Homogenization of the Global Radiosonde Temperature Dataset through Combined Comparison with Reanalysis Background Series and Neighboring Stations

LEOPOLD HAIMBERGER, CHRISTINA TAVOLATO, AND STEFAN SPERKA

Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria

(Manuscript received 16 November 2011, in final form 30 May 2012)

ABSTRACT

This article describes progress in the homogenization of global radiosonde temperatures with updated versions of the Radiosonde Observation Correction Using Reanalyses (RAOBCORE) and Radiosonde Innovation Composite Homogenization (RICH) software packages. These are automated methods to homogenize the global radiosonde temperature dataset back to 1958. The break dates are determined from analysis of time series of differences between radiosonde temperatures (obs) and background forecasts (bg) of climate data assimilation systems used for the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) and the ongoing interim ECMWF Re-Analysis (ERA-Interim).

RAOBCORE uses the obs−bg time series also for estimating the break sizes. RICH determines the break sizes either by comparing the observations of a tested time series with observations of neighboring radiosonde time series (RICH-obs) or by comparing their background departures (RICH-τ). Consequently RAOBCORE results may be influenced by inhomogeneities in the bg, whereas break size estimation with RICH-obs is independent of the bg. The adjustment quality of RICH-obs, on the other hand, may suffer from large interpolation errors at remote stations. RICH-τ is a compromise that substantially reduces interpolation errors at the cost of slight dependence on the bg.

Adjustment uncertainty is estimated by comparing the three methods and also by varying parameters in RICH. The adjusted radiosonde time series are compared with recent temperature datasets based on (Advanced) Microwave Sounding Unit [(A)MSU] radiances. The overall spatiotemporal consistency of the homogenized dataset has improved compared to earlier versions, particularly in the presatellite era. Vertical profiles of temperature trends are more consistent with satellite data as well.

1. Introduction

The radiosonde network is a central part of the global upper air observing system. Since many stations have operated since the late 1950s or longer, radiosonde temperature records have been extensively used also for climate studies. It became clear that these records need to be homogenized before their trends and low-frequency variability can be interpreted (Parker et al. 1997).Various homogenization approaches for the global radiosonde network have been put forward (Luers and Eskridge 1995; Lanzante et al. 2003; Thorne et al. 2005b; Free et al. 2005; Sherwood et al. 2008; McCarthy et al. 2008; Titchner et al. 2009) but none of them could explain and remove the apparent pervasive cooling bias in the radiosonde temperature data compared to satellite data. In particular, the vertical trend profiles in the tropics did not show the enhanced upper tropospheric amplification as predicted by climate models, with the exception of the ensemble described in Thorne et al. (2011a) that included this possibility in its uncertainty bounds. Only temperature trend assessments based on changes in the zonal mean thermal wind structure support enhanced upper tropospheric warming in the tropics (Allen and Sherwood 2008).

Discrepancies between layer mean atmospheric temperatures derived from radiances of the Microwave Sounding Unit (MSU) and MSU-equivalent temperatures calculated from radiosonde data can be attributed at least partly to inhomogeneities in the raw radiosonde data. Any remaining inconsistency would be caused by inhomogeneities and uncertainties in the MSU brightness temperatures as described, for example, by Thorne

Denotes Open Access content.

Corresponding author address: Leopold Haimberger, Department of Meteorology and Geophysics, University of Vienna, Althanstrasse 14, A-1090 Vienna, Austria.
E-mail: leopold.haimberger@univie.ac.at

DOI: 10.1175/JCLI-D-11-00668.1

et al. (2011b) or Mears et al. (2011). Temperature trends from raw radiosonde data are also inconsistent with climate models, which project an upper tropospheric warming maximum, especially in the tropics (Santer et al. 2005; Trenberth et al. 2007; Santer et al. 2008).

Haimberger (2007) introduced a new homogenization method [Radiosonde Observation Correction Using Reanalyses (RAOBCORE)] that analyzed not only 0000 − 1200 UTC difference time series and station history information but also time series of background departures from the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005). These background departures, also referred to as innovations, are the differences between observations $\mathbf{y}$ and the state vector $\mathbf{x}_b$ of the background forecast by an assimilating model. Here $\mathbf{x}_b$ is mapped to the observation location by the observation operator H. Following the notation in data assimilation literature (see, e.g., Courtier et al. 1998; Lewis et al. 2005) we write $\tau = \mathbf{y} - H\mathbf{x}_b$. The background state $H\mathbf{x}_b$ is considered independent of the radiosonde observations $\mathbf{y}$, which is generally a good assumption except that persistent common biases at several neighboring radiosonde stations may have a noticeable effect on the background state (Haimberger 2007).

Analysis of daily innovation time series proved highly efficient for break detection, but break size estimation from reanalysis innovations is complex because of a few inhomogeneities of the ERA-40 background forecast and analysis time series (Uppala et al. 2005; Haimberger 2007; Grant et al. 2008; Screen and Simmonds 2011). RAOBCORE, which uses innovation series for both break detection and innovation, has also been criticized for being not independent of satellite data and of the assumptions made in the assimilating model. This aspect certainly limits the value of radiosonde data homogenized by RAOBCORE for comparison with satellite datasets. Nevertheless, RAOBCORE adjustments have been used as radiosonde bias correction in the interim ECMWF Re-Analysis (ERA-Interim; Dee et al. 2011b; Haimberger and Andrae 2011), in the Modern-Era Retrospective Analysis for Research and Applications (MERRA; Rienecker et al. 2011), and in the Japanese 55-yr reanalysis (Ebita et al. 2011).

To overcome the dependence problem, Haimberger et al. (2008) developed a method called Radiosonde Innovation Composite Homogenization (RICH), which uses the breakpoint date information from RAOBCORE but calculates the break size estimates by comparison with neighboring radiosonde temperature records. Since these reference records are independent of satellite data, satellite data can affect RICH estimates only through the breakpoint dates provided by RAOBCORE. RICH worked quite well for homogenizing the time series of the satellite period 1979 onward. Also it revealed an upper tropospheric warming maximum in the much debated period 1979–99 (Santer et al. 2008).

The present paper is motivated by the fact that Haimberger et al. (2008) provided only rudimentary documentation of RICH, by the availability of very much enhanced background forecasts due to the advent of ERA-Interim (Dee and Uppala 2009; Dee et al. 2011a,b), and by substantial extensions and improvements made to the RICH algorithm itself. Another motivation was the desire to quantify the uncertainties in the homogenization approach through the use of ensemble techniques. This has been pioneered by McCarthy et al. (2008) and Thorne et al. (2011a) for radiosonde temperatures. Ensembles of MSU brightness temperatures have become available recently as well (Mears et al. 2011) and are also employed for assessing uncertainties in SST data (Kennedy et al. 2011a,b). They allow for much better assessment of whether differences between various datasets are statistically significant.

The next section describes the RICH method and its parameters in some detail. Section 3 describes input data for RICH as well as the datasets used for comparison and validation. Section 4 explains the methodology used to quantify parametric and to a certain extent also structural adjustment uncertainties. Section 5 shows results for selected stations and large-scale means. Implications of the results for a general climate data improvement strategy are discussed in the conclusions.

2. Description of the RICH adjustment method

The basic idea of RICH—homogenization of tested time series through comparison with neighboring reference series of the same observation type—is not new. A large variety of upper air data and surface data homogenization methods work with this idea (see, e.g., Thorne et al. 2005b; Sherwood et al. 2005; Venema et al. 2012).

The main novelty of RICH is that it tries to make optimal use of the output from a dynamical climate data assimilation system for break detection and adjustments. For this purpose it uses the break detection part of the RAOBCORE algorithm, which analyzes 0000 and 1200 UTC daily time series of background departures $\tau = \mathbf{y} - H\mathbf{x}_b$ from reanalyses.

For radiosonde temperature measurements H is specified as a simple interpolation from the ECMWF model grid to the observation location. The innovations are an important standard diagnostic for time series models or data assimilation systems. Haimberger (2007) demonstrated that statistical analysis of daily innovation time series with homogeneity tests is quite efficient in

finding breaks in these time series because of their small variance. The RAOBCORE algorithm, which does this analysis, has not changed appreciably since publication in 2007. However, the background forecast data used as reference have improved (see section 3).

RAOBCORE yields the date of potential breaks in the radiosonde time series. These dates are input for the Radiosonde Innovation Composite Homogenization algorithm, where the word ''innovation'' in the acronym should remind the user that the used break dates have been calculated by analyzing innovation time series.

a. *Definition of averaging operators and break size estimates*

For break estimation, RICH compares either neighboring observation (obs) time series (RICH-obs) or neighboring τ time series (RICH-τ). The obs time series are independent of satellite or other nonradiosonde observing systems. RICH-obs assumes that a sufficiently long homogeneous reference time series can be constructed from neighboring radiosonde temperature time series for each break date in a tested station time series so that the size of the shift can be accurately estimated.

While the essence of break size estimation is rather simple (comparing means), we formally define the estimates to make the details of the break size estimation process transparent. In this paper we denote differences between tested and reference stations as follows:

$$ {}^{i}\nabla_{x}^{jk} = {}^{i}x^{j} - {}^{i}x^{k}, \tag{1}$$

where ${}^{i}x^{j}$ denotes either obs or bg temperature values or their difference τ = obs − bg at station j, and i is the launch index that stands for the date and time when the x values were calculated or measured. Also, ${}^{i}\nabla_{x}^{jk}$ denotes the difference of x values between stations j and k at the same observation time i. Since the stations j and k are spatially separated, it seems suitable to use the gradient symbol for this difference.

In general we do not want to compare individual values but averages over a time period. We define the time average $\overline{x^{j}}(a)$:

$$ \frac{1}{n}\sum_{i=1}^{n} {}^{i}x^{j} = \overline{x^{j}}(a). \tag{2}$$

The choice of the interval a depends on the dates of other breaks in the tested or the reference series as well as on data availability. Its length is 0.5–8 yr. The sample size n is the number of launches (130–2920) at a specific time of day (0000 or 1200 UTC); n depends on the length of a and data availability at station j in this interval. While smaller sample sizes are possible they have not been found advantageous. Details for the choice of the intervals used for break size estimation are given in section 2b below. Note also that $\overline{x^{j}}(a)$ is seasonally invariant, as are all means below. Therefore seasonal variations of biases, especially in polar regions, are not taken into account. Andrae et al. (2004) and Haimberger and Andrae (2011) discuss methods to estimate the radiation error as a function of solar elevation. They can complement the seasonally invariant adjustments calculated in this paper but this is not pursued further here.

1) Mean differences of observations and background

The observation difference is the mean temperature difference between observations from two neighboring radiosonde stations and is defined as

$$ \overline{\nabla_{\mathrm{obs}}^{jk}}(a) = \overline{\mathrm{obs}^{j}}(a) - \overline{\mathrm{obs}^{k}}(a), \tag{3}$$

where obs are the measured temperatures at stations j and k in the interval a, which contains n pairs of observations; $\overline{\mathrm{obs}^{j}}(a)$ and $\overline{\mathrm{obs}^{k}}(a)$ are mean values estimated at stations j and k in interval a. It is important that only those data are counted where values at both stations j and k are available. This reduces the risk of unrealistic differences due to unequal sampling at stations j and k. The same quantity can be defined for the bg:

$$ \overline{\nabla_{\mathrm{bg}}^{jk}}(a) = \overline{\mathrm{bg}^{j}}(a) - \overline{\mathrm{bg}^{k}}(a). \tag{4}$$

Let us now consider the situation where an interval a with n numbers of observation pairs and an interval b with m numbers of observation pairs are separated by a break in the time series of station k. Then we expect that the observation difference to station j in the two intervals is different: $\overline{\nabla_{\mathrm{obs}}^{jk}}(a) \neq \overline{\nabla_{\mathrm{obs}}^{jk}}(b)$. The size of the change between intervals a and b can be written as

$$\begin{aligned} \Delta_{\mathrm{obs}}^{jk}(a,b) &= [\overline{\mathrm{obs}^{j}}(b) - \overline{\mathrm{obs}^{k}}(b)] - [\overline{\mathrm{obs}^{j}}(a) - \overline{\mathrm{obs}^{k}}(a)] \\ &= \overline{\nabla_{\mathrm{obs}}^{jk}}(b) - \overline{\nabla_{\mathrm{obs}}^{jk}}(a). \end{aligned} \tag{5}$$

If the stations j and k are close to each other, and if the reference time series at station j is homogeneous, $\Delta_{\mathrm{obs}}^{jk}(a,b)$ already represents an unbiased estimate for the size of the break at station k occurring between intervals a and b. As explained in section 2b below, RICH-obs uses weighted means of $\Delta_{\mathrm{obs}}^{jk}$ from several neighboring stations for estimating the break sizes at station k. The black profiles in the left panels of Fig. 1 are $\Delta_{\mathrm{obs}}^{jk}(a,b)$ estimates for different reference stations j and different pressure levels.

