Robust Tropospheric Warming Revealed by Iteratively Homogenized Radiosonde Data

STEVEN C. SHERWOOD, CATHRYN L. MEYER, AND ROBERT J. ALLEN

Yale University, New Haven, Connecticut

HOLLY A. TITCHNER

Met Office Hadley Centre, Exeter, United Kingdom

(Manuscript received 7 November 2007, in final form 11 March 2008)

ABSTRACT

Results are presented from a new homogenization of data since 1959 from 527 radiosonde stations. This effort differs from previous ones by employing an approach specifically designed to minimize systematic errors in adjustment, by including wind shear as well as temperature, by seasonally resolving adjustments, and by using neither satellite information nor station metadata. Relatively few artifacts were detected in wind shear, and associated adjustments were indistinguishable from random adjustments. Temperature artifacts were detected most often in the late 1980s–early 1990s. Uncertainty was characterized from variations within an ensemble of homogenizations and used to test goodness of fit with satellite data using reduced chi squared.

The meridional variations of zonally aggregated temperature trend since 1979 moved significantly closer to those of the Microwave Sounding Unit (MSU) after data adjustment. Adjusted data from 5°S to 20°N continue to show relatively weak warming, but the error is quite large, and the trends are inconsistent with those at other latitudes. Overall, the adjusted trends are close to those of MSU for the temperature of the lower troposphere (TLT). For channel 2, they are consistent with two analyses (Remote Sensing Systems, $p = 0.54$, and the University of Maryland, $p = 0.32$) showing the strongest warming but not with the University of Alabama dataset ($p = 0.0001$). The troposphere warms at least as strongly as the surface, with local warming maxima at 300 hPa in the tropics and in the boundary layer of the extratropical Northern Hemisphere (ENH). Tropospheric warming since 1959 is almost hemispherically symmetric, but since 1979 it is significantly stronger in ENH and weaker in the extratropical Southern Hemisphere (ESH). ESH trends are relatively uncertain because of poor sampling. Stratospheric cooling also remains stronger than indicated by MSU and likely excessive.

While this effort appears not to have detected all artifacts, trends appear to be systematically improved. Stronger warming is shown in the Northern Hemisphere where sampling is best. Several suggestions are made for future attempts. These results support the hypothesis that trends in wind data are relatively uncorrupted by artifacts compared to temperature, and should be exploited in future homogenization efforts.

1. Introduction

The question of whether tropospheric temperatures are participating as expected in climate change has been controversial, with some observing systems reporting changes that are inconsistent with the models (CCSP 2006; National Research Council 2000) and inferences from melting tropical glaciers (Thompson et al. 2006). Recent work indicates that early estimates of atmospheric warming were too low, but that discrepancies between expected and measured tropospheric warming rates have not been fully explained in the tropics (CCSP 2006), although some analyses of the satellite record have reported concordance with surface warming (Fu et al. 2004; Vinnikov et al. 2006). Radiosonde and many satellite analyses indicate relatively weak tropospheric warming since 1979 (when satellite monitoring began), though the radiosonde record indicates stronger warming before 1979. Thus, attention will be paid here to the two time periods before and after 1979.

Our longest and most detailed record of "upper air"

Corresponding author address: S. Sherwood, Yale University, New Haven, CT 06520.
E-mail: ssherwood@alum.mit.edu

DOI: 10.1175/2008JCLI2320.1

temperatures comes from the radiosonde record, which achieved significant global sampling in the years after the 1958 International Geophysical Year. Trends in this record have been reported by a number of studies (e.g., Angell and Korshover 1975), but always with caution because of the numerous changes (many of them undocumented) in observing practice and instrumentation that likely affected climatic variations and trends. Several recent studies have attempted to detect and correct steplike artifacts in records from selected stations. Lanzante et al. (2003, hereafter LKS) embarked on a painstaking, subjective analysis of 87 individual station time series through 1997, occasionally using neighbor time series and/or climate indices, such as the ENSO index, to help identify natural variability. The resulting records were used by Thorne et al. (2005) as a backbone to quantify natural variability and aid in detection at a much larger number of stations. Haimberger (2007) made similar use of forecasts from a forecast model driven by many observing systems. The LKS results were extended through more recent years by an automated procedure "Radiosonde Atmospheric Temperature Products for Assessing Climate" ("RATPAC"; Free et al. 2005). Most recently, Christy et al. (2007) attempted to homogenize tropical stations since 1979 using satellite data.

The difficulties in this type of homogenization are foreshadowed by earlier studies (Free et al. 2002; Gaffen et al. 2000) finding that the process tends to remove whatever trend is present in the data. Indeed, doubts remain about the ability of the methods employed so far to fully recover climate signals. Small and hard-to-detect artifacts may be pervasive in the record (Randel and Wu 2006; Sherwood et al. 2005), and false detections can easily cause problems. Recent investigations at the Hadley Centre indicate systematic underestimation of trends in simulated tests (Titchner et al. 2008), while recent revisions of the Haimberger (2007) methodology have found the global mean trend to be sensitive to errors in the reanalysis background field used (Haimberger et al. 2008). These findings indicate that new methods may still be needed.

A detailed exploration by Sherwood (2007, hereafter S07) using statistical simulations revealed that standard methods were often unable to estimate trends reliably. Three problems were identified. First, even with liberal detection criteria not all changepoints are found; this is the "missed artifact" problem. On the other hand, even with very strict criteria, false changepoint detections are unavoidable when time series have realistic serial correlation. Subsequent adjustment of the time series tended to eliminate trends (or, in the case where a satellite reference is used, trends in the sonde–satellite difference); this is the "greedy artifact" problem. Finally, when reference information from nearby stations was used, artifacts at neighbor stations tend to cause adjustment errors; this is the "bad neighbor" problem. In this case, after adjustment, climate signals became more similar at nearby stations even when the average bias over the whole network was not reduced.

S07 concluded that the best approach was to detect changepoints liberally and employ a method designed to minimize the impacts of false detections and bad neighbors. He recommended, in particular, an approach called iterative universal kriging (IUK). Key elements of this approach are the use of an underlying model to impute missing data values, and the fitting of the data to a model that simultaneously treats artifacts and natural fluctuations. Here we briefly discuss further tests of this method and present the results of applying it to the global radiosonde network. Another advance on previous work is the use of observed wind shear fluctuations to help identify natural temperature variability, as recommended by Allen and Sherwood (2007).

2. Data

We analyzed twice-daily data, for several reasons. First, individual observations are expected to have homoscedastic error behavior (similar variance for all observations) as assumed by all methods, while monthly means will not because of large variations in sampling rate at some stations. Second, it was hoped that sampling biases (such as "foul weather" biases or balloon-burst biases resulting from bursting of the balloon in cold air) could be removed though skillful imputation of missing values. Finally, series of 0000 minus 1200 UTC temperature differences from adjacent launch pairs are an especially sensitive indicator of bias changes (Haimberger 2005; Sherwood et al. 2005).

We began with all available 0000 and 1200 UTC launches from the Integrated Global Radiosonde Archive (IGRA; Durre et al. 2006; dataset available online at http://www.ncdc.noaa.gov/oa/cab/igra/index.php) from 1959 to 2005. Sampling prior to 1959 is too poor to attempt climatic analysis, and near-modern coverage is not achieved until the late 1960s. The IGRA archive includes some 2000+ stations, though only a minority of these collected data with any regularity over substantial periods of time. At many stations, especially in the tropics, data are sparse or completely absent at one of the two nominal observing times (usually the one falling during local nighttime). For simplicity, we began with data only from the "mandatory reporting" levels of 850, 700, 500, 400, 300, 250,

200, 150, 100, 70, 50, and 30 hPa. We omitted 20 hPa and above because of insufficient data, and 1000 hPa because of the poor quality of the data (LKS). The variables we examined were temperature T, zonal shear $S_x \equiv du/dz$, and meridional shear $S_y \equiv dv/dz$ at each level. Shears were obtained by the finite difference of winds. For target levels from 700 through 150 hPa we used centered differencing of the nearest two mandatory levels, while elsewhere we employed a noncentered difference between the target level and the one closer to the midtroposphere (this to avoid using data at 1000 hPa and to minimize scarcity problems in the stratosphere). Finally, we omitted the 700- and 400-hPa levels from further analysis to bring closer parity to the number of tropospheric and stratospheric levels, for reasons discussed in appendix B.

For each variable, we first discarded any data values differing from the median at that station and level by more than six pseudo–standard deviations (see Lanzante 1996, hereafter L96). We then removed any station with more than 90% of either T or **S** data missing in the upper troposphere during either the first or last third of the time period; thus, a long-term average of six observations per month was deemed sufficient to at least carry the station through the analysis.

