Online Supplement

Methods

As described in main text, each ΔT sample is the difference between two consecutive soundings. We discarded ΔT values more than six pseudo-standard deviations from the median (*1*), which eliminated fewer than 0.5% of observations at any station. For the 1979-97 time period, least-squares linear trends were calculated at each station that had at least 30 ΔT observations in each of the first and last thirds of that time period (for the 1959-97 period, the first and last fourths). The uncertainty of a station's trend was calculated assuming one half of a degree of freedom per observation used from that station (not one per ΔT sample, which would be an overestimate (*2*)). Autocorrelation analysis confirmed that ΔT sample values separated by 24 hours were essentially uncorrelated (see below), so no adjustment was needed for serial correlation. Trend values having uncertainty $> 0.5\,^{\circ}$C per decade were discarded, though this only occurred at a few station-levels.

The median uncertainty (1-σ) among all stations was near 0.05-0.06 K/decade at all levels. This is much smaller than the typical linear trend uncertainty in an individual stations's temperature itself, due to the greater variability of the latter and its significant temporal coherence (*3*). Consequently, we were able to use data from 21 tropical LKS (*4*) stations (18 after excluding Indian and those within 10° of 90E or 90W) for 1979-97 trends, and 11 (10) for 1959-97, even though LKS included nighttime trends from only eight at all longitudes. Many stations report sporadically, especially at night and in the Tropics. Since our sampling technique strongly filters out variability on time scales longer than one day, including the dominant time scale of ~ 3 days that arises due to the passage of synoptic waves, as well as climate variability, we were able to detect much smaller changes in day-night temperature difference at these stations than would be possible by examining differences between (e.g.) monthly means of daytime and nighttime data. This is because the 0000 and 1200 UTC monthly means will generally come from soundings on different days that sampled different weather.

The day-night difference quantity $\Delta T'$ was computed according to the rule

$$\begin{aligned} \Delta T' &\equiv -\Delta T \qquad (\text{longitudes} < 90) \\ \Delta T' &\equiv +\Delta T \qquad (\text{longitudes} > 90) \end{aligned}$$

so that $\Delta T'$ is always equal to (6 am through 6 pm) minus (6 pm through 6 am). Though we take this as the "day minus night" difference, this will not always be true except on the equator or at equinox. At stations near 90E and 90W, launches at either observing time will include some after and some before sunrise. This problem is exacerbated by the fact that launch times can be anytime within three hours of the nominal launch time, and it takes roughly an hour or more for the sonde to complete its ascent. We try to minimize this problem by discarding all stations within 10° of these meridians in making averages of the trend in $\Delta T'$. This is not really sufficient, especially outside the deep Tropics, due to seasonal variations in the length of day, but tests with more aggressive data screening indicate that the resulting errors are modest compared to other uncertainties. Nonetheless, it is likely that effects of solar heating may be underestimated in parts of the world such as India, the eastern US, western China, and central

Russia where significant numbers of daylight observations can occur at both launch times. A strong east-west gradient in the effect appears among Indian stations, which is consistent with an increase toward the west of the country in the degree of day-night separation between the synoptic observing times (though, as noted, we do not actually use these stations, this pattern is nonetheless instructive).

In making belt-average trends $\langle \Delta T' \rangle$, we simply averaged the trends at all qualifying stations with equal weight. This reproduces the impact that the error would have on the daytime data in a climatology in which all stations were averaged equally. To calculate the impact $\delta_{\rm sol}$ on the all-time-of-day trend, we used

$$\begin{aligned} \delta_{\rm sol} &= -f \frac{d\langle \Delta T' \rangle}{dt} \\ f &= \frac{2N_{\rm Day} + N_{\rm All}}{2(N_{\rm Day} + N_{\rm All})} \end{aligned} \tag{1}$$

where $N_{\rm Day}$ and $N_{\rm All}$ are the numbers of these stations where only daytime data, and where both day and night data were used, respectively, to calculate trends. This holds for work such as LKS where separate trends were calculated for each available synoptic time and then averaged with equal weight. For other ways of combining data from different observing times, other formulas would be needed to estimate f.

In computing the tropical time series (Fig. 3), only the 10 stations away from the 90° meridians and providing good trend data over the 1959-97 period were used. Since the CARDS dataset ends in 2000, we extended each station's time series through 2004 using data from the IGRA archive. Based on cursory inspection, the IGRA and CARDS data appear consistent at the stations examined. Monthly means of $\Delta T'$ were calculated at each station, then an average was made for each month from whatever stations were available that month; this number fell as low as three for a few months in the early 1960's, but remained at least five after 1965 and was at least eight for most months. Some of the rapid variations, particularly the sharp change near the beginning of the time series, may be artifacts of changes in the mix of stations available, each of which has a different mean value of $\Delta T'$. The downward trend after 1962, however, appears to be due to a combination of secular variations at some stations and large step-like changes at others, all occurring at different times so as to produce a fairly steady decrease in the mean value. This decrease is close to that calculated (Fig. 4) by the more robust method of averaging the trend from each station with equal weight. The uncertainty of this curve may be readily assessed from the "noise" in the graph, since each monthly mean is completely independent. Similar behavior is seen for the SH and NH time series (not shown).

