used to detect breakpoints. Easterling and Peterson (1995) developed a variation of the TPR in which the regression lines were not constrained to meet, and a linear regression was fitted to the part of the difference series before the point being tested and another part after the year being tested:

$$T(i) = T_{CAN} - T_{REF} \text{ and } T(i) = \mu + \alpha i \qquad \text{(A1)}$$

T_{CAN} and T_{REF} represent candidate (station merged time series) and reference (station night-time series adjusted at recorded metadata events) series, respectively. The residual sum of squares from a single regression through the entire time series was also calculated with a critical value $U(i)$:

$$U(i) = \frac{(RSS_1 - RSS_2)/3}{RSS_2/(n-4)} \qquad \text{(A2)}$$

for which RSS_1 is the residual sum of squares for $i = 1, \ldots .c$, and RSS_2 is the residual sum for $i = c + 1, \ldots .n$. The significance of the two-phase fit was tested with a likelihood ratio statistic using the two residual sums of squares and with the difference in the means of the difference series before and after the discontinuity as evaluated by a Student's t-test and the multi-response permutation procedure (MRPP; Mielke, 1991). If the breakpoint was significant at the 95% level (probability: $P = 0.05$), it was considered a true discontinuity. If the discontinuity was significant, the time series was subdivided into two at that year and a break point was assigned to this point.

A homogenized array was created by

$$T^1(i) = \begin{cases} \mu_1 + \alpha_1 i & 1 \le i \le c \\ \mu_2 + \alpha_2 i & c < i \le n \end{cases} \qquad \text{(A3)}$$

with α_1 and α_2 calculated by

$$\alpha_1 = \frac{\sum_{i=1}^{c}(i - \bar{i})(T_1 - \overline{T})}{\sum_{i=1}^{c}(i - \bar{i})^2}, \quad \mu_1 = (\overline{T} - \alpha_1 \bar{i}) \text{ and}$$

$$\alpha_2 = \frac{\sum_{i=c+1}^{n}(i - \bar{i})(T_1 - \overline{T})}{\sum_{i=c+1}^{c}(i - \bar{i})^2}, \quad \mu_2 = (\overline{T} - \alpha_2 \bar{i}) \qquad \text{(A4)}$$

The adjustment that was applied to all data points prior to the discontinuity was the difference in the means of the two windows of the difference series. This test was repeated for all years of the time series up until the discontinuity became insignificant.

Appendix B – Quarc Methodology

A breakpoint is defined as a change in the mean value of the time series that is a direct result of a change in instrumentation or observing practice. In order to test the null hypothesis we use two pieces of information, the probability of rejecting the null hypothesis from a statistical breakpoint identification, S, and a probability from the (known to be incomplete) metadata record of changes at a given station, M. We define the joint probability of obtaining M and S given the null hypothesis as,

$$P(M \cap S|H_0) = P(S|H_0)P(M|H_o) \qquad \text{(B1)}$$

Equation (1) assumes that M and S are dependent only through the breakpoints, i.e. they are conditionally independent.

A non-parametric Kolmogorov–Smirnov test (Press *et al*., 1992, and hereafter K–S test) is applied to the time series of seasonal mean differences between station data and weighted composites of data from neighbouring stations to provide S. The approach assumes that the neighbour reference series is a reasonable estimate of the common natural variability between the target station and its neighbours. Multiple K–S test statistics are calculated, one for each pressure level, $P(L_1), P(L_2), \ldots, P(L_{\mathrm{n}})$. In order to maintain consistency between the individual pressure levels, the $P(S|H_0)$ component of the breakpoint detection algorithm is estimated from the geometric mean of the available K–S statistics.

$$P(S|H_0) \propto \left(\prod_{k=1}^{n} P(L_k)\right)^{\frac{1}{n}} \qquad \text{(B2)}$$

We use a simple subjective probability model to estimate $P(M|H_0)$. We set a background value of 1, with each metadata event assigned a given date represented as a minimum with a cut-off point six seasons either side of the reported timing of the event. This accounts for potential uncertainty in the reported date.

Breakpoints are identified from the product of the K–S statistic and metadata statistic (Equation 1), referred to as the breakpoint score, since it is not, strictly speaking, a probability. The lower the score the greater confidence we

DOI: 10.1002/joc

have that a potential breakpoint exists at that time point. Minima below a specified critical value are assigned as breaks.

Adjustment factors are derived from the time series of station minus neighbour differences, calculated as the difference in the medians of pre-defined periods either side of a breakpoint. If another breakpoint exists within this window, the window is reduced accordingly. At least five data points are required either side of the breakpoint; otherwise it is ignored. A number of tests are made to confirm that the adjustment estimates are well defined and not influenced by breakpoints in the neighbour composite or outliers in the neighbour or station time series.

The system is run iteratively. The adjusted data from each iteration are fed back through the system, re-calculating neighbour composites. In early iterations we set a very low critical value threshold for breakpoints so that only the worst breakpoints are identified. In later iterations, after these worst offenders have been removed, we have relaxed this threshold to detect smaller breakpoints, or re-calculate adjustments that were rejected in earlier iterations but that are now better constrained. The automated system is critically reliant on a number of parameters that will directly or indirectly influence the number of breakpoints detected, false detection rates, and adjustment estimates (Appendix A of McCarthy *et al*., 2007). These are varied randomly within the ensemble considered here with values derived from a random number generator.

DOI: 10.1002/joc