The situation becomes more complicated, however, if the distance between stations is large. In this case the true

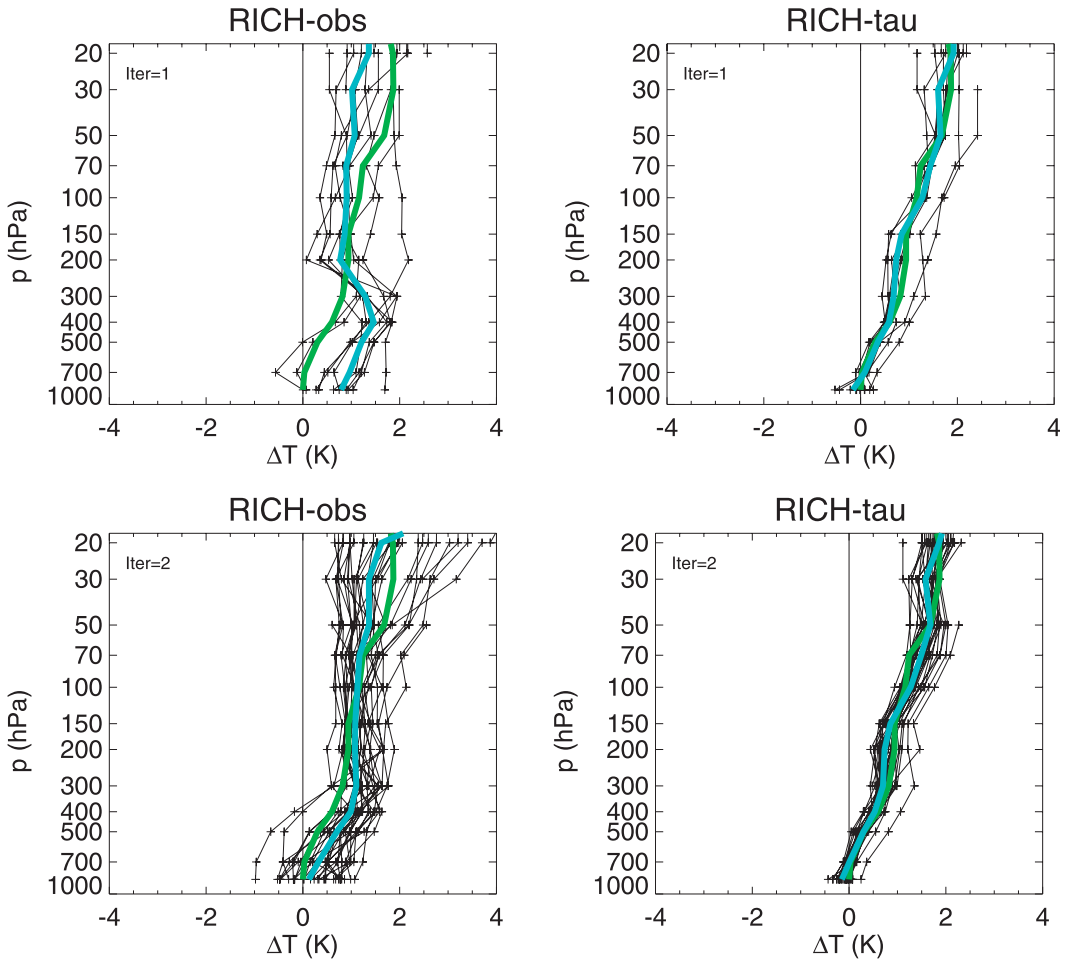

FIG. 1. RICH break size estimation: black curves are individual (left) Δ^{jk}_{obs} and (right) Δ^{jk}_{τ} profiles for a break in the radiosonde record of Bethel (Alaska, 70219) in July 1989. Thick green profile (identical in all four panels) is the RAOBCORE estimate, thick blue profiles are (left) RICH-obs and (right) RICH-τ estimates, respectively. (top) Profiles from first iteration of RICH (see Fig. 3; 10 neighbors used); (bottom) profiles from second iteration with 30 neighbors.

mean temperatures at stations j and k may evolve differently due to regional climate anomalies. In such situations $\Delta^{jk}_{\text{obs}}(a,b)$ could be different from zero even if both records at stations j and k were free of artificial jumps. Interpreting $\Delta^{jk}_{\text{obs}}(a,b)$ as break size estimate would lead to false adjustments in this case and must be avoided.

The problem can be circumvented if the short-term climate anomalies can be realistically represented by an independent dataset. ERA-40/ERA-Interim background fields (bg) are generally, with known exceptions in polar regions (e.g., Grant et al. 2008), of sufficient quality that they can represent regional climate anomalies well. If this is true, the difference Δ^{jk}_{bg}, defined as

$$\Delta^{jk}_{\text{bg}}(a,b) = [\overline{\text{bg}^j}(b) - \overline{\text{bg}^k}(b)] - [\overline{\text{bg}^j}(a) - \overline{\text{bg}^k}(a)]$$
$$= \overline{\nabla^{jk}_{\text{bg}}}(b) - \overline{\nabla^{jk}_{\text{bg}}}(a), \qquad (6)$$

describes a climate change in the temperature gradient between stations j and k, which needs to be subtracted from the break size estimate $\Delta^{jk}_{\text{obs}}(a,b)$ gained from comparison of the observations.

2) INNOVATION DIFFERENCES BETWEEN INTERVALS AND STATIONS

If one considers the bg temperature gradient to be a true estimate of the state of the atmosphere, the innovations, or obs − bg differences must contain information about the systematic difference of the radiosondes that were used. For a single station (k) and one interval (a) the obs − bg difference is defined as

$$\overline{\tau^k}(a) = \overline{\text{obs}^k}(a) - \overline{\text{bg}^k}(a). \qquad (7)$$

If the background provides a spatially and temporally homogeneous field, a change in the systematic bias of radiosonde data, caused by a change of observation systems, must change $\tau^k(a)$. This value can be estimated as

$$\Delta_\tau^k(a,b) = [\overline{\text{obs}^k}(a) - \overline{\text{bg}^k}(a)] - [\overline{\text{obs}^k}(b) - \overline{\text{bg}^k}(b)] = \overline{\tau^k}(a) - \overline{\tau^k}(b). \quad (8)$$

This estimate is used to calculate the adjustments performed by RAOBCORE (Haimberger 2007). As one can see, no reference station j is necessary to get a break size estimate in this case. However, evaluation of Eq. (8) puts high demands on the homogeneity of the bg. The green curve in all panels of Fig. 1 is the profile of $\Delta_\tau^k(a,b)$ for a specific break in the time series of Bethel, Alaska.

The RICH-τ method makes different use of the bg. It requires only that the gradient of the bg between stations j and k be realistic. The absolute value of the bg does not need to be unbiased or at least homogeneous, as is required for RAOBCORE. It uses the information provided by both the background and the time series of a neighboring station for break size estimation. An innovation difference $\nabla_\tau^{jk}(a)$ is defined as

$$\overline{\nabla_\tau^{jk}}(a) = [\overline{\text{obs}^j}(a) - \overline{\text{bg}^j}(a)] - [\overline{\text{obs}^k}(a) - \overline{\text{bg}^k}(a)] = \overline{\tau^j}(a) - \overline{\tau^k}(a). \quad (9)$$

It thus combines observations and background information of two neighboring radiosondes, and is a modification of $\overline{\nabla_{\text{obs}}^{jk}}$. If only station k changes its upper air observation system, and the background provides a homogeneous field, the difference of $\overline{\nabla_\tau^{jk}}$ before and after the artificial break must contain information about the systematic error in the observations:

$$\Delta_\tau^{jk}(a,b) = [\overline{\text{obs}^j}(b) - \overline{\text{bg}^j}(b)] - [\overline{\text{obs}^k}(b) - \overline{\text{bg}^k}(b)] - [\overline{\text{obs}^j}(a) - \overline{\text{bg}^j}(a)] + [\overline{\text{obs}^k}(a) - \overline{\text{bg}^k}(a)] = \overline{\nabla_\tau^{jk}}(b) - \overline{\nabla_\tau^{jk}}(a). \quad (10)$$

In contrast to $\Delta_{\text{obs}}^{jk}(a,b)$, this difference takes a possible regional climate anomaly into account, provided the bg gradients are realistic. Using definitions (5) and (6), $\Delta_\tau^{jk}(a,b)$ can also be written as

$$\Delta_\tau^{jk}(a,b) = \Delta_{\text{obs}}^{jk}(a,b) - \Delta_{\text{bg}}^{jk}(a,b). \quad (11)$$

RICH-τ uses weighted means of Δ_τ^{jk} from several neighboring stations to estimate the break sizes at station k. The black profiles in the right panels of Fig. 1 are $\Delta_\tau^{jk}(a,b)$ estimates for different reference stations j and different pressure levels.

3) Interpretation

If the bg is correct at all times and places and the reference radiosonde station time series is homogeneous throughout the combined interval $[a,b]$, the break size is given as the value of Δ_τ^{jk}, which is related to Δ_{obs}^{jk} and Δ_τ^k as

$$\Delta_\tau^{jk}(a,b) = \Delta_{\text{obs}}^{jk}(a,b) - \Delta_{\text{bg}}^{jk}(a,b) = \Delta_\tau^k(a,b) - \Delta_\tau^j(a,b). \quad (12)$$

In the ideal case, if both bg and neighboring radiosonde record j are accurate and homogeneous, Δ_τ^k should be equal to Δ_τ^{jk}. For an ideal reference radiosonde station Δ_τ^j is zero, if the bg is free of artificial shifts, and an optimal bg would have zero τ^j in case of true observations.

If Δ_τ^k and Δ_τ^{jk} are different, however, it is difficult to tell where the discrepancies are coming from. If the background is biased and the reference sonde j is homogeneous, one can speculate that even a biased background would deliver a realistic ∇_{bg}^{jk}, making Δ_τ^{jk} a better estimate for the systematic bias between radiosondes than Δ_τ^k. If the reference record is inhomogeneous, it may be better to use Δ_τ^k for break estimation since this difference is not affected by inhomogeneities in record j.