We next formed a series of adjacent 0000 minus 1200 UTC temperature difference (dT) values at each level and station, again discarding any more than six pseudo–standard deviations away from the median. Those stations having at least 25 dT values at 200 hPa in each of the first and last third of the time period [judged to be the minimum number needed to detect changes, see Sherwood et al. (2005)] were put into "group A," with the remaining stations, which will be harder to homogenize, relegated to "group B."

Data were divided into the following three latitude belts for homogenization and analysis: the tropics (30°N–30°S), the extratropical Southern Hemisphere (ENH), and the extratropical Northern Hemisphere (ESH). Table 1 shows the number of stations in each group and latitude band. Because of the large number of group A stations in ENH, the 37 group B stations there were not used.

TABLE 1. Station count by region. Groups A and B are those with and without, respectively, sufficient twice-daily data to detect solar heating changes. Group B stations in the ENH region were not used in this study.

Region	No. group A	No. group B
ENH	309	(37)
Tropics (30°S–30°N)	132	48
ENH	19	19

3. Overview of methodology

We assume at the outset, as have most others, that artifacts in the record consist of stepwise changes in bias. The first task is then to detect when and where these occur; the second is to estimate the bias changes. We iterated through these tasks several times in a series of "rounds" designed to address the easiest problems first. All detections were performed on deseasonalized monthly mean time series using one of the following two multiple changepoint detection schemes: two-phase regression (Wang 2003) and the nonparameteric method of Lanzante (1996) (see appendix A for more details). Bias changes were estimated from twice-daily data using a maximum-likelihood fitting method.

First, the activity in each "round" is summarized, with details of the different steps given later:

- Round 1, blending 0000 and 1200 UTC data: We assume that 1) any observable, long-term change in reported dT must have been due to instrumental changes altering the net effect of sunlight on the instrument or balloon (Sherwood et al. 2005), and 2) these instrumental changes could also have affected nighttime readings. Accordingly, we began by detecting changepoints in dT at group A stations; applying offsets to the daytime data prior to each changepoint to eliminate changes in dT at any level; and, finally, empirically removing the diurnal cycle from each station. This enabled us to combine both observing times into a "diurnally adjusted" series, which was used for the remainder of the analysis. Times of detected dT change were retained for use in the subsequent steps.
- Round 2, preliminary homogenization of group A: Local changepoints (LCPs) in temperature or wind shear were detected (independently) at each level of each group A station. To help avoid false detections, for each variable a clustering procedure (detailed in appendix B) was applied to the set of LCPs detected at a station; for temperature this set was the combination of LCPs detected from T and from dT (retained from round 1), while for **S** it was the LCPs detected in either component of **S**. Clusters spanning at least four levels were assumed to correspond to a single changepoint (CP) affecting all levels. Others were discarded as being probable false detections (this assumption is supported by a decreasing agreement, not shown, with other methods when only one or two LCPs are required, and by a test described in section 4a showing that artifacts tend to be vertically coherent). We made an exception, however, for iso-

lated shifts in dT detected during round 1, keeping those changepoints at the level(s) found. This was because changes in the application of radiation corrections could cause discontinuities in dT at only one or two levels resulting from, for example, changing the altitude range for an applied correction (LKS).

Level shifts were estimated independently for each level and season using IUK (see section 3a). Season-specific shift amounts crudely allow for any dependence of biases on either climate or solar zenith angle.

Note that because level shifts were reestimated here at most points where a daytime adjustment was made in round 1, those adjustments were effectively overridden, serving only to help blend the two observing times. This means that the final adjustments to the data are nearly independent of the day–night difference trends noted by Sherwood et al. (2005); in other words, our methodology does not strongly prefer nighttime over daytime data.

- Round 3, second homogenization of group A: The IUK fitting procedure decomposes the data into coherent variations, artificial changes, and local anomalies (see below). Artifacts should appear in the second or third components, and should be more evident if the first component is left out. Thus, following S07, we reapplied the above detection procedure to data reconstructed from *those components only,* to improve detection and estimation for both temperature and shear.
- Rounds 4 and 5, homogenization of group B: We finally repeated the procedures of rounds 2–3 with group B stations, retaining the homogenized group A data from round 3 to aid in estimation (both rounds) and detection (round 5). Because dT data are not available for these stations, the round 1 procedure could not be repeated, and the aggregation procedure for temperature was applied only to LCPs detected from T. No further detections were done for group A, but level shifts were reestimated both times at the group A stations (as it happened, generally with little change from their round 3 values).

Level-shift estimation

Level shifts were estimated using IUK (Sherwood 2000, hereafter S00). This involves regressing the (incomplete) data at one pressure level onto a model that includes both natural and artificial variability,

$$\mathbf{Z} = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 + \boldsymbol{\epsilon}, \tag{1}$$

where each of these quantities is a vector function of station and time. In this study $\mathbf{Z}$ is the vector T, S_x, S_y. The term $\boldsymbol{\mu}_1$ represents large-scale variability of the true field as a linear superposition of basis functions:

$$\boldsymbol{\mu}_1(s,t) = \sum_{i=1}^{m} \mathbf{a}_i \mathbf{f}_i(s,t) + \sum_{i=1}^{n} \mathbf{b}_i \mathbf{g}_i(s,t), \tag{2}$$

while $\boldsymbol{\mu}_2$ represents artificial changes, and $\boldsymbol{\epsilon}$ is small-scale and other unmodeled variability. This representation follows S07, except for the notational difference that in that study the entire basis was included in a single μ. The variables s and t in (2) are the discrete location and time coordinates, respectively, at which measurements of $\mathbf{Z}$ are nominally available. After regression (M step), this model is used to impute missing values (E step), whereupon the model is refitted; iteration of this procedure converges to the maximum-likelihood values of all regression parameters given the incomplete data and model structure (see S00). Simultaneous regression onto natural and artificial variability avoids the "greedy artifact" problem (Wang 2003; S07).

The function $\mathbf{f}$ represents the signal patterns (linear trend, ENSO response, etc.), while $\mathbf{g}$ represents other natural variations. As in previous applications we adopt a linear function of time for $\mathbf{f}$ and a truncated series of empirical orthogonal functions (EOFs) for $\mathbf{g}$ (see appendix C for more details). In S00's application of the method to seasonal wind data in the tropical upper troposphere and lower stratosphere, six empirical functions $\mathbf{g}$ were retained; the first corresponded to the quasi-biennial oscillation and the second to a residual seasonal cycle. An advantage of using EOFs is that, because they seek to explain variance at many stations at once, they are relatively unlikely to be affected significantly by artifacts at one or a few stations; this feature is further strengthened by the use of wind data, whose artifacts are likely to be at different times than those of temperature. Thus, $\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2$ should serve as a nearly homogeneous model for the variability at a station, minimizing the "bad neighbor" problem. If all stations in a large country experienced simultaneous bias shifts in the same direction some "bad neighbor" behavior would probably remain, but probably less than with other neighbor-based approaches.

The second term in (1), $\boldsymbol{\mu}_2$, is a linear superposition of station-specific Heaviside step functions, one centered on each changepoint. The corresponding amplitude parameter is the level shift to be estimated. Thus, the homogenized data are $\boldsymbol{\mu}_1 + \boldsymbol{\epsilon}$. The field with coherent variability removed, $\boldsymbol{\mu}_2 + \boldsymbol{\epsilon}$, is what was used in round 3 above.