Supporting Text

A key question in interpreting these results is whether previous efforts to "homogenize" the radiosonde record (detect and remove spurious changes by inspecting the data) have already found and taken account of the problems reported here. These efforts (cited in the main text; see also (*5*)) have, to our knowledge, all concluded that the net effect of all detectable errors is significantly smaller in the troposphere than what we obtain here from just the solar heating error. On the other hand, these studies have obtained corrections in the stratosphere that are

much closer to those reported here. Why the difference in the troposphere? Either (a) we are exaggerating the tropospheric effect somehow, (b) previous efforts have found it but also found other errors that cancel most of its impact on trends, or (c) previous efforts either missed most of this problem or made spurious corrections elsewhere that canceled it out. Possibility (a) is refuted through the various uncertainty figures quoted in the text, and the robustness of the effect across so many stations. Although the magnitude of the impact is uncertain, the uncertainty does not accomodate a magnitude as small as previous corrections. Possibility (b) cannot be ruled out, but it requires that the other errors have a systematic trend throughout the global record and that this nearly cancel the one documented and explained here in the troposphere—but not in the stratosphere, where corrections due to solar heating and/or other effects are acknowledged to be too *small* (*5*). Large errors have been reported in isolated cases due to changes in instrument manufacturer (e.g., VIZ to Vaisala), analysis software, operator procedure, and numerous other details. It would be quite fortuitous, however, for the cumulative effect of these worldwide to meet the above conditions. Since these other types of error are independent of the solar heating trend we must, absent definite quantitative knowledge, add them as independent sources of uncertainty. This only reinforces our basic conclusion that the total uncertainty in radiosonde trends equals or exceeds the trend expected in the troposphere on the basis of reported surface observations.

Though (b) seems unlikely, (c) is plausible in our judgment given the difficulty of the homogenization process and the poor sampling, particularly in the Tropics. LKS found that day-night difference information was the most useful way of identifying artifacts (*4*), and the main secondary strategy involves buddy checks (*9*). Yet many stations have little or no nighttime data, few or no nearby neighbors for buddy checks, and/or sporadic reporting that significantly increases sampling errors in monthly mean temperature. Most tropical stations suffer from at least one of these problems.

Several statistical issues complicate the effort to internally estimate artifacts, even when sampling is good. First, simple tests show that discontinuities cannot be identified reliably in a time series with a few dozen degrees of freedom (as is roughly the case for these radiosonde temperature series) unless they are of order 1σ of the time series (the record in question minus any available "reference" time series based on independent data) or larger. Apparent jumps below the safe detection limit will be found by a sufficiently aggressive algorithm even when none actually exist (Type I errors). Because Type I errors will coincide with large natural changes, it only takes a few of these to wreak havoc on trend statistics. A less aggressive strategy may avoid these errors, but will also fail to detect all but the largest artifacts. Second, typical procedures aimed at removing level shifts will alias natural variability or trends onto the estimated shift, with the result that (a) real trends are partly removed (trends biased toward those in the reference series), and (b) accurate corrections tend to be compensated by Type I errors or amplitude errors of opposite sign. Examples of natural fluctuations that could easily lead to a type I error are documented in the climate literature (*4*) (Section 5a on Majuro), as has the tendency for genuine trends to be removed through absorption into the adjustments (*6, 7*). Finally, there is evidence from comparisons with colocated satellite data (*8*) that spurious temperature changes at a station sometimes occur by the accumulation of many small increments that resemble natural variations. We find similar behavior for ΔT changes in some cases.

LKS presented evidence that fluctuations in their adjusted time series agreed better with independent satellite data; large local problems were surely reduced. However, much of this

success probably derived from the large stratospheric adjustments and does not guarantee that small systematic problems in the troposphere were successfully removed. In fact LKS pointed out themselves that they probably did not remove many of the smaller artifacts. This would also affect efforts (*9*) that use LKS as their backbone or "ground truth." A recent data-assimilation-based study (*10*), while an impressive effort that was fully independent of LKS adjustments, still relied implicitly on a background field that would be contaminated by low-level, widespread errors and susceptible to all of the above problems. Unfortunately, the basic limitations of change-point analysis are probably not adequately appreciated in the climate community. The development of unbiased correction methodologies would be an important advancement toward the goal of confident homogenization of the record.

Data files

The online supplement contains two .tar archives, one for each time period considered. Each expands to a folder containing a separate ASCII file for each pressure level. Each such file contains five columns listing, respectively: the station ID, latitude, longitude, trend in ΔT (in K/decade), and one-sigma uncertainty of fit in this trend (also in K/decade).

Supporting References and Notes

1. J. R. Lanzante, *Int. J. Climatology* **16**, 1197 (1996).

2. Each ΔT sample requires two independent observations, providing one degree of freedom (DOF) for ΔT and one for the mean of T (which is also independent for each sample). Since the same observation is frequently used for two ΔT samples, not all the ΔT samples are independent. A perfectly reporting station with N days of complete twice-daily data would provide $2N$ independent temperature observations yeilding $2N$ ΔT samples, but only N DOF.

3. B. D. Santer, *et al.*, *J. Geophys. Res.* **105**, 7337 (2000).

4. J. R. Lanzante, S. A. Klein, D. J. Seidel, *J. Climate* **16**, 224 (2003).

5. J. R. Christy, W. B. Norris, *Geophys. Res. Lett.* **31**, Art. No. L06211 (2004).

6. D. J. Gaffen, M. A. Sargent, R. E. Habermann, J. R. Lanzante, *J. Climate* **13**, 1776 (2000).

7. M. Free, *et al.*, *Bull. Amer. Meteor. Soc.* **83**, 891 (2002)

8. W. J. Randel, F. Wu, *J. Climate* (Submitted manuscript).

9. P. W. Thorne, *et al.*, *J. Geophys. Res.* (In Press).

10. L. Haimberger, Homogenization of radiosonde temperature time series using ERA-40 analysis feedback information, *Tech. rep.*, ECMWF (2005). ERA-40 Project Report Series #23, 68 pp.