It is instructive to compare the standard errors of Δ_τ^k, Δ_τ^{jk}, and Δ_{obs}^{jk}. These are calculated for each of these Δ_m as

$$\sigma_m = \sqrt{\frac{1}{n_a + n_b}[n_a \sigma_x(a)^2 + n_b \sigma_x(b)^2]}, \quad (13)$$

where $\sigma_x(a)$ is the standard deviation of

$$\begin{cases} \overline{\tau^k}(a) & \text{for} \quad m = \text{RAOBCORE}, \\ \overline{\text{obs}^k}(a) - \overline{\text{obs}^j}(a) & \text{for} \quad m = \text{RICH-obs}, \\ \overline{\tau^k}(a) - \overline{\tau^j}(a) & \text{for} \quad m = \text{RICH-}\tau. \end{cases}$$

The same definition applies to $\sigma_x(b)$. Note that the size of a possible break between intervals a, b has no effect on the combined standard deviation σ_m. As is shown in Fig. 2, σ_m is smaller for RAOBCORE than for RICH-τ and much smaller than for RICH-obs. The main reason is the smallness of the individual background departures (${}^i\tau^k$) compared to background departure differences (${}^i\tau^k - {}^i\tau^j$) and to temperature differences (${}^i\text{obs}^k - {}^i\text{obs}^j$). A second reason is the smaller sample sizes in RICH due to missing data at individual reference stations.

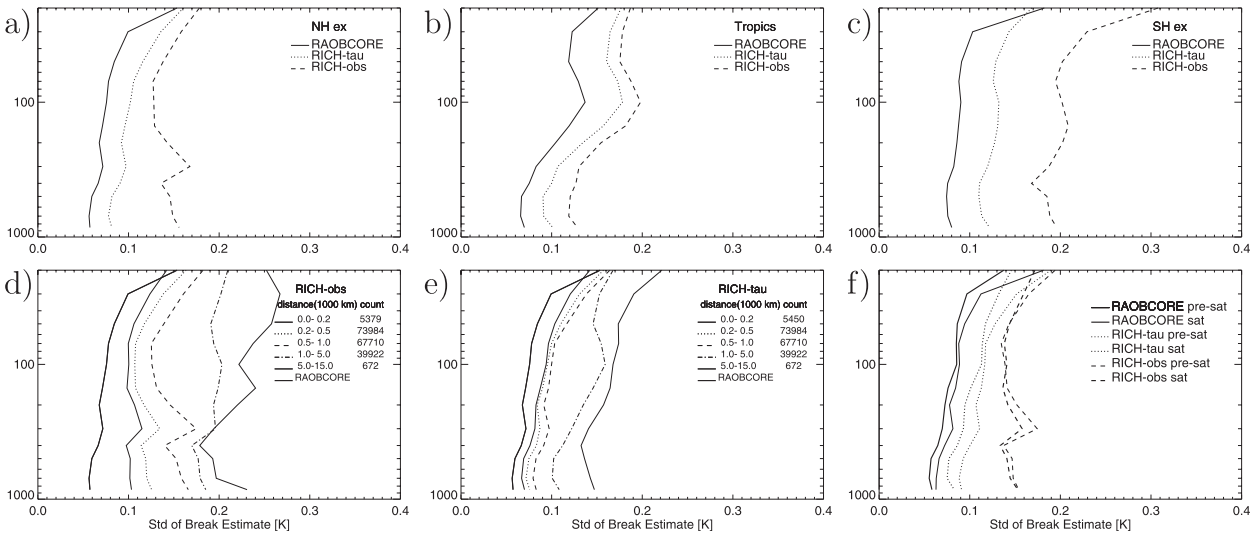

FIG. 2. Log-pressure profiles of sample error standard deviation of break size estimates as defined in Eq. (13) for RAOBCORE (solid), RICH-obs (dashed), and RICH-τ (dotted). First row shows dependency on region: (a) 20°–90°N, (b) 20°S–20°N, and (c) 90°–20°S. Second row shows dependency on (d),(e) distance between stations and epoch [(f) presatellite (1958–78) and satellite (1979–present) periods].

Figure 2 shows also to what extent standard errors are smaller in the NH extratropics than in the tropics and in the SH extratropics. For RICH there is a strong dependence of standard errors on the distance to the neighboring station. There is relatively weak dependence on the period considered. Estimates in the satellite era are only slightly more accurate than in the presatellite era. The vertical profile of the standard errors is interesting in that it does not increase monotonically with height. Especially for RICH-obs the tropopause region is the most challenging, since at those levels near jet streams it can happen that data in the troposphere at one station are compared with data at a neighboring station at the same pressure level that is already in the stratosphere.

It should be noted that only Δ_τ^k immediately yields the break size, whereas in RICH values of Δ_τ^{jk} or $\Delta_{\rm obs}^{jk}$ have to be averaged over several stations j to yield the break size estimate. This reduces the standard error to a certain extent, depending on the distribution of neighboring stations, but not much since the ${}^i{\rm obs}^k - {}^i{\rm obs}^j$ differences are not independent of, for example, the ${}^i{\rm obs}^k - {}^i{\rm obs}^{j+1}$ differences.

b. Sampling strategy for constructing the reference series

The expressions above describe break estimates from comparison of a test station k with a single reference station j. To reduce noise it is necessary to consult several reference stations and to build a weighted average over the break estimates. For RICH-obs, the break estimate at a known breakpoint date t at pressure level p is

$$\hat{\Delta}_{\rm obs}^k(p,t) = \frac{1}{\left(\sum_{j=1}^{J} w^{jk}\right)} \sum_{j=1}^{J} \Delta_{\rm obs}^{jk}(a^{jk}, b^{jk}, p, t) w^{jk}. \qquad (14)$$

The hat symbol denotes averaging over the break size estimates from comparison with a sample of neighboring stations j. Note that the intervals a^{jk}, b^{jk} are dependent on the tested station k and on reference station j. This dependency exists also in the formulas above but has not been made explicit to keep the notation simple. The location of breakpoints at reference station j as well as at the tested station k near the breakpoint to be adjusted determines the length of intervals a^{jk}, b^{jk}. The break estimates from the individual comparisons with reference series are then averaged using weights w^{jk}. For RICH-τ, we use the same averaging procedure.

While several weighting procedures are possible, we chose weights decreasing exponentially with distance:

$$w^{jk} = \exp(-d^{jk}/1500\,{\rm km}). \qquad (15)$$

The parameter d is not the usual spherical distance but has been chosen as

$$d^{jk} = r_E \Delta\varphi^{jk} + r_E(0.1\Delta\lambda^{jk}), \qquad (16)$$

where r_E is Earth's radius and $\Delta\varphi^{jk}$, $\Delta\lambda^{jk}$ are latitude and longitude differences between stations j and k. Note that the meridional distance is weighted much higher than the east–west distance, especially at low latitudes. This choice reflects the fact that climate zones depend mostly

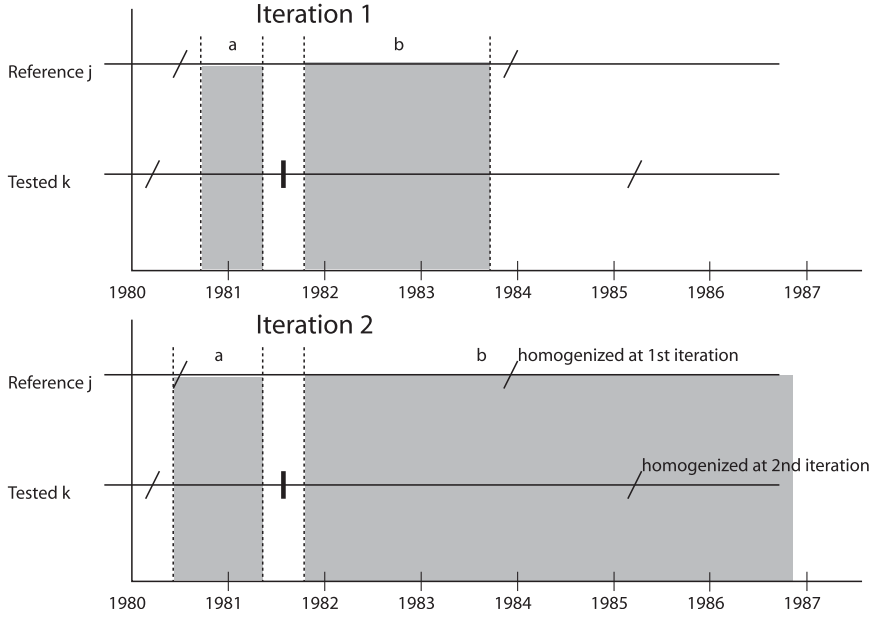

FIG. 3. Illustration of data selection when adjusting break (thick vertical line) at tested series. Slashes are other breakpoints at tested station and at reference station. In first iteration, time series must be considered inhomogeneous at all breakpoints. Gray areas indicate intervals used for break estimation. Data near breakpoints are not used to account for uncertainty in break date detection. In the second iteration, the reference series from the first iteration are considered homogeneous. Also the tested series is considered homogeneous after the breakpoint just tested, since the adjustment procedure works backward, starting from the most recent breakpoint. Thus the interval *b* is often much larger in second iteration. Interval *a* is bounded only by earlier breakpoints in the tested series but not in the reference series and thus may also be longer.

on latitude, at least in zeroth order (Kottek et al. 2006; Rougier 2007) and is consistent with findings of Wallis (1998) and McCarthy (2008). It also reduces the risk to make composites of stations where the tropopause height is different. The parameter d is used not only for weighting the composites, but also for choosing the stations used in the composites. Neighboring stations are sorted according to d^{jk}. Only the nearest stations that have enough data for break size estimation at station j are used for the composite. The number of neighbors varies between 3 and 30 (see section 4b below). Depending on the region, the most distant neighbor considered is between 300 km (over densely populated areas) and 10 000 km (in the tropical Pacific or the Southern Ocean) away.

Other choices for w^{jk} such as linear correlation of temperature series from reanalyses are also possible (Thorne et al. 2005b). However, correlation is rather height dependent. The linear temperature correlation patterns in the lower troposphere look very different to the correlation patterns at the tropopause level. We chose to avoid different weights at different levels.

The validity of the break size estimates relies on the assumption that the difference time series for estimating Δ^{jk}_{obs} or Δ^{jk}_{τ} are temporally homogeneous before and after a breakpoint at the test station. In other words, it is assumed that each difference in the series belongs to the same statistical population. To guarantee this, it is essential to know the dates of breakpoints of the tested stations and of the reference stations and to choose a^{jk} and b^{jk} accordingly. Only then averaging over inhomogeneities in the records can be avoided. Figure 3 shows how the averaging intervals a and b are influenced by the dates of neighboring breakpoints at the tested station and the reference station.

In the case of RICH, the dates of breakpoints have been determined with RAOBCORE. As described by Haimberger (2007), RAOBCORE analyzes ${}^{i}\tau^{k}$ (innovation) time series using a version of the standard normal homogeneity test (Alexandersson and Moberg 1997) that has been adapted to work well with difference series that have an annual cycle and data gaps, a common case for radiosonde temperatures especially at high levels. About 8000 breaks could be detected between 1958 and 2011 (an average of about seven breaks per station). With more liberal parameter settings, even more breaks could be identified, but too many breakpoints

restrict the length of the time series available for break size estimation and could lead to unnecessarily high noise levels in the break size estimates.

c. Adjustment procedure

The break sizes at station k are estimated, beginning with the most recent break. The record is then adjusted at this breakpoint. Then the next earliest break is adjusted, working backward in time. This procedure is performed for every station k on the globe with at least two years of data.

It has proven useful to perform the adjustment of the global radiosonde temperature dataset in two steps. In the first iteration, averaging over known breakpoints has been strictly avoided (see Fig. 3a). This idea is consistent with the concept of pairwise intercomparison together with using only homogeneous subperiods for estimating break sizes (Della-Marta and Wanner 2006; Sperka 2007; Caussinus and Mestre 2004; Menne and Williams 2009). Also, the number of reference series has been limited to a very small number (3 or 10) to reduce the risk of using of neighboring stations that are far away. However, the spatial noise in trend estimates after the first iteration is relatively large because of the small number of reference stations and the sometimes short time intervals permitted for break size estimation.