Less coherent variability is represented in (1) by the field $\boldsymbol{\epsilon}$, which was modeled as a Gaussian random field having a spatially homogeneous and stationary, though

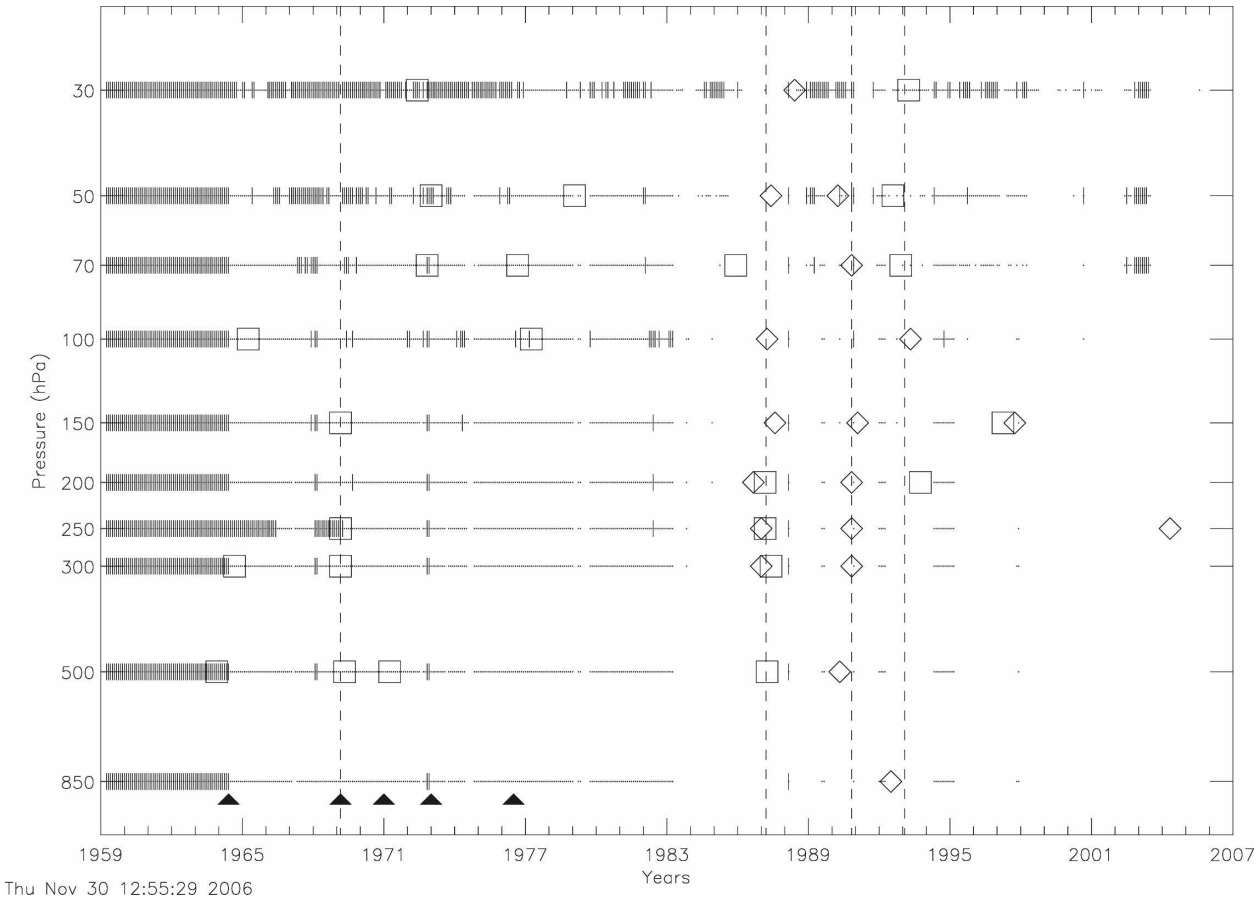

FIG. 1. Results of changepoint detection at Niamey Aerodrome in Niger (WMO 61052) with the L96 scheme. Diamonds and squares represent LCPs found in rounds 1 and 2, respectively, and vertical dashed lines show locations of final CPs. Vertical hatching by level shows months with insufficient data, while small dots show months with sufficient data at only one time of day. Solid triangles at bottom show times of CPs detected by LKS at this station at 300 hPa.

anisotropic, autocovariance function $\sigma_\epsilon(d\mathbf{x}, dt)$ (e.g., Daley 1991) characterized by four independent parameters for each variable (T, S_x, and S_y).

In principle, all levels could have been analyzed at one time, but this would have been problematic for the computer's memory. To the extent that natural and artificial changes are vertically similar, the benefits of this would be limited. However, it is possible that stratospheric variability could have been better estimated with the assistance of tropospheric data, something that could be exploited in future studies.

4. Performance evaluation

a. Detection

Sensitivity tests indicate that many, if not most, detected changepoints in T occurred at times of identifiable changes in dT. Occasionally, level shifts in T opposed those applied in the previous round to the daytime data. Day–night differences are thus useful in detection, but nighttime data should not be assumed as being homogeneous.

In Fig. 1 we show the T detection results at one illustrative group A station (Niamey). Data were available only after mid-1964, were sparse in the stratosphere, and were usually collected only once per day until the second half of the period. Detections at individual levels (LCPs) showed a promising tendency to cluster near certain times. These characteristics were fairly typical.

More atypical was the poor correspondence at this particular station between our changepoints and those of LKS (only one match). This particular station emphasizes some methodological differences between the studies: LKS posited several changepoints early in the record on the basis of time-of-observation changes reported in the station metadata, while our detections were mostly later in the record and were found in day–night difference data. LKS did not examine nighttime data at this particular station because of the small

amount of it. Among the 64 group A stations examined by LKS, we found 169/160 CPs (using L96/two phases, respectively) compared to 60 detected by LKS at 300 hPa; 25/25 of ours corresponded (within ±6 months) to an LKS detection. Thus, our approach was more aggressive than LKS, but we did often agree on CP locations despite cases like Fig. 1.

Because we have made no use of station metadata to aid in detection, we can test performance by comparing detected changepoint times with the times of known changes. One well-documented example of an instrument change producing a significant bias change is the change of radiosonde type in Australia during the late 1980s (LKS; Parker and Cox 1995). We picked up changepoints from 1986 to 1988 at nearly all Australian A stations, but only a minority of B stations, suggesting that we may not have achieved thorough detection at B stations. The overall detection rate is shown over time in Fig. 2, and shows a peak in the late 1980s to early 1990s. This corresponds roughly to a similar maximum in station metadata events (P. Thorne 2007, personal communication), as does the sharp peak in 1969.

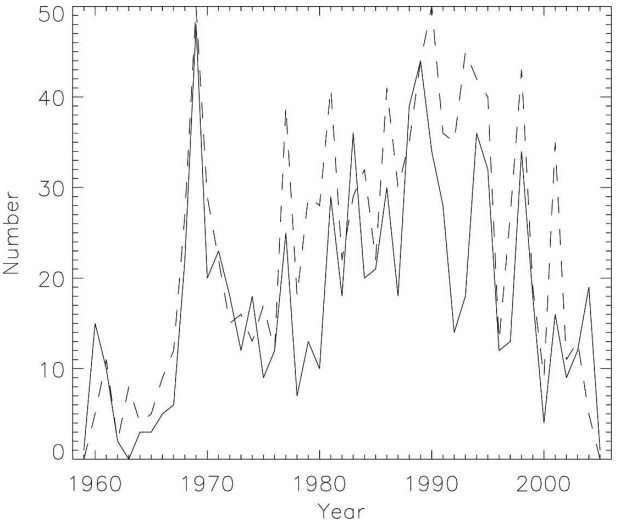

FIG. 2. Number of detected T changepoints per year with two-phase regression (solid) and L96 (dashed) methods, among group A stations at round 3.

It is not clear how well detections should correspond to metadata in detail because, unfortunately, many events recorded in the metadata do not significantly affect observing bias, and events that are significant may be unrecorded. We compared our detections at 12 Comprehensive Aerological Reference Data Set (CARDS) stations [records examined by Free et al. (2002)] with metadata from the CARDS dataset and with the results of other detection methods evaluated in that study. Only 23% of our detections corresponded with recorded metadata events within 6 months. We calculated the same percentage for detections of the only other method that did not use metadata (that of the University of Alabama at Huntsville (UAH), which used satellite comparisons to detect artifacts). Either both methods performed poorly, or metadata are at best an incomplete guide in finding important artifacts, despite the broad agreement in time variation of metadata and detections.

One way to discriminate between correct and false detections should be to examine the vertical coherence of estimated shifts. Natural variability is anticorrelated in the troposphere and stratosphere (CCSP 2006; Kiladis et al. 2001; Riehl 1954; Wu et al. 2006), so a false detection resulting from a natural fluctuation should usually end up with level shifts of opposite signs in these two layers. By contrast, instrumental changes may reasonably be expected to cause shifts having, more often than not, the same sign at any two levels (e.g., Randel and Wu 2006; Sherwood et al. 2005). We computed the correlation coefficient r_{ij} between level shifts at a given pair of pressure levels i, j over all detected changepoints at the 12 stations above, and then averaged over $i \neq j$. We found r ranging from 0.41 to 0.47, compared with a value of 0.17 for randomly chosen times. Given the noise levels in the data, it is unlikely that $r > \sim 0.5$–0.7 would be possible even with perfect vertical coherence of bias changes. Our r thus suggests that correct detections are dominating false ones and demonstrates that the shifts do tend to be at least somewhat vertically coherent. This result also supports our decision to insist on detections at multiple levels for assigning a changepoint. We conclude that the 23% correspondence with metadata reflects inadequacy of the metadata more than failure of our method.

Finally, we detected 30/13 changepoints in wind shear at the same set of stations. Thus, we find much less evidence for heterogeneities in wind shear than in temperature, consistent with the conclusions of Allen and Sherwood (2007) and with the results of another new study (Gruber and Haimberger 2008).

CP detection rates in round 3 were 16% lower than in round 2 for T and more than 50% lower for **S**. Interestingly, the number of T detections matching those of LKS held steady (25/27 compared to 25/25 in round 2) despite this drop. Because LKS used a very different methodology, it is reasonable to suppose that their errors are largely independent of ours, implying that the CPs "lost" in this round were probably false detections. This suggests that round 3 results will indeed be an improvement over those of round 2.

b. Level-shift estimation

In a set of idealized simulations, S07 found level-shift and trend estimation by IUK to be fully unbiased when changepoints were known a priori and nearly unbiased when they had to be detected from the data. This stood in contrast to other methods examined in that study, whose level-shift estimates were contaminated either by the underlying trend or by artifacts at neighbor stations. However, the simulated datasets were generated from a model structurally identical to that used for analysis, so the robustness of IUK to the character of the natural variability was not tested.

Recently, the IUK method has been set up to run at the Met Office's (UKMO's) Hadley Centre, and a number of tests have been performed using more elaborate simulations. An atmospheric general circulation model forced with observed sea surface temperatures was sampled at the times and locations of actual sonde launches, and several thousand artificial jumps were introduced to simulation station records. These simulations will be described in detail elsewhere (Titchner et al. 2008). IUK was used to estimate level shifts given either (a) perfect knowledge of changepoints, or (b) changepoints detected by the automated UKMO system (McCarthy et al. 2008).

In the latter tests IUK performed reasonably, particularly with regard to the precision of individual shift estimates (~0.3°C rms error in the troposphere globally). Also, as in S07, estimation errors were fully independent of either the mean trend or mean level shift in the network. However, in some tests with known changepoints estimated trends were still significantly off from the truth when averaged over the network. This was because the EOFs turned out to capture intraseasonal and interannual variability reasonably well, but not decadal variability. Decadal variations were then aliased onto μ_2 instead of μ_1, which affected the trend when changepoint times happened to correlate with decadal fluctuations. Because the decadal fluctuations were globally coherent, the resulting errors were as well (by contrast, other sources of error in estimating trends at individual stations did indeed prove to be independent).

Fortunately, one can test the fit of the model by examining ϵ. In particular, we found that the mean of ϵ in the months before versus after changepoints did not always match, and that the mismatch in ϵ consistently agreed (to within sampling uncertainty) with the mean shift error. This renders the problem detectable in practice, at least approximately, to an accuracy arguably no worse then could be achieved with any similar approach given that the limiting factor is sampling uncertainty near the changepoint. This diagnosis applied to the actual data revealed no detectable bias in the tropics or ESH, but a systematic underestimation of shifts in the ENH troposphere of about 0.05°–0.10°C. This implies a small trend overestimate of ~0.03°C decade^{-1} in the ENH troposphere. Note that this is only the bias resulting from shift estimates at detected changepoints, and says nothing about possible biases remaining resulting from undetected artifacts.

It is tempting to test a homogenization effort by comparing retrieved trends at neighboring stations, to see if they are more consistent. A simple measure for our purposes is the standard deviation of the trend across stations; at most levels this turns out to be similar or slightly greater in the adjusted data than in the raw data (e.g., increasing from 0.21° to 0.28°C decade^{-1} at 300 hPa). Does this mean our effort failed? Tests of a variety of methods at the Met Office's Hadley Centre indicate that those procedures most successful in rms correction tend to be no more successful than others, and sometimes less so, at estimating large-scale trends (see Sherwood et al. 2008, manuscript submitted to *Geophys. Res. Lett.*). Thus, the failure to reduce rms error is not a sign that our procedure has failed to reduce large-scale trend errors. It does suggest, however, that for anyone interested in trends at only one or two stations, the current dataset may offer little improvement over the raw data other than perhaps to help quantify uncertainty.

The scatter of trends was about 3 times as large at stations in India as elsewhere. This is mostly due to the scatter in the raw data, but also because relatively large numbers of changepoints were detected, which add to the estimation error. Past studies have consistently reported problems with Indian data (e.g., Parker and Cox 1995).

5. Results

Diurnal T adjustments tended to increase trends at all latitudes, especially in the tropics, as anticipated from previous work (Sherwood et al. 2005). Subsequent T adjustment using IUK negated much of this warming in the deep tropics and ESH, but increased warming in the ENH region and the subtropics of both hemispheres. The adjustments in ENH showed a seasonal dependence, being roughly twice as large in summer as in winter with both schemes. There was no consistent seasonal dependence in the other two latitude bands.

Adjustments to **S** did not appear to be very important. As mentioned previously, few artifacts were found (on the order of 1 per 10 stations in round 3). Furthermore, the mean level shift assigned in any latitude did

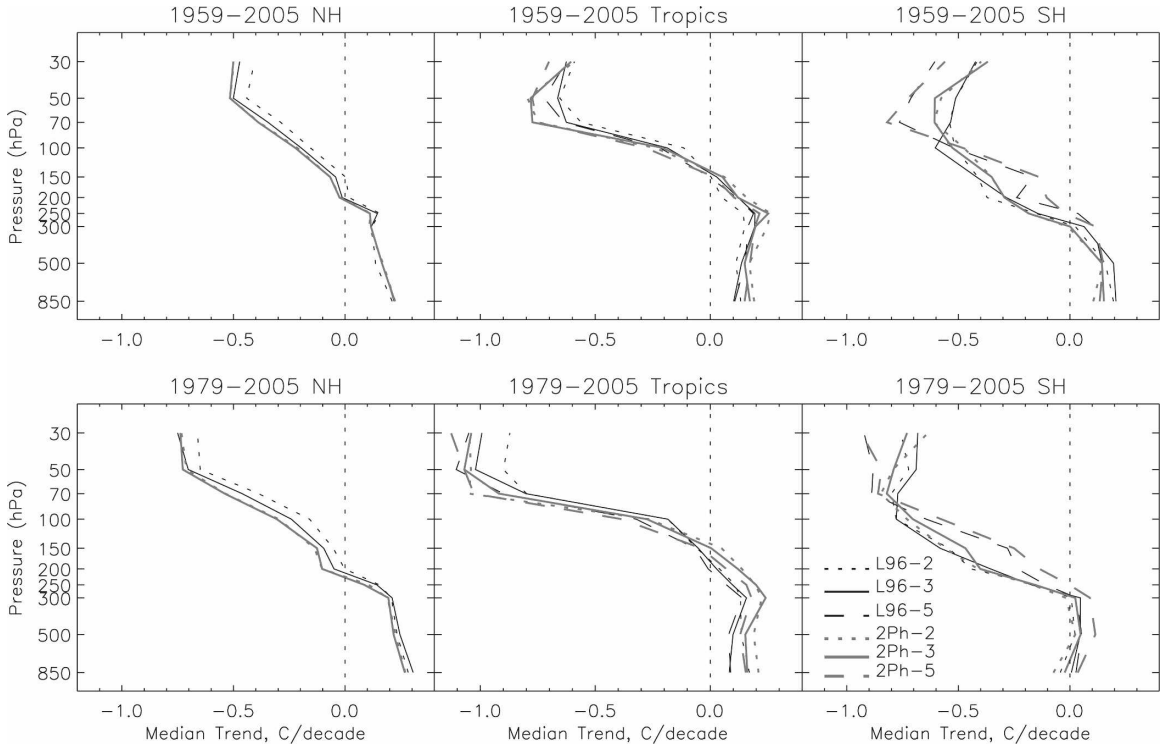

FIG. 3. Median trend vs pressure (from left to right) within three regions, (from top to bottom) for two time periods. Line type denotes round or stage of the procedure (see legend), with solid lines indicating our best estimate (round 3) and dashed lines showing results with group B stations added. Line density (black or gray) shows changepoint method.

not differ significantly from zero. Adjustments at individual stations were often significant compared to the trend, but, as for temperature, these adjustments did not necessarily make individual station records more accurate. Thus, we do not consider the adjusted shear dataset to be superior to the raw one and do not discuss it further.

a. Warming trend profiles

Because of the likelihood of artifacts remaining at some stations and the variable degree of uncertainty across stations, and hence the non-Gaussian distribution of resulting errors, we adopt the median as our location statistic for all purposes, judging this to be a more robust and objective estimate of the true mean than the alternative of discarding "bad" stations and taking the sample mean of what is left (sample means did not differ from medians by more than the uncertainties quoted below). Median trend profiles are shown for each region in Fig. 3. These profiles are encouragingly coherent with height, latitude, and time period, and peak near 300 hPa in the tropics as expected on physical grounds. Use of anomalies to aid in CP detection (round 3) led to the strongest trends. Tropical tropospheric warming rates were about the same over the two time periods, but in ENH the warming accelerated after 1979 and in ESH it approximately ceased. These changes are roughly consistent with the surface record (CCSP 2006), although tropical warming at the surface is thought to have accelerated somewhat during the second interval, which does not appear in our data.