When the first homogenization has been applied at all radiosonde stations, one could finish. It proved highly advantageous, however, to perform a second adjustment step where the tested series are compared with neighboring series that have been homogenized in the first iteration. In the second iteration, it is assumed that the pervasive biases in the reference stations have been removed in the first iteration. If this is the case, it is allowed to average over breakpoint dates in the reference stations (see Fig. 3b). If it is known that this is not the case, because the adjustment failed in the first iteration, the RICH algorithm avoids averaging over those breakpoint dates in the reference stations. The main reason for failure in the first iteration at some pressure level is lack of reference data. In many cases only the uppermost pressure levels are affected. The default averaging interval of 8 years in the second iteration is reduced only by breaks in the test station occurring earlier than the break just estimated, and by data gaps. Since the conditions on break dates are less strict in the second iteration, more stations can be used as reference stations without the need to interpolate over large distances. Thus the number of stations used for break size estimation is set 3 times higher in the second iteration.

The two-step approach minimizes the risk of averaging over inhomogeneous samples. Nevertheless, the following additional measures have been taken to avoid averaging over breakpoints and thus noisy break estimates or even divergence of the iterative procedure:

- A neighboring station is used as reference only if it has no breakpoint 180 days before and 180 days after the breakpoint diagnosed by RAOBCORE at the tested station.
- At least 130–330 good values must be available for comparison before/after the breakpoint and the next breakpoint at the tested or the reference station.
- At least 30 days of data next to the breakpoints at the tested station and at the reference station are discarded to avoid inhomogeneities due to inaccurately detected breakpoints. If the averaging interval is long enough, 180 days of data are discarded. If the sample size is small (i.e., close to the minimum number specified above), the number of discarded data is reduced to 120, then 60, then 30.
- Care is taken to avoid unequal sampling of the annual cycle in the intervals before/after the breakpoint of the tested station, as described in Haimberger (2007). If, for example, no January values are available in interval a^{jk}, January values in interval b^{jk} are deleted.
- When calculating the break size estimate $\hat{\Delta}^{k}_{\mathrm{obs}}(p,t)$ from the sample of neighboring stations, the maximum and minimum break size estimate are discarded if the sample size is larger than three. This trimming of the mean leads to more robust break size estimates. We refrained from using the even more robust median or the interquartile range average, however, since we found that spatiotemporal consistency is better when using the mean or only slightly trimmed means (not shown).

All these measures are designed to ensure that the members of the compared samples in intervals a^{jk} and b^{jk} belong to the same populations so that the estimated means are meaningful. It may seem trivial to mention this, but we think that clean separation between homogeneous and inhomogeneous parts of the analyzed records is the key to successful homogenization. The good performance of the pairwise intercomparison methods of Caussinus and Mestre (2004) and Menne and Williams (2009) in a recent intercomparison of surface data homogenization methods (Venema et al. 2012) supports this.

It should also be noted that the above measures are only successful if the vast majority of breaks have been detected by RAOBCORE. Undetected breaks still may contaminate the break size estimates. However, the settings of RAOBCORE have been quite liberal and the sensitivity experiments below indicate that there are only few undetected breaks.

The consequent avoidance of averaging over inhomogeneities is most likely a key advantage compared to

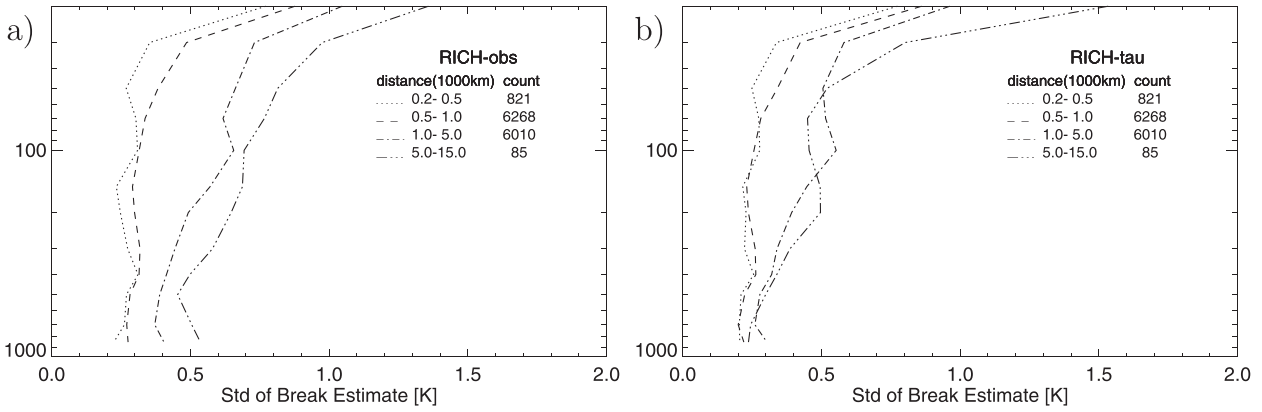

FIG. 4. Log-pressure profiles of rms difference between break size estimates from RAOBCORE and (a) RICH-obs and (b) RICH-τ for different maximum distances between stations.

Thorne et al. (2005b) and McCarthy et al. (2008) and to Sherwood et al. (2008), who have also tried to avoid it but used monthly data.

The use of daily data leaves much more choices in data selection and allows us to cope with breaks separated by only a few months. It also helps to avoid unjustified adjustment of sampling biases that may occur, for example, if balloons burst earlier if the stratosphere is colder. While daily data could be used to make more sophisticated adjustments that modify the probability distribution functions of the observations (e.g., Dai et al. 2011; Della-Marta and Wanner 2006), we restricted ourselves to modifying the mean, since this is challenging enough, as is clearly shown in Figs. 2 and 4.

ADJUSTMENT OF THE CLIMATOLOGY

As already discussed by Haimberger (2007), the availability of background departure time series allows efficient adjustment of the mean of short time series that cannot be analyzed for breakpoints and of time series that are known to have sizeable biases even in their most recent part. The basic assumption is that the mean background departure of the most recent part (at most 8 years, with a minimum 1/2 year for very short records) should be equal to the background departures of neighboring stations. The mean difference between the background departures $\nabla_\tau^{jk}(a)$ of stations j, k has already been defined in Eq. (9). The time interval a is now the most recent part of the series.

The adjustment of the most recent part is now calculated as

$$\hat{\Delta}_{mr}^{k}(p,t) = \frac{1}{\left(\sum_{j=1}^{J} w^{jk}\right)} \sum_{j=1}^{J} \nabla_\tau^{jk}(a^{jk}, p, t) w^{jk}. \quad (17)$$

This is formally quite similar to the break size calculation used in RICH-τ. The weights have the same meaning as above and of course only one time interval must be considered per intercomparison. Since the adjustment of the climatology is performed after the homogeneity adjustments, the involved time series are considered homogeneous.

While the adjustment of the climatology appears formally simpler, it works only if the neighboring stations are unbiased. This strong assumption is justifiable only for a few of the most recent radiosonde types in use. We chose stations using Vaisala RS90, RS92, Meisei, and current Sippican radiosondes (type codes 37, 52, 53, 60, 61, 62, 63, 66, 67, 71, 72, 73, 74, 78, 79, 80, 81, 82, 83, 47, 55, 56, 26, 76, 85, 86, and 87 according to the World Meteorological Organization (WMO) Binary Universal Form for data Representation (BUFR) common code table C-2) as reference types stations (about 400). These are also fairly well distributed over Earth.

The adjustment of the climatologies does not affect trend analysis. However, it allows to bias-correct even very short records that are normally not considered for climatological purposes but are certainly useful for reanalysis efforts, and it reduces the rejection rate of records during assimilation.

3. Input data and comparison strategy

The following input data have been used for this intercomparison:

- Daily (0000 and 1200 UTC) radiosonde data on 16 standard pressure levels (10, 20, 30, 50, 70, 100, 150, 200, 250, 300, 400, 500, 700, 850, 925, 1000 hPa) from the Integrated Global Radiosonde Archive (IGRA; Durre et al. 2006) and the ERA-40/ERA-Interim radiosonde archives from 1958 onward. Metadata information is mainly based on the IGRA metadata file, whose origins go back to Gaffen (1996) and which

is constantly updated (http://www1.ncdc.noaa.gov/pub/data/igra/igra-metadata.txt). This valuable source of information reports 8715 events regarding the radiosonde type alone. Additional metadata information is retrieved from ERA-40/ERA-Interim radiosonde records, which often include the radiosonde type as transmitted by the Global Telecommunication System from about 1990 onward.

- Background forecasts from the ERA-40/ERA-Interim reanalyses interpolated to the station locations. The properties of the forecasts are quite different between ERA-40 (Uppala et al. 2005; used for the period 1958–78) and ERA-Interim (Dee et al. 2011a,b; for the period 1979–2010). Most notably, ERA-40 uses 6-hourly cycling and a three-dimensional variational data assimilation (3D-VAR) system, whereas ERA-Interim uses a considerably more advanced 4D-VAR assimilation system with a better forecast model, 12-h cycling, and variational bias correction of satellite data. No bias correction has been applied to radiosonde temperatures in ERA-40 during the period 1958–79. In ERA-Interim, the radiosondes have been adjusted to remove both the annual mean bias, using RAOBCORE v1.3 adjustments, and the annual cycle of the radiation error. For a more detailed description, see Haimberger and Andrae (2011).

To avoid a shift in the reference at the change point from ERA-40 to ERA-Interim in 1979, zonal mean background departures between reanalysis and radiosonde observations have been calculated for the years 1978 and 1980. The difference in these zonal mean departures has been taken as estimate for the shift between ERA-40 and ERA-Interim temperatures. 1979 has been avoided since it had atypical data coverage due to the Global Weather Experiment in 1979. The shift has been subtracted from the ERA-40 background departures. While this procedure may reduce the break detection power of RAOBCORE during these years, it has been found to yield the smoothest transition from ERA-40 to ERA-Interim. Alternative approaches, such as using twentieth-century reanalysis (Compo et al. 2011) data or using regional instead of zonal means, left some shifts in the time series of zonal mean background departures after the adjustment and also led to less spatiotemporal consistency of trend estimates in time intervals that include the year 1979. Further investigations of the uncertainties introduced through this transition will nevertheless be necessary in the future.

Solely for intercomparison, we use the following datasets:

- Remote Sensing Systems (RSS) MSU v3.3 brightness temperatures (Mears and Wentz 2009) on the standard 2.5° × 2.5° lat/lon grid at lower stratospheric (LS) and upper tropospheric (TS) and midtropospheric (MT) layers. An ensemble of realizations of this dataset using a Monte Carlo technique is available as well (Mears et al. 2011). From this ensemble we use the mean and the 5% and 95% percentiles, where the ensemble members have been sorted after their global mean trend.
- University of Alabama at Huntsville (UAH) MSU v5.4 data (Christy et al. 2003) on the standard 2.5° × 2.5° lat/lon grid at the LS and MT layers and National Environmental Satellite, Data and Information Service (NESDIS) Center for Satellite Applications and Research (STAR) (Zou et al. 2009) version 2.0 data at LS, TS, and MT layers. The RSS, UAH, and STAR temperature datasets have been calculated from the same raw radiance datasets, though with quite different analysis methods.
- The updated Hadley Centre Atmospheric Temperature (HadAT2) dataset (Thorne et al. 2005a; McCarthy et al. 2008). It is a well-known homogeneity-adjusted pure radiosonde dataset containing 676 stations. The input data resolution is monthly and the adjustments are provided with seasonal resolution. An ensemble of realizations of this dataset is described by Thorne et al. (2011a). Minima, 25%/50%/75%, and maxima of zonal belt mean trends 1979–2003 are available. Some of the variations in this ensemble appear very strong (e.g., a removal of signal experiment). Since the trend interval is also shorter than the one analyzed in the present paper, the uncertainties are likely smaller than the values given for this ensemble. Therefore we preferred to use the interquartile range of this ensemble as uncertainty estimate in the comparisons below.