Lower-tropospheric trends in ESH were highly heterogeneous since 1979 according to data from the Microwave Sounding Unit (MSU) satellite series, with some regions warming and others cooling (Mears and Wentz 2005). Given this, and the small number of ESH stations, our ESH trends must be treated with caution. Uncertainty is also indicated by the relatively large differences between trends with and without Group B data in ESH, especially in the stratosphere.

Stratospheric cooling rates were about 0.25°C decade^{-1} faster since 1979 than 1959 in all three latitude belts, implying a substantial acceleration. This acceleration is qualitatively consistent with accelerated ozone losses during the 1980s, but the cooling rates may still be too strong (see section 5b). The strongest seasonal cooling (not shown) was during September–November (SON; the Antarctic ozone hole season) from 100 to 50 hPa poleward of 70°S, reaching −2.8°C decade^{-1} for 1959–2005. Cooling there was near zero in the March–

May (MAM) and June–August (JJA) seasons. This seasonal variation was evident to some extent throughout the ESH stratosphere. Seasonal trend variations were small elsewhere.

Adding the Group B stations slightly reduced tropospheric warming trends in the tropics and increased them in ESH. Given the importance of *dT,* and the failure to detect suspected artifacts at many Australian Group B stations, we judge that the Group A results are more reliable. We therefore adopt the average of the L96 and two-phase round 3 results as our best estimate of the trend. Group B results do have some value in helping quantify structural plus sampling uncertainty, and may prove useful in subsequent studies. The uncertainty in tropospheric trends evident from these comparisons is $\sim$0.07°C decade^{-1}; in stratospheric trends, the uncertainty is $\sim$0.1°C decade^{-1}. The likely high bias of 0.02°C decade^{-1} in the ENH troposphere should also be recalled, but is well within the above uncertainty. Neither figure includes the effect of undetected changepoints, which we cannot quantify.

b. Trend comparison with MSU satellite

We next compare our results with trends from three published analyses of the MSU dataset (Fig. 4) by computing MSU-equivalent temperature trends using the static weighting functions.[1] Because global mean trends from MSU have been controversial and are sensitive to small calibration errors, we focus primarily on the horizontal variations in warming rate.

The following three MSU products are available: channel 4 (Fig. 4a) mainly observes the lower stratosphere (30–100 hPa), but receives some radiance from the upper troposphere; channel 2 mainly observes the free troposphere, but extends slightly into the lower stratosphere and down to the surface; and temperature of the lower troposphere (TLT) pseudochannel (obtained by differencing two zenith angles of channel 2) observes the lower troposphere and surface. Zonal mean sampling biases associated with the radiosonde network (not shown) turned out to be $<$0.05°C decade^{-1}, except near Antarctica (where they depend on which MSU product is used), and were much less when averaged over several neighboring latitudes, so we compare our data with the fully sampled MSU products. To illustrate the impact of adjustments, the results for unadjusted data are shown in Fig. 5.

In the lower troposphere (Fig. 5c), the sonde results are scattered about those of the two available MSU analyses. Adjustment brings the data closer to MSU, especially in ENH, although the warming is still slightly less than that shown by the MSU there (a discrepancy that increases slightly if diagnosed adjustment biases are accounted for). In the tropics the two MSU products diverge somewhat, with the sonde data appearing closer to that results of UAH, as found by Christy et al. (2007), albeit with large uncertainty. MSU TLT trends should be treated with some caution at higher northern latitudes because of the significant land coverage, which causes problems resulting from microwave emission from the land surface and elevated terrain (e.g., Mears and Wentz 2005). This problem is considerable also over Antarctica; hence, Remote Sensing Systems (RSS) does not provide a product poleward of 70°S. A second caution arises from our assumption that the surface trend matches that at 850 hPa; if observed surface warming rates (CCSP 2006) were used instead, with linear interpolation, the trends in high midlatitudes would increase by up to 0.04°C decade^{-1}.

For channel 2 (Fig. 5b), the adjustments have less effect because of the opposing directions of adjustments in the troposphere and lower stratosphere. For this channel the MSU analyses diverge more, showing similar meridional variation in warming but a well-known difference in the mean (see CCSP 2006). The meridional variation from sondes shows scatter but is roughly consistent with that of MSU, except from 5°S to 20°N where sonde trends dip substantially. This discontinuity with trends on either side is clearly incorrect, because it conflicts with all MSU analyses and is physically inconsistent with the small trend in wind shear (shown in the figure; see also Allen and Sherwood 2008). Only 3% of our stations fall in the 5°S–20°N range, accounting for the very large error bars there. In fact no trend from 20°S to 20°N departs by more than 2σ from any of the MSU products.

To quantify the overall consistency of our data with each MSU product, we computed the p value (probability that the mean-squared discrepancies would, given the uncertainty of each point, exceed those observed) from a reduced chi-squared goodness-of-fit test, allowing an intrinsic uncertainty for each MSU product of 0.04°C decade^{-1} at each latitude. For channel 2 this yields 0.56, 0.33, and 0.0001 for University of Maryland (UMd), RSS, and UAH, respectively. *While UAH is closer to our data in the deep tropics than are the other two products, the large error bars there mean that this region has little influence on overall goodness of fit, which is instead dominated by the ENH region.* It is not our intention to pass judgment on the MSU datasets

[1] To apply the weights, we set trends below 850 hPa equal to those at 850, and those at 10 hPa and above equal to half the trend at 30 hPa (based on results of LKS, whose analysis extended to 10 hPa). All interpolations were linear in log-pressure.

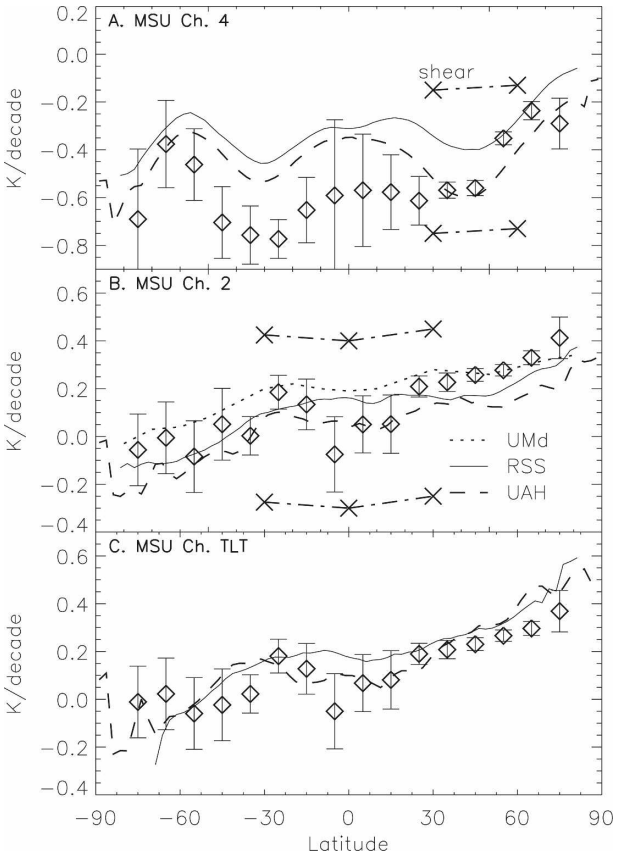

FIG. 4. Comparison of trends, 1979–2005. Lines show MSU channels (a) 4, (b) 2, and (c) TLT linear trends from RSS (solid) channel 2, version 3.0 (Mears et al. 2003)/TLT version 5.0 (Mears and Wentz 2005); UAH (dashed) TLT version 5.2/channel 2, version 3.0 (Christy et al. 2003); and UMd (dotted, channel 2 only; Vinnikov et al. 2006). Symbols show simulated MSU trends from round 3 averaged over the two detection schemes, with approximate 1-sigma error bars, based on the median trend at each pressure from stations at that latitude. The *differences* in vertical position of two crosses joined by dot–dashed lines show the meridional *differences* between the trend at the two latitudes implied by that of wind shear via thermal wind balance. Their vertical positions are arbitrary.

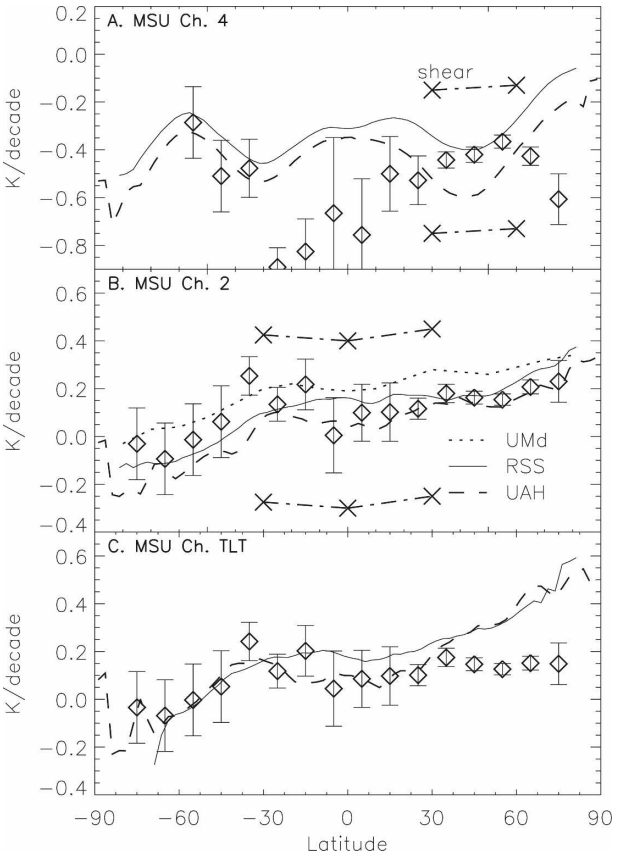

FIG. 5. As in Fig. 4, but for unadjusted data.

per se, but to assess what overall level of global tropospheric warming is most consistent with our data, using the various MSU estimates as straw-man hypotheses. We conclude that, for the channel 2 weighting region, an amount close to that of either RSS or UMd is most likely. The UAH channel 2 product may contain an error that still was not corrected as of this writing, but was corrected in TLT (C. Mears 2007, personal communication), and comments here may not apply to future versions of these datasets.

A similar test for the TLT product is more ambiguous, with p values of 0.27 and 0.09 for RSS and UAH. Thus, our data are marginally consistent with either TLT dataset, though UAH is more of a stretch. Given the problems with this channel noted above, we judge that we have achieved reasonable agreement with MSU and cannot really distinguish between the two MSU products in this case.

For channel 4 (Fig. 5a), unfortunately, significant disagreement between the adjusted sonde data and both MSU analyses remains. From 20°S to 20°N, we obtain -0.55°C decade^{-1}, falling between the -0.69°C decade^{-1} from RATPAC/LKS and -0.29 to -0.37°C decade^{-1} from MSU (CCSP 2006). Consensus has been that radiosonde errors are responsible for most of the discrepancy (CCSP 2006). Indeed, trends in lower-stratospheric wind shear from 30°N to 60°N (shown also in the figure) indicate that, according to thermal wind balance, the meridional gradient of the temperature trend is roughly correct in the satellite data rather than the sondes. Thus, we judge that our cooling trends are still probably too strong at near-tropical and southern latitudes, by at least 0.1°C decade^{-1}. One caution here is that we do not have sonde trends at 10 hPa, and if unexpected trends are occurring in the middle stratosphere, these will throw the comparison off.

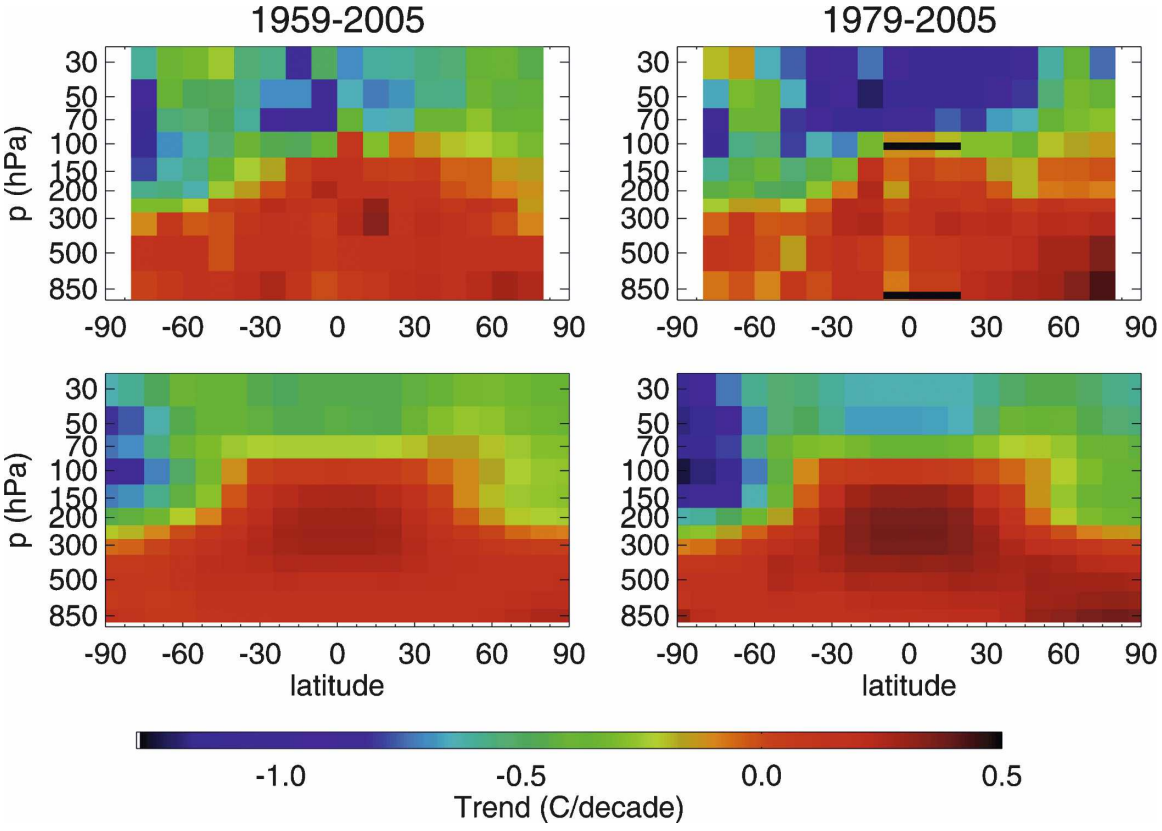

FIG. 6. (top) Latitude–height temperature trend cross sections during both periods; short thick bars indicate latitudes discussed in text where sonde adjustments in the troposphere still appear inadequate. (bottom) Average hindcast trends over the same periods from 11 coupled climate models (those including ozone depletion among the forcings) from the World Climate Research Program (WCRP) Coupled Model Intercomparison Project (CMIP3) multimodel dataset.

Given that (outside 5°S–20°N) the sonde warming is equal to or less than MSU for both TLT and channel 4 vertical weights, one would also expect it to be on the low side in channel 2, which straddles them. In fact it falls near the top of the reported MSU results for channel 2. This peculiarity occurs for both the RSS and UAH datasets, though much more extremely so for UAH. Insofar as the vertical distributions shown in Fig. 3 are very close to moist adiabatic, as, for example, predicted by GCMs (Fig. 6), this suggests a systematic bias in at least one MSU channel that has not been fully removed by either group. The discrepancy could in principle be explained by a surface temperature trend greater from that at 850 hPa, but this trend would have to be nearly 1.0°C decade^{-1}, which is far greater than that indicated by surface records. It could also arise from sonde overestimation of warming near the tropopause, where a priori physical expectations are less clear, or could be associated with biases evidently remaining in the stratosphere in our data. Thus, a firm statement is not yet possible.

6. Conclusions

Full homogenization of a dataset is probably impossible, so we have sought a procedure whose errors should not have systematic impacts on climate signals. Our procedure differs in several important ways from those applied previously, namely, by (a) making use of day-minus-night series for the precise removal of artifacts associated with solar heating, with other artifacts detected in a subsequent step; (b) using a shift estimation algorithm (IUK) that has been tested on both idealized data and data simulated by a climate model, and whose estimation biases should be relatively small and can be diagnosed a posteriori from residuals of fit; (c) estimating level shifts from twice-daily, rather than monthly or seasonal, data; (d) estimating seasonally dependent bias changes; (e) using wind shear data to aid in identifying natural variability, while also homogenizing the shear data; and (f) using only radiosonde data, with no auxiliary input from satellites, forecast models, station metadata, or previously homogenized datasets

as references. Because of characteristic (f) we can test our results against satellite and metadata. The resulting dataset was apparently not improved with respect to trend variance at individual stations, but was improved with respect to the meridional warming pattern shown by the MSU satellite. This is consistent with the expectation that systematic errors were reduced, with a small penalty paid on the side of random error, as shown in tests of the method on simulated datasets.