For comparison with satellite data, the radiosonde and reanalysis temperature profiles have been converted into brightness temperatures using the Radiative Transfer for TOVS version 10 (RTTOV v10) software package (Saunders et al. 2011). This is an improvement to earlier studies that used static weighting functions that leads to up to 20% less variance in the difference series between brightness temperatures from adjusted radiosonde data and MSU data, especially in the lower stratosphere.

We did comparisons also with Global Positioning System radio occultation data for the period 2001–10, but these are published elsewhere (Ladstätter et al. 2011; Steiner et al. 2011).

4. Estimation of adjustment uncertainty

Homogenization is useful if the adjustment uncertainties are smaller than the time variation of biases that need to be estimated. That has to be shown for the

adjustments of individual breaks as well as for the global mean adjustments that affect estimates of the global climate change signal.

For this purpose we need to understand the uncertainties in the adjustments that are calculated by RICH/RAOBCORE. We reduce the uncertainty of the low-frequency variability estimates only if we can show that the adjustment uncertainties are appreciably smaller than the estimated time-varying biases. This section tries to highlight some manifestations of the adjustment errors due to sampling of neighbor stations and due to setting of parameters in the adjustment methods. Differences between RICH-obs/RICH-τ and RAOBCORE may be seen as manifestation of structural uncertainties (Thorne et al. 2005a), whereas differences within the RICH-ensembles are interpreted as parametric uncertainties.

About 8000 breaks are found by RAOBCORE in the period 1958–2011, and for each of these breaks, samples of neighboring records to adjust this break have to be determined. Individual break profiles tell much about the uncertainties involved in estimating the adjustments. Figure 1 shows the estimated profiles of a break caused by transition from VIZ-A to Space Data radiosondes in 1989 at station Bethel (70219) in Alaska. The thin lines in the upper-left panel of Fig. 1 are individual break size estimates from comparison of neighboring anomaly records (Δ^{jk}_{obs}). The thick green line is the RAOBCORE estimate for this break, and the thick blue line is the distance weighted mean of the Δ^{jk}_{obs} profiles, which comprises the RICH-obs break size estimate. The break size estimates are all clearly nonzero, and one can say that the adjustment uncertainties are much smaller in this case than the estimated shift. While the agreement between RAOBCORE and RICH-obs estimates is reasonable, there is considerable spread of the Δ^{jk}_{obs}.

When estimating the same break by comparing neighboring innovation records (Δ^{jk}_{τ}), the spread is much smaller. For this plot it appears clearly advantageous to use innovations instead of observations of neighboring stations for comparison. In particular, the reduced spread indicates that regional climate anomalies have substantial influence on the break size estimates if they are not accounted for. This effect is largest for remote stations where one has to compare stations that are far apart. In the case of this specific break at Bethel the maximum distance to a reference station is 3070 km.

Note also that no vertical smoothing has been applied at any stage of break size estimation, which is an important feature of the present adjustment system. Earlier experiments with vertical smoothing generally yielded unsatisfactory results, since the vertical profiles of the biases can be rather complex. At some occasions there are individual profiles that are clear outliers. This may indicate that the reference station record is not as homogeneous as thought, or it may come from a very different region than the other reference profiles. Of course, there may also be issues with data density in the time series that may cause spread. As mentioned above, the maximum and minimum individual estimates are removed to increase the robustness of the estimate.

While Fig. 2 shows the sampling errors involved in estimating the break sizes in individual profiles, it is instructive to see also how much RAOBCORE and RICH break size estimates differ. In the example in Fig. 1 this difference between the thick blue and green curves is on the order of 0.5 K. Figure 4 depicts the rms difference between RAOBCORE and RICH-obs adjustment estimates averaged over stations with different ''remoteness,'' measured as the distance between the tested station and the most distant station used for comparison. As long as this distance is less than 1000 km the rms difference is quite small for both RICH-obs and RICH-τ. If it is larger the difference between RAOBCORE and RICH estimates can become relatively large, especially for RICH-obs. The difference becomes also large for the highest levels ($<$50 hPa), mainly because reference stations with enough data are hard to find there.

a. Spatiotemporal consistency of adjusted datasets

Inhomogeneities in an observation dataset often manifest themselves in large trend discrepancies at nearby stations. Trends from homogenized records should therefore be spatially more consistent than the unadjusted datasets. Figure 5 indeed shows improved spatiotemporal consistency of trend estimates for the period 1979–2006 for the MSU-equivalent lower stratospheric layer for all three homogenization methods. The adjusted datasets are more consistent than those shown in Fig. 1 of Haimberger et al. (2008) with parameter settings similar to those used in this paper. The improvements mainly come from the improved background and changes in the neighbor station selection. The use of RTTOV for calculating brightness temperatures also contributes to the better consistency, which has been quantified with a cost function[1] introduced by Haimberger (2007).

While we think it is clear that a homogenization algorithm should improve spatiotemporal consistency, this parameter is not necessarily an indicator of improved temporal homogeneity of large scale means. For example, MSU brightness temperatures are spatiotemporally

[1] $\text{Cost}[1/N(N-1)]\sum_i^N \sum_{j=i+1}^N \Delta_{ij} \exp^{-d_{ij}/1000km}$, where i and j are radiosonde stations, d_{ij} is distance in km and Δ_{ij} is the trend difference in K decade^{-1}.

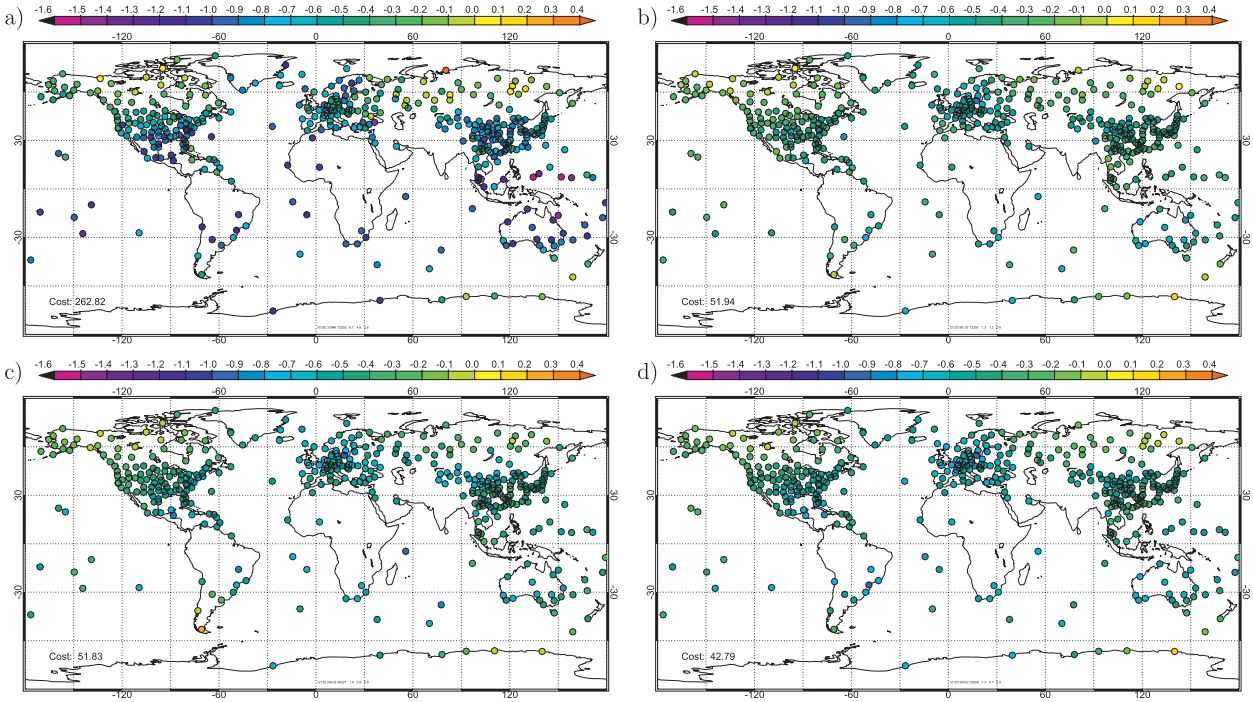

FIG. 5. Daily mean TLS trends 1979–2006: (a) unadjusted, (b) adjusted with RAOBCORE, (c) adjusted with RICH-obs ensemble mean, and (d) adjusted with RICH-τ ensemble mean. Reduced cost values (in lower left corners of panels) indicate improvement compared to Fig. 1 of Haimberger et al. (2008). Only for this plot the 30-hPa level has been omitted in the brightness temperature calculation to allow a better comparison.

very consistent, but nevertheless can have significant inhomogeneities or at least uncertainties (Mears et al. 2011). The temporal homogeneity can only be guaranteed with measurement practices that allow traceability to SI standards (Seidel et al. 2009). These are lacking for historic radiosonde observations. One can, however, constrain the uncertainties by comparison with independent data and with sensitivity experiments where uncertain parameters of the adjustment system are varied.

b. Sensitivity experiments and ensembles

In RICH, the following parameters have been varied to estimate their influence on adjustment results when the breakpoints from RAOBCORE are fixed:

- Using either RICH-obs or RICH-τ for break adjustments. Both versions have their merits as discussed above, but they can yield quite different results particularly at remote locations.
- The test station and a reference station must have a minimum number of data points in both intervals *a* and *b* (see Fig. 3) to deliver a valid break size estimate at one level. This minimum data threshold for adjustment is varied between 130 and 330 values for interval *a* and ~200–900 values for interval *b*. The minimum number is dependent on the pressure level. For the higher levels fewer data are required. A smaller number increases the chance to find nearby neighbors for adjustment calculation but produces higher noise and makes the estimates more susceptible to effects of short-term climate anomalies, especially for RICH-obs.
- The number of neighbors used for calculating the composites is another important parameter. It is varied between 3 and 10 in the first iteration and between 9 and 30 in the second iteration (see Fig. 3). A small number reduces the risk that information from very distant stations can enter the break size estimates but leads to noisier estimates due to the small sample of profiles.
- We used inverse distance weighting of the neighbors. The 1/*e* decay of the weighting function defined in (15) was varied between 3000 and 5000 km.
- When calculating adjustments, one can start with the uppermost or the lowermost pressure level. When one starts with the uppermost level, suitable neighbors are probably farther away but those likely have a full profile. The same set of stations is used for all levels of an adjustment profile in this case.

When one starts with the lowermost level, nearby stations are more likely to be used, even if they have few data in the uppermost levels. Data for the uppermost levels are then taken from more distant but complete

records. This means that for one adjustment profile, different sets of stations may be used for different levels.

- Treatment of the Phillips-RS4-MKIII to Vaisala-RS80 transition over Australia/Malaysia and the surrounding islands. This transition during the years 1987–89 was a major change in the radiosonde observing system in the tropics and the Southern Hemisphere. Because of the sparseness of reference stations and due to the fact that a major ENSO event (1987 El Niño followed by a strong La Niña in 1988) caused large regional anomalies over the tropics, there is much uncertainty involved in adjusting with neighbor composites. We used two strategies:

  a) Adjust the Phillips-Vaisala breaks using a profile estimated from an international radiosonde intercomparison (Nash and Schmidlin 1987). In particular, RICH-obs has problems getting reliable break size estimates when two breakpoints are relatively close to each other. As an additional measure, some (about 50) breakpoints at remote stations have been deleted to get longer intervals for break size estimation.
  b) Let RICH work as usual (i.e., strictly use neighbor intercomparison for break size estimation; do not delete any breakpoints detected in RAOBCORE).

In total, two choices have been tried for six parameters, yielding an ensemble with $2^6 = 64$ members. This approach is less exhaustive than, for example, employing a Monte Carlo technique but nevertheless allows us to explore at least part of the parameter space of the present adjustment methods.