Adjustments here are computed from data at two synoptic observing times (0000 and 1200 UTC) folded together. These adjustments override most of the daytime adjustments made in the previous step, whose purpose was to enable the two observing times to be folded together in a consistent manner. Thus, our adjustments are nearly independent of those implied by Sherwood et al. (2005), who used nighttime data as a reference.

We obtained a few new results relevant to future radiosonde studies. Artifacts in wind shear were far fewer than those of temperature and had no statistically significant impact on trends. Temperature artifacts were much more frequent in the late 1980s and early 1990s than at other times, corroborating independent evidence from metadata of more frequent changes at this time. In the Northern Hemisphere extratropics (ENH), level shifts were roughly twice as large in summer as in winter, indicating that future homogenization efforts may also want to consider seasonally dependent bias adjustments.

It does not appear that the homogenization effort described here was completely successful; the meridional trend differences in the tropics since 1979 were physically unrealistic and inconsistent with those shown by the MSU satellite. Artifacts evidently remain in the troposphere in some of the stations from 5°S to 20°N, because trends there are too low compared to those at other latitudes. Trends at other latitudes agree fairly well with those of MSU for the lower troposphere. For channel 2 (which peaks near 500 hPa), they fall between those two analyses (RSS and University of Maryland) showing the most warming. A statistical test showed that based on our data, a warming profile for channel 2 similar to that of RSS or UMd could not be rejected, but that one resembling a third product (UAH) could be rejected at high significance. In the stratosphere, the adjusted data show a more realistic latitudinal variation at most latitudes, but more cooling than either available MSU analysis. This probably also indicates artifacts in the stratospheric sonde data that have not successfully been removed. Because there is no evidence of significant adjustment biases, we conclude that undetected artifacts are probably behind these evident failures. Such artifacts might take the form of spurious drifts, which seem to occur at some tropical stations (Randel and Wu 2006) and are very hard to distinguish from natural variations. We are not sure what to recommend to ameliorate this problem. The use of winds, however, seems to be promising as an additional source of information to constrain natural variability and thereby improve the detection of unnatural variations in the dataset; future efforts should consider employing dynamic constraints such as geostrophy, which might improve on the purely statistical approach here.

Despite this, the adjusted tropospheric temperature trends agree roughly with physical expectations. We find, in particular, tropical (30°N–30°S) warming that is as fast since 1979 as during the longer 1959–2005 period and that increases with altitude from 850 through about 300 hPa. Extratropical warming was about the same in both hemispheres since 1959, but was strongly asymmetric since 1979, with stronger warming in the ENH and little in the ESH. Previous homogenized datasets generally showed some, but not all, of these characteristics.

The structural uncertainty in our trends, quantified here by taking half the full range of results at different stages of a multistage analysis, with two different changepoint detection schemes, is >0.05°C decade^{-1} in the tropical troposphere and >0.1°C decade^{-1} for the stratosphere and the Southern Hemisphere extratropics. Our 1979–2005 trends for 850–300 hPa in the tropics are $0.15° \pm 0.07$°C decade^{-1}. This is within uncertainty of the roughly 0.17°–0.22° expected on the basis of surface trends of 0.12°–0.14°C decade^{-1} (CCSP 2006; Santer et al. 2005), and the agreement would improve if one were to remove the deep tropical stations whose behavior is inconsistent with the rest of the network. This reinforces similar previous findings of consistent trends (Fu et al. 2004; Mears and Wentz 2005; Sherwood et al. 2005; Vinnikov et al. 2006) but remains unsatisfying until errors are further reduced. Our homogenized data and homogenization parameters are available online (http://earth.geology.yale.edu/sherwood/radproj/).

Acknowledgments. The efforts of Peter Thorne, Holly Titchner, and Mark McCarthy at the Met Office's Hadley Centre to set up tests of IUK were critical in helping to understand the performance of the method. We also thank John Lanzante for his educational input and helpful comments on the manuscript. RSS and UAH MSU data and static weighting functions were provided by RSS with help from Carl Mears and the support of the NOAA/Climate and Global Change Program (CGCP). UMd MSU data were obtained from the

UMd Web site courtesy of K. Vinnikov. We acknowledge the modeling groups—the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM)—for their roles in making available the WCRP CMIP3 multimodel dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy. The work was funded by NOAA CGCP NA03OAR4310153 and NSF ATM-0134893. Holly Titchner was supported by the Joint Defra and MoD Programme, (Defra) GA01101 (MoD) CBC/2B/0417_Annex C5.

APPENDIX A

Detection Scheme Details

To detect changepoints in a time series for a given station and level, we removed outliers (points differing from the median by more than six pseudo–standard deviations), and then binned the data into monthly means (detection on daily data yielded what was subjectively judged to be an excessive number of detections). Months with fewer than 10 observations were denoted missing, and levels with fewer than 30 good months were discarded (this mainly occurred in the stratosphere). These series were then deseasonalized by fitting and removing annual and semiannual sinusoids, before running a changepoint algorithm. Detected changepoints were assigned to either the first of the month after the change, or midmonth if the detection coincided with a missing month. Detections were not allowed in gaps of more than 1 month for T and **S**, because we expected this to produce many false detections (although that was not tested); this eliminated about 4% of the detections. This requirement was waived for dT because this variable was expected to have little serial correlation.

We employed the following two iterative detection schemes: that of L96, based on the nonparametric Mann–Wilcoxon–Whitney test, and the restricted two-phase regression method (Wang 2003). Both find multiple changepoints by iteratively testing homogeneous segments until no segment shows evidence of a significant level shift. L96 reported favorable performance compared to previous methods, including that of Easterling and Peterson (1995), which was based on two-phase regression. Issues related to the changepoint detection scheme were investigated in more depth by S07, who found somewhat better detection by the two-phase method, but similar trend estimation properties of both schemes for cases relevant to the present study. That study also recommended a liberal significance threshold of 0.99, which is adopted here for both schemes.

A key point made by S07 was that, especially when using L96 and especially with liberal significance settings, it is essential to vet the detected changepoints (as suggested by L96) with a further test that requires that the variability near the changepoint more closely resemble a step change (i.e., artifact) than a linear trend (i.e., genuine change). This was quantified via the residual variance produced by two appropriate regressions of data within 30 months on either side of the point being tested. We applied this test to the T and S_x series, which each possess significant natural variability that could cause false changepoint detections, but not the dT series, because of its presumed deficit of coherent natural variability. About 8% of the detected T changepoints subjected to this test were thus rejected.

APPENDIX B

Changepoint Aggregation Procedure

To reduce false detections and/or multiple detections of the same event at different times, and to avoid detections too close together given the ability of the available data to distinguish their effects, we performed the following "aggregation procedure" to boil a set of changepoints detected at individual pressure levels (LCPs) down to a smaller set of consensus CP times for each station:

1) We placed LCPs from all levels in chronological order.
2) We used the first LCP to initiate a "cluster," and then stepped through the remaining LCPs, calculating the time t elapsed since the previous LCP for each. For $t < 6$ months, the LCP was assigned to the previous cluster; otherwise, it initiated a new cluster. Upon completion, every LCP belonged to a cluster and every cluster contained at least one LCP.
3) If a cluster's LCPs spanned more than 3 yr, the LCP farthest from the median date was permanently discarded. This was repeated until the LCPs in the cluster spanned less than 3 yr.
4) A summary changepoint (SCP) was assigned to each cluster, with a date equal to the first day of the month of the median of its LCP times. If the number of LCPs in the cluster was at least equal to a threshold N_{sig}, the SCP was judged highly significant and was assumed potentially to have affected all levels. In this case we call it a "consensus changepoint," or simply a CP.