The combination of all these parameter settings yields moderate spread in the RICH adjusted trend profiles. While there is not enough space to show the effect of each particular parameter setting, it can be said that all contribute appreciably to the spread. The effect of using either RICH-obs or RICH-τ is indicated by the gray shades. For the interval shown in the tropics, RICH-obs leads to vertically smoother trend profiles than RICH-τ, which yields trend profiles that are roughly between RAOBCORE and RICH-obs. Below we often split the ensemble into 32-member RICH-obs and RICH-τ ensembles because of their fundamental difference concerning data dependency. Separate ensemble means of the RICH-obs and RICH-τ members have been used in Figs. 6–10 below.

Since RICH results are also dependent upon the breakpoint dataset provided by RAOBCORE, we conducted sensitivity experiments with breakpoints from different RAOBCORE versions:

1) breakpoints from RAOBCORE v1.5 (these are used for most plots in the present paper);
2) breakpoints from RAOBCORE v1.5 where no prior adjustment of ERA-40 background between 1971 and 1978 has been applied, similar to RAOBCORE v1.3 in Haimberger et al. (2008);
3) breakpoints from RAOBCORE v1.5 where no metadata from documented changes of equipment are taken into account;
4) breakpoints from RAOBCORE v1.5 with neither background adjustment nor metadata; and
5) RAOBCORE v1.4 breakpoints, as in Haimberger et al. (2008). This is for comparison with older findings. The main differences between RAOBCORE v1.4 and RAOBCORE v1.5 are the absence of breaks after 2005 in RAOBCORE v1.4 and the use of ERA-Interim as reference from 1979 onward in RAOBCORE v1.5.

5. Selected results

Adjustment results are documented online (http://www.univie.ac.at/theoret-met/research/raobcore/) with thousands of adjustment profile plots, trend maps, and time series. Only a few are highlighted here. Figure 5 shows how RAOBCORE and the two RICH versions improve the spatial consistency of TLS equivalent trends. This figure can be compared with Fig. 1 of Haimberger et al. (2008). A clear improvement is noticeable compared to these earlier versions of RICH-obs and RAOBCORE.

Much improved spatiotemporal consistency can be found also in the presatellite period. This period was characterized by several quite stable national observing systems, like in the United States, but also by observing systems with extreme changes. Most notably the stratospheric temperature biases in early MESURAL radiosondes shifted by up to 10 degrees K over Europe as well as at several stations in the Pacific. There were also strong changes over the former Soviet Union, and these shifts lead to the rather noisy trend pattern in Fig. 6a. The mean RICH-obs adjustments are able to remove most breaks, yielding a surprisingly consistent map of trend estimates even at 100 hPa, as shown in Fig. 6b. Chinese radiosondes are available but are missing in this plot since early Chinese ascents did not reach higher than 300 hPa. The trend map (trend cost value 221) is even more consistent than RAOBCORE adjusted trends (cost 355) and trends from ERA-40/ERA-Interim (cost 248). If we restrict the ensemble to members that use 30 neighbors, the cost is reduced even to 185.

One should not ignore, however, that there are a few remote stations whose trends still look unrealistic after adjustment. We could have removed them but we found it instructive to see the limitations of the method as well. It is also important to recognize that we show only the ensemble mean. Many of these station trends that look

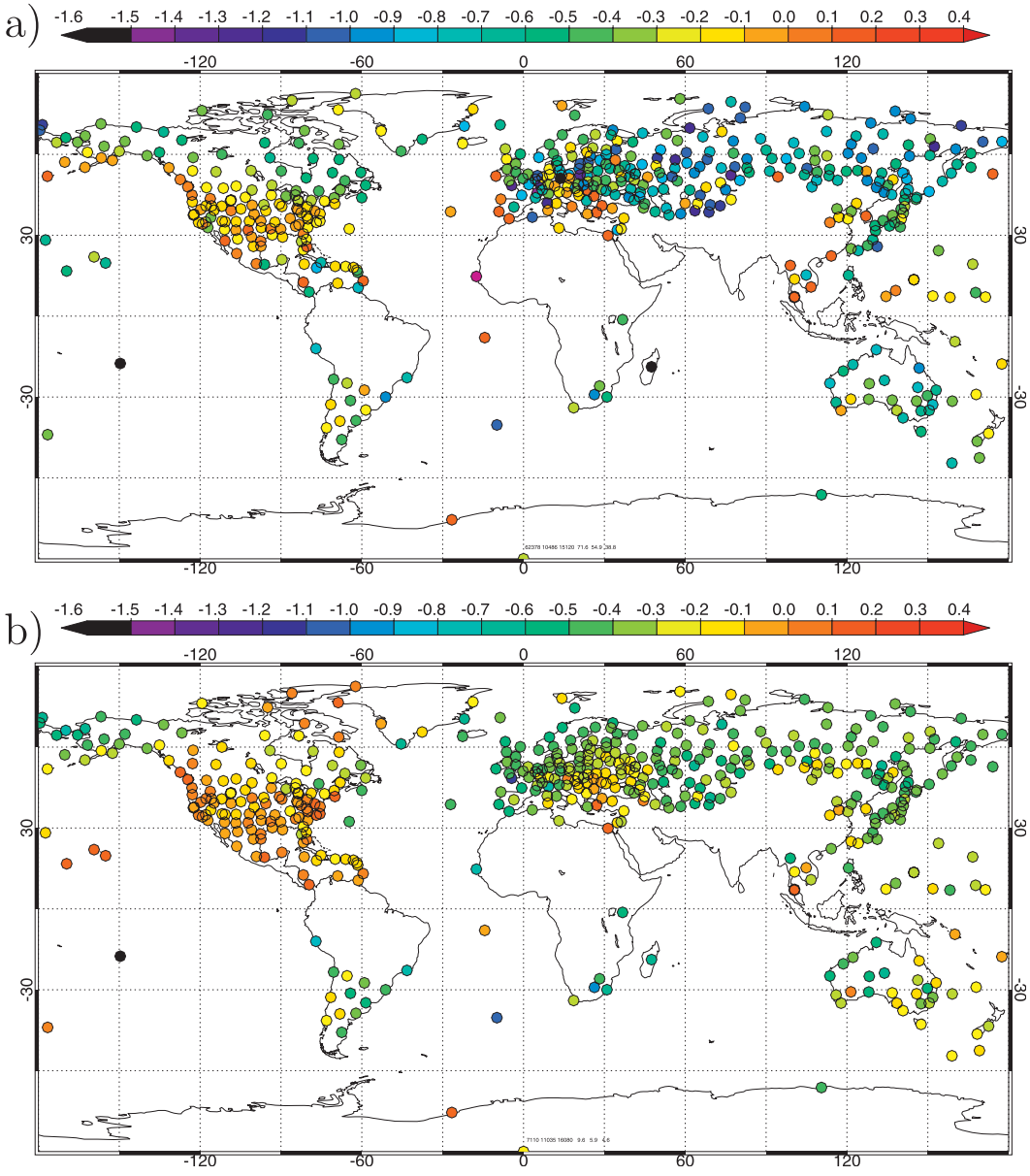

FIG. 6. Daily mean 100-hPa trends 1964–84 in units K decade^{-1}: (a) unadjusted, trend cost 1564, (b) adjusted with ensemble mean RICH-obs, trend cost 221. Especially in the early period, RICH has been improved compared to Haimberger et al. (2008). The RICH cost value with this past version of RICH was 508.

suspicious in the ensemble mean are consistent with their neighbors in some ensemble members. Certainly these stations need to be worked on in future releases. It is also interesting to see that trends over America tend to be close to zero or even positive whereas trends over Europe/Asia tend to be negative. It is quite possible that the breaks that lead to strong negative trends in the unadjusted data over the former Soviet Union have not been fully removed.

Since there are practically no other upper air data to compare with, the internal consistency of the adjusted dataset, the spread of adjustment ensembles, and comparison with other adjusted radiosonde datasets, such as Thorne et al. (2011a) must be the main quality measures. The sensitivity experiments below and comparison with HadAT are some steps in this direction, although more efforts in this direction are needed.

a. Comparison with satellite data

For the satellite era from 1979 onward, there are more possibilities for intercomparison. Figures 7 and 8 show Hovmöller plots of zonal mean temperature anomaly

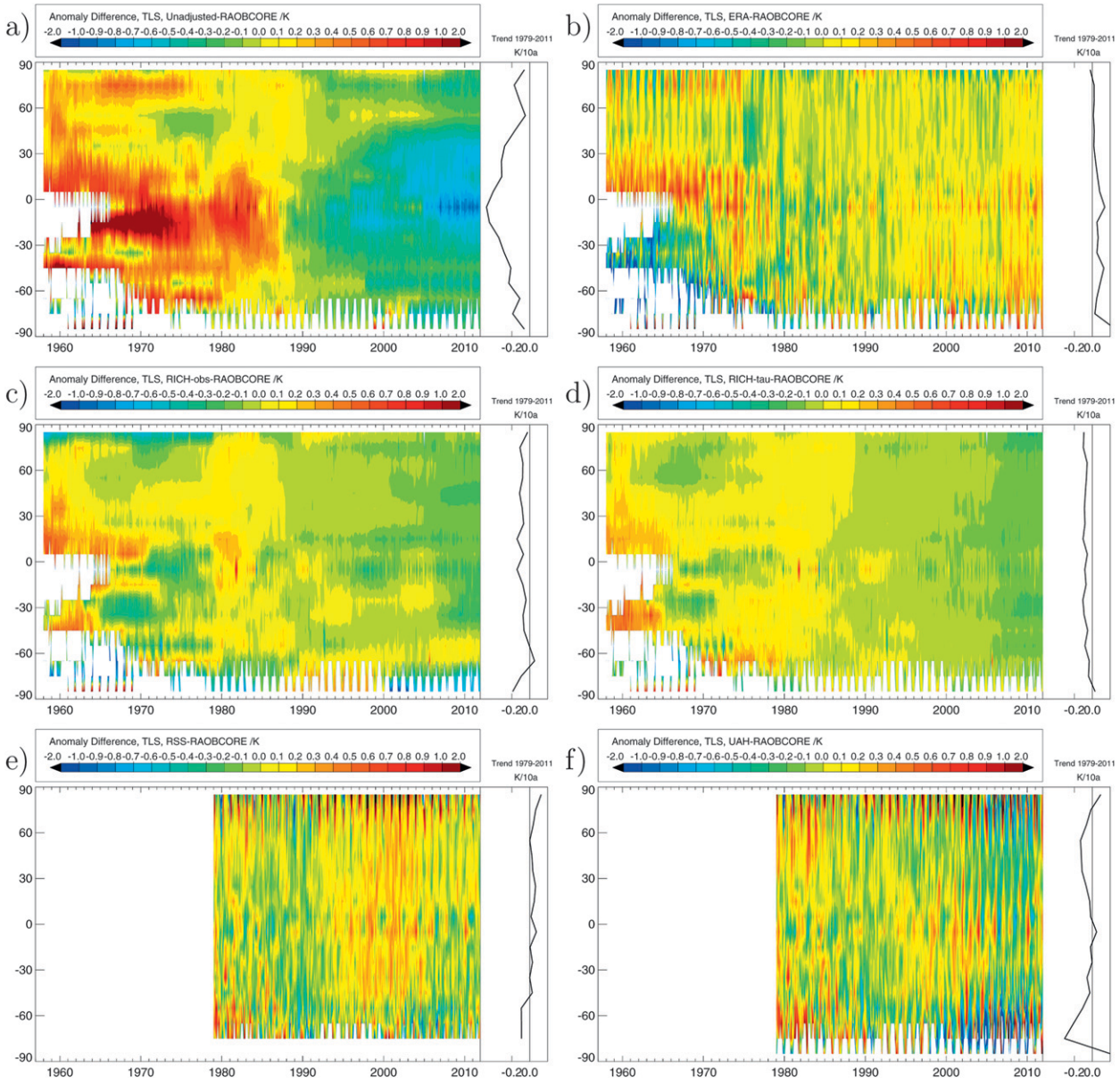

FIG. 7. Hovmöller plots of zonal mean temperature anomaly differences (anomalies relative to 2001–10 climatology, all differences relative to RAOBCORE v1.5 adjusted temperatures) for the MSU LS layer. Black lines at right of plots show trend difference 1979–2010. Differences to (a) unadjusted radiosonde data, (b) to ERA-40/ERA-Interim reanalyses, (c) mean of RICH-obs ensemble, (d) mean of RICH-τ ensemble, (e) RSS brightness temperatures, and (f) UAH brightness temperatures.

differences at the lower stratospheric and midtropospheric layers. RAOBCORE v1.5 adjusted temperature anomalies are taken as reference.[2] Figure 7a shows how substantial the temperature adjustments are in the lower stratosphere, particularly in the tropics. The absolute lower stratospheric cooling trends are essentially halved by the adjustments. The strong warm bias of the unadjusted radiosondes in the tropics in the presatellite era mainly stems from many stations using MESURAL equipment in the tropical Pacific. Another important period of changes was 1987–89 when Australia changed the radiosonde type and the early to mid-1990s when several former French colonies and the United States changed the radiosonde type in the tropical Pacific. Relatively small adjustments are made at polar regions, at least in the zonal mean.

[2] This does by no means imply that the RAOBCORE-adjusted records are the most reliable dataset.

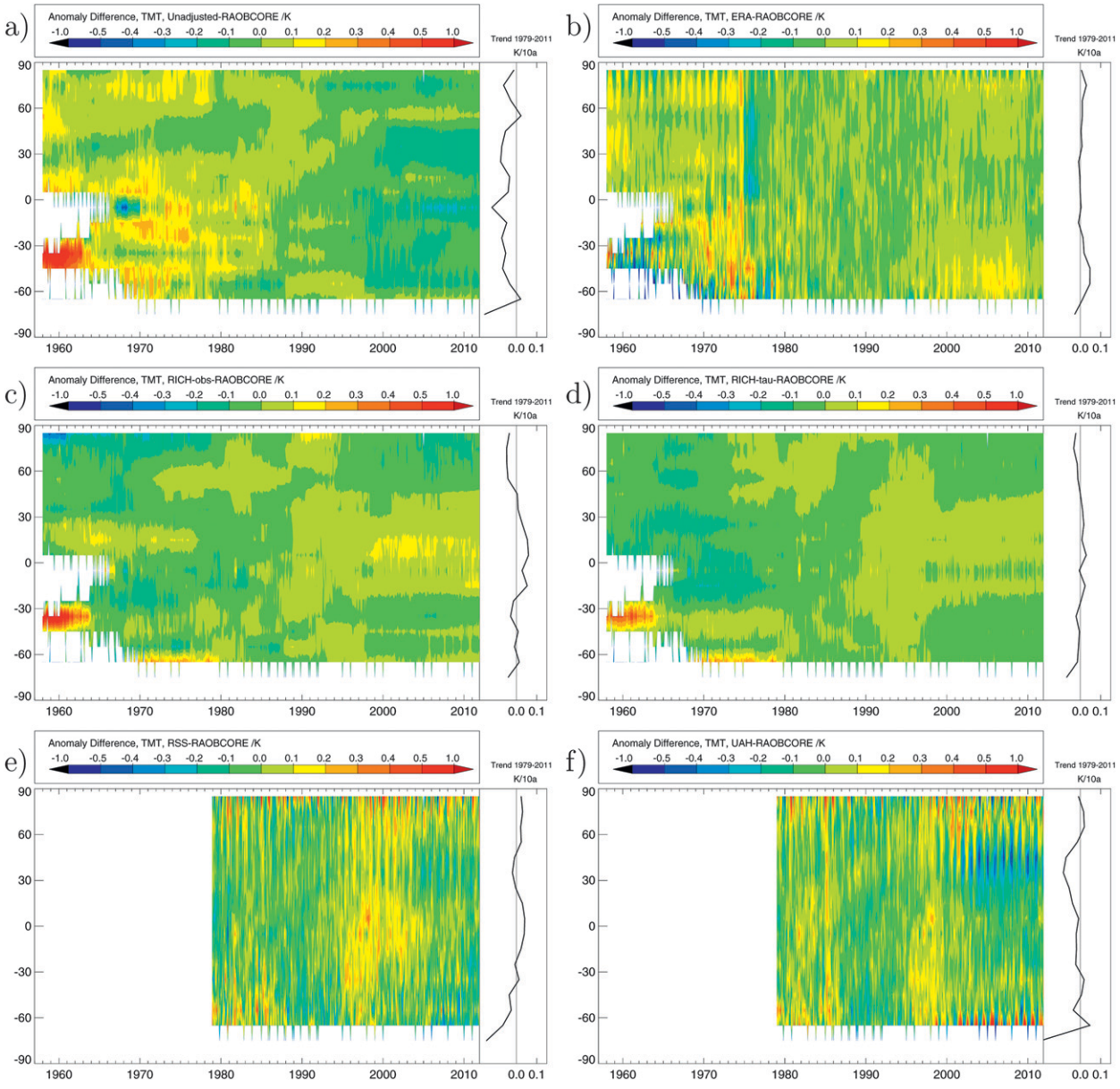

FIG. 8. As in Fig. 7, but for the MSU MT layer. Note different contour interval and scaling.

The ERA-Interim analysis shows less cooling than RAOBCORE in the LS in the satellite era, particularly in the tropics and the Southern Hemisphere. Warming relative to RAOBCORE occurs in 2007, when data from the Constellation Observing System for Meteorology, Ionosphere and Climate (COSMIC) have been introduced in ERA-Interim (Poli et al. 2010). Both RICH versions (mean RICH-obs and mean RICH-τ) show more cooling than RAOBCORE at all latitudes, the difference being on the order of 0.1 K decade^{-1}.

RSS and STAR (not shown) data yield less cooling than RAOBCORE in the LS, although RSS temperatures have recently (since ca. 2002) cooled compared to RAOBCORE. UAH shows about the same cooling as RAOBCORE/RICH. Cooling from UAH in the extratropics is rather strong compared to the other datasets.

The biases in the MT layer (Fig. 8) are much weaker but still the unadjusted data show cooling relative to RAOBCORE and the satellite datasets. Agreement between RAOBCORE, RICH-τ, and ERA-Interim reanalysis is excellent in the satellite era. In the pre-satellite era, RAOBCORE anomalies are cooler than ERA-40, and known homogeneity problems in ERA-40 in the 1970s (Uppala et al. 2005) are evident. There is

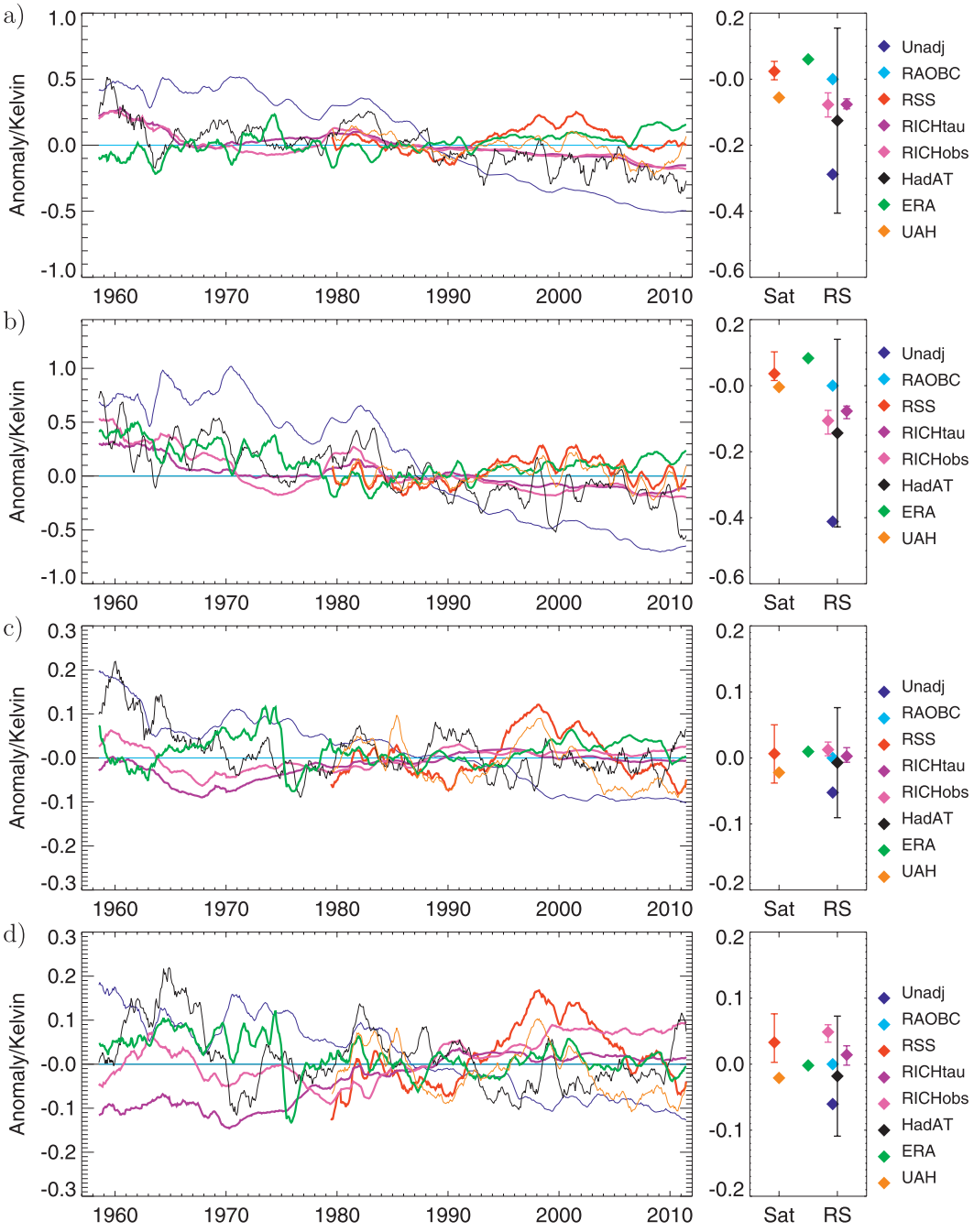

FIG. 9. MSU equivalent temperature difference time series relative to RAOBCORE v1.5. Trend difference for period 1979–2011 in K decade^{-1} is indicated in right panels. (a) LS global, (b) LS tropics (20°S–20°N), (c) MT global, (d) MT tropics. Note scaling differences. Uncertainty bars for RSS, HadAT, RICH-obs, and RICH-τ are 5% and 95% percentiles. Original HadAT trend spread valid for 1979–2003 has been reduced by a factor of 1.32 for the longer interval 1979–2011. Absolute trends of RAOBCORE v1.5 can be found in Fig. 13.

better consistency with HadAT (not shown), although there is more warming in RAOBCORE/RICH than in HadAT. The RSS and UAH satellite data show generally stronger warming than the radiosonde datasets and ERA-Interim until 2002 and substantial cooling compared to these datasets thereafter, especially in midlatitude regions. This cooling in the past decade is noticeable not only compared to radiosondes but also compared to GPS radio occultation (GPS-RO) data (Ladstätter et al. 2011; Steiner et al. 2011). Those authors argue that residual