Tests described in section 4a were repeated with different values of N_{sig}, with apparent performance plateauing at $N_{\mathrm{sig}} = 4$. This conservative value also ensures that

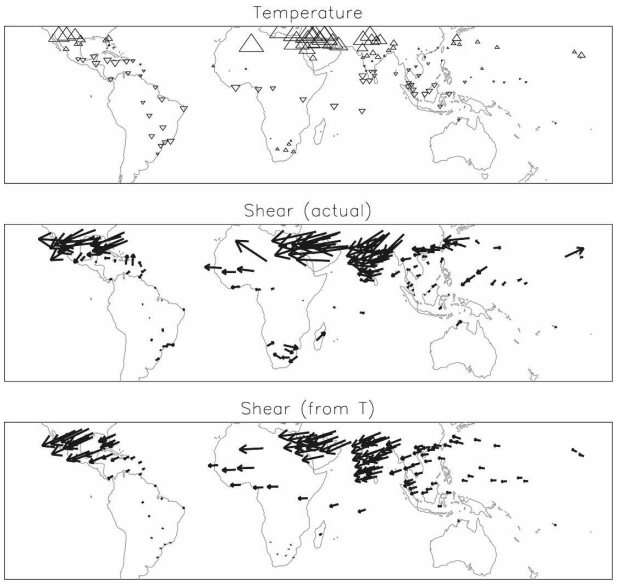

FIG. C1. (a) Temperature, (b) vertical shear, and (c) estimated thermal wind vertical shear loadings of the leading mode for JJA in the tropics. Triangle sizes are proportional to the temperature magnitude and point up or down for positive or negative loading, respectively; arrow lengths are proportional to shear $\times f$, where f is the Coriolis parameter.

at least one LCP must lie in both the troposphere and stratosphere. Accordingly, this was chosen for the T homogenization. Note that because dT is assumed to have no actual trend, it was homogenized (round 1) by adjusting every LCP without performing the aggregation.

APPENDIX C

Estimating the Natural Variability Basis for IUK

The function $\mathbf{g}$ is obtained here by principal component (EOF) analysis of the currently estimated natural variability $\boldsymbol{\mu}_1 + \boldsymbol{\epsilon}$, including imputed missing values. Issues associated with this choice are discussed by S07. Tests indicated insensitivity of results to the number of iterations N_g in which the EOFs were recomputed as long as this exceeded two, so we set $N_g = 7$. S07 found occasions where results deteriorated unless each station's own data were excluded from its EOFs, but sensitivity tests here indicated that this was unimportant (probably because of the large number of stations), so this was not done.

An issue that did not arise in S07 is how to combine wind and temperature data. We simply rescaled all wind shear data at a given season and level to give S_x the same variance as T. This typically gave S_y less variance, but gave the winds together more variance than T. Results were not sensitive to modest changes in this scaling factor. Data were smoothed in time with a 13-element window, as in S00, to reduce the influence of synoptic fluctuations on the modes. In accord with tests described below, we decided to truncate at six modes in the tropics (as chosen also by S00), three in ESH, and nine in ENH, when analyzing group A; for groups A and B we used nine in the tropics.

These numbers were subjectively determined by first running test cases at 300 hPa with 10 EOFs and examining the modes for physical plausibility. In particular, it is expected that the EOF loadings (T, S_x, and S_y) will be geostrophically balanced. This was roughly true for the leading modes at each season and level; Fig. C1 shows one example (JJA in the tropics), in which the wind shear loadings resemble those obtained from the thermal wind equation by numerically estimating the gradient of the temperature field. The correspondence is not expected to be exact, because of noise in the data. This problem tends to grow worse as one goes down through the modes, forming the basis for the truncation decision.

REFERENCES

Allen, R. J., and S. C. Sherwood, 2007: Utility of radiosonde wind data in representing climatological variations of tropospheric temperature and baroclinicity in the western tropical Pacific. *J. Climate,* **20,** 5229–5243.

——, and ——, 2008: Warming maximum in the tropical upper troposphere deduced from thermal wind observations. *Nature Geosci.,* **1,** 399–403.

Angell, J. K., and J. Korshover, 1975: Estimate of global change in tropospheric temperature between 1958 and 1973. *Mon. Wea. Rev.,* **103,** 1007–1012.

CCSP, 2006: Temperature trends in the lower atmosphere: Steps for understanding and reconciling differences. NOAA/NCDC, Climate Change Science Program and the Subcommittee on Global Change Research Rep. 1.1, 164 pp.

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.

——, W. B. Norris, R. W. Spencer, and J. J. Hnilo, 2007: Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements. *J. Geophys. Res.,* **112,** D06102, doi:10.1029/2005JD006881.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

Durre, I., R. S. Vose, and D. B. Wuertz, 2006: Overview of the Integrated Global Radiosonde Archive. *J. Climate,* **19,** 53–68.

Easterling, D. R., and T. C. Peterson, 1995: A new method for detecting undocumented discontinuities in climatological time-series. *Int. J. Climatol.,* **15,** 369–377.

Free, M., and Coauthors, 2002: Creating climate reference datasets: CARDS workshop on adjusting radiosonde temperature data for climate monitoring. *Bull. Amer. Meteor. Soc.,* **83,** 891–899.

——, D. J. Seidel, J. K. Angell, J. Lanzante, I. Durre, and T. C. Peterson, 2005: Radiosonde Atmospheric Temperature Prod-

ucts for Assessing Climate (RATPAC): A new data set of large-area anomaly time series. *J. Geophys. Res.,,* **110,** D22101, doi:10.1029/2005JD006169.

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., M. A. Sargent, R. E. Habermann, and J. R. Lanzante, 2000: Sensitivity of tropospheric and stratospheric temperature trends to radiosonde data quality. *J. Climate,* **13,** 1776–1796.

Gruber, C., and L. Haimberger, 2008: On the homogeneity of radiosonde wind time series. *Meteor. Z.,* in press.

Haimberger, L., 2005: Homogenization of radiosonde temperature time series using ERA-40 analysis feedback information. ECMWF Technical Report, ERA-40 Project Report Series 23, 68 pp.

——, 2007: Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate,* **20,** 1377–1403.

——, C. Tavolato, and S. Sperka, 2008: Toward elimination of the warm bias in historic radiosonde temperature records—Some new results from a comprehensive intercomparison of upper air data. *J. Climate,* **21,** 4587–4606.

Kiladis, G. N., K. H. Straub, G. C. Reid, and K. S. Gage, 2001: Aspects of interannual and intraseasonal variability of the tropopause and lower stratosphere. *Quart. J. Roy. Meteor. Soc.,* **127,** 1961–1983.

Lanzante, J. R., 1996: Resistant, robust and nonparametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.,* **16,** 1197–1226.

——, S. A. Klein, and D. J. Seidel, 2003: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. *J. Climate,* **16,** 224–240.

McCarthy, M. P., H. A. Titchner, P. W. Thorne, S. F. Tett, L. Haimberger, and D. E. Parker, 2008: Assessing bias and uncertainty in the HadAT adjusted radiosonde climate record. *J. Climate,* **21,** 817–832.

Mears, C. A., and F. J. Wentz, 2005: The effect of diurnal correction on satellite-derived lower tropospheric temperature. *Science,* **309,** 1548–1551.

——, M. C. Schabel, and F. J. Wentz, 2003: A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Climate,* **16,** 3650–3664.

National Research Council, 2000: *Reconciling Observations of Global Temperature Change.* National Academy Press, 85 pp.

Parker, D. E., and D. I. Cox, 1995: Towards a consistent global climatological rawinsonde database. *Inter. J. Climatol.,* **15,** 473–496.

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

Riehl, H., 1954: *Tropical Meteorology.* McGraw Hill, 392 pp.

Santer, B. D., and Coauthors, 2005: Amplification of surface temperature trends and variability in the tropical atmosphere. *Science,* **309,** 1551–1556.

Sherwood, S. C., 2000: Climate signals from station arrays with missing data, and an application to winds. *J. Geophys. Res.,* **105,** 29 489–29 500.

——, 2007: Simultaneous detection of climate change and observing biases in a network with incomplete sampling. *J. Climate,* **20,** 4047–4062.

——, J. R. Lanzante, and C. L. Meyer, 2005: Radiosonde daytime biases and late-20th century warming. *Science,* **309,** 1556–1559.

Thompson, L. G., and Coauthors, 2006: Abrupt tropical climate change: Past and present. *Proc. Natl. Acad. Sci. USA,* **103,** 10 536–10 543.

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.

Titchner, H. A., P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, and D. E. Parker, 2008: Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments. *J. Climate,* in press.

Vinnikov, K. Y., N. C. Grody, A. Robock, R. J. Stouffer, P. D. Jones, and M. D. Goldberg, 2006: Temperature trends at the surface and in the troposphere. *J. Geophys. Res.,* **111,** D03106, doi:10.1029/2005JD006392.

Wang, X. L., 2003: Comments on "Detection of undocumented changepoints: A revision of the two-phase regression model." *J. Climate,* **16,** 3383–3385.

Wu, W., A. E. Dessler, and G. R. North, 2006: Analysis of the correlations between atmospheric boundary-layer and free-tropospheric temperatures in the tropics. *Geophys. Res. Lett.,* **33,** L20707, doi:10.1029/2006GL026708.