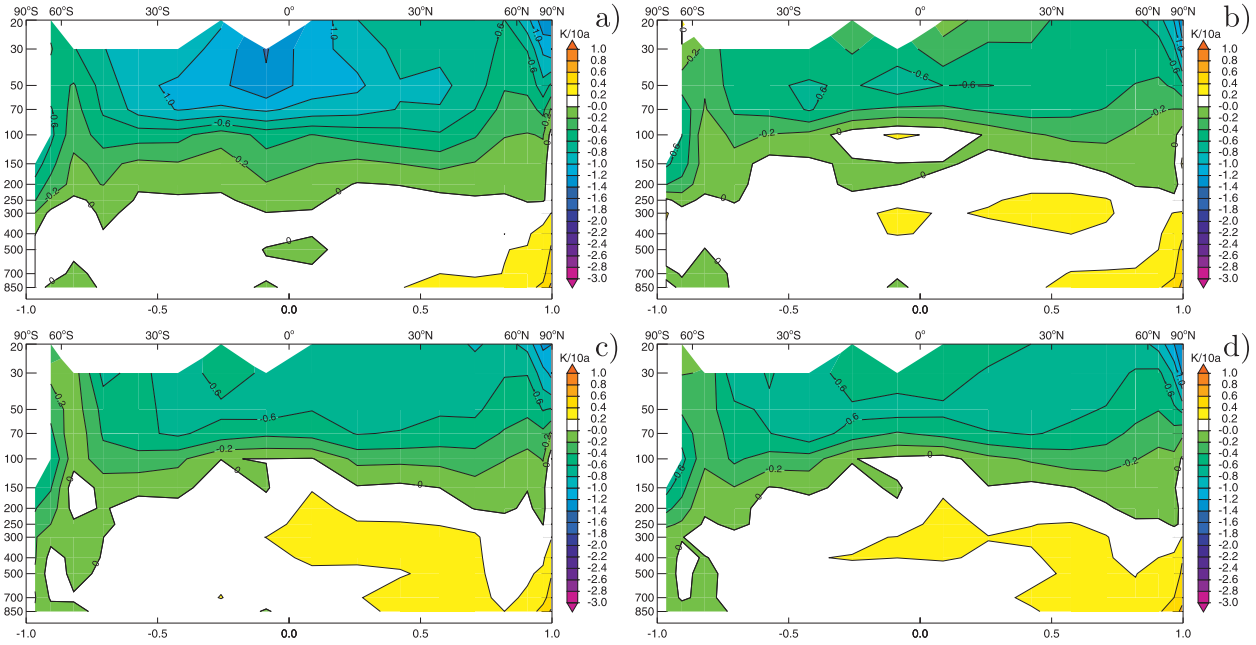

FIG. 10. Zonal mean trends 1979–2010 from (a) unadjusted, (b) RAOBCORE adjusted, (c) mean RICH-obs adjusted, and (d) mean RICH-τ-adjusted radiosondes.

inhomogeneities in the MSU datasets are the most likely cause given the excellent agreement between GPS-RO data from various platforms and also given the good agreement between radiosondes and GPS-RO after 2002.

Figure 9 shows global belt mean MSU (equivalent) brightness temperature anomaly differences with respect to RAOBCORE v1.5. RAOBCORE and RICH show better agreement with satellite data in the LS than HadAT in the satellite era. In the presatellite era, there is no upper air dataset independent of radiosondes. We can only compare with other radiosonde datasets and reanalyses. During this period RAOBCORE adjusted data have least cooling in the tropics, whereas they show good agreement with the ERA-40 reanalysis when averaged over the whole globe. In the tropics, the ERA-40 analysis may be biased warm in the LS due to the effect of unadjusted MESURAL observations before 1975.

The lower panels show the comparison in the MT layer. Note that the temperature scale is very fine: biases of 0.1 K appear large in this plot. There is generally excellent agreement of between RICH/RAOBCORE and the ERA-40/ERA-Interim analysis. Only in the tropics RICH-obs shows more warming in the satellite era, most likely because its profiles do not exhibit the warming minimum in the tropics around 700 hPa that is evident in ERA-Interim and RAOBCORE adjusted data. The transient warming feature of the UAH and RSS MT temperatures relative to RAOBCORE with a peak in 1998 that has been noted above for the LS is very evident also in the MT. It could be related to enhanced uncertainty in the MSU record at this time (Mears et al. 2011).

The zonal mean trend spread within the RICH-obs and RICH-τ ensembles is rather small in the satellite era compared to the spread given by the HadAT and RSS ensembles. Larger spread may be achievable by varying the breakpoints, but Figs. 11c and 11d suggest that this also has limited impact. We found that too strong variation of parameters inevitably led to strongly reduced spatiotemporal consistency of trend estimates and is therefore not justifiable. If one considers the differences among RAOBCORE, RICH-obs, and RICH-τ as crude uncertainty estimates, it is comparable to the uncertainties estimated from the RSS and HadAT ensembles.

In the presatellite era, Fig. 9 shows more spread between the radiosonde datasets. It should be noted, however, that observation density is quite limited such that different treatment of one breakpoint at one station in the tropics can have quite an effect on the curves shown in Fig. 9. RICH-τ yields the coolest MT temperatures. The difficulties in ERA-40 analyses due to assimilation of Vertical Temperature Profile Radiometer (VTPR) temperatures (Uppala et al. 2005) are evident in the analysis time series despite measures to adjust the global mean background forecasts between 1971 and 1978.

Despite some improvements in spatiotemporal consistency also for the time intervals 1973–2006 and 1958–2006

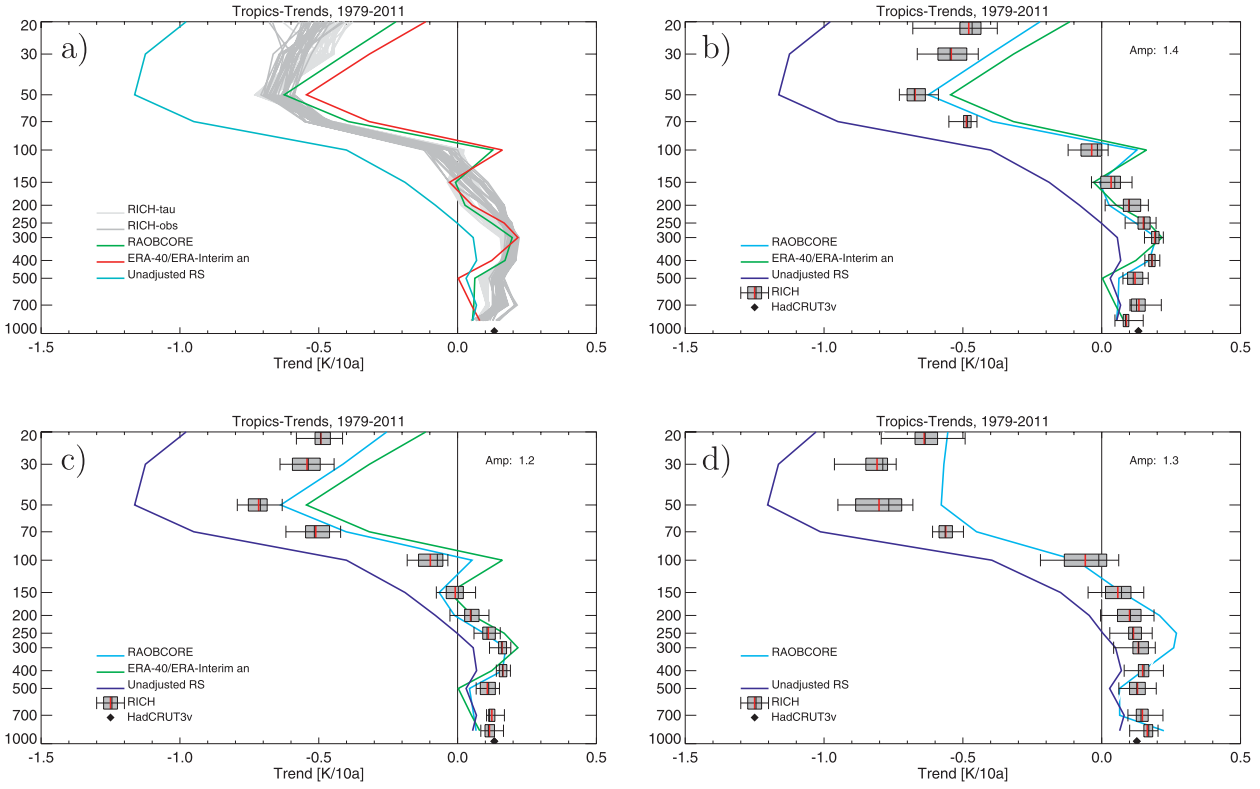

FIG. 11. Plots of tropical belt mean trends, using different RAOBCORE versions for break detection, using (a) RAOBCORE v1.5, (b) the same data as in (a) but ensembles are depicted with whisker quantile plots [markers in whiskers are minimum, maximum, 25%, 50%, and 75% quantiles, and mean (red)], (c) RAOBCORE 1.5 but without background correction between 1971 and 1978 and without metadata, and (d) the earlier RAOBCORE version v1.4.

(not shown) at lower levels in the tropics compared to RAOBCORE 1.4, one should therefore still be cautious when interpreting time series from single stations or small regions. One should also note that the adjustments only affect temperatures, not dewpoints. This is important for the analysis of radiosonde humidity measures such as dewpoint depression time series (see, e.g., Dai et al. 2011). For example, the difference between the various radiosonde datasets is 0.2–0.3 K in the 1960s in the tropical MT layer in Fig. 9d). A dewpoint depression error of 0.2 K converts to about 1% uncertainty in relative humidity.

b. Vertical structure of temperature changes

One strength of radiosonde data is their vertical resolution, which allows us to resolve the sharp transition from tropospheric heating to stratospheric cooling much better than MSU data can.

Figure 10 shows that all three homogenization methods yield a warming maximum in the tropical midtroposphere in the period 1979–2010 in contrast to unadjusted data, which show cooling in the deep tropics' troposphere. RAOBCORE yields the weakest maximum in the tropical midtroposphere; however, it has an additional warming maximum at the 100-hPa level that appears to be inherited from the ERA-Interim bg (not shown) that has a similar spike. The medians of the RICH-τ and RICH-obs ensembles yield vertically smoother trend profiles and quite similar zonal mean trends, with RICH-obs showing slightly less stratospheric cooling in the southern extratropics. The RICH estimates in the Southern Hemisphere, also near Antarctica, look much more reasonable than they did in the first version of RICH and RAOBCORE, where Haimberger et al. (2008) found that RICH showed only slightly more warming than unadjusted data and RAOBCORE adjusted trends had unrealistic vertical variations.

Figure 11a shows tropical belt mean vertical trend profiles of unadjusted data, ERA-Interim analysis data, RAOBCORE adjusted temperatures, and individual profiles from the RICH-obs and RICH-τ ensembles. In Fig. 11b the RICH data are presented as quantile plots that contain both RICH-obs and RICH-τ ensembles. The belt means for the tropics show a robust upper tropospheric warming maximum that is gentler and looks more plausible than the RAOBCORE maximum.

The RICH profiles also look more plausible than those from ERA-Interim, which shows a strong warming maximum at 100 hPa. This maximum is likely related to a problem in the assimilation of satellite data near this level in the early 1980s since the maximum disappears when analyzing trends from periods starting five years later.

The impact of using breakpoints from different RAOBCORE versions on RICH is relatively small, as can be seen from Figs. 11c and 11d. Figure 11c shows the profiles of RAOBCORE and RICH if no metadata and no background adjustment in 1971–78 are applied. There is slightly stronger cooling in the stratosphere and a less pronounced warming maximum in the troposphere.

The trend profiles from RAOBCORE v1.4 had quite different shapes compared to RAOBCORE v1.5 (cf. Figs. 11b and 11d), with more warming in the upper troposphere and less cooling in the stratosphere. The different breakpoints have some influence on the RICH trends as well but these show less variation than is the difference between RAOBCORE versions 1.4 and 1.5. This underlines the robustness of both RICH-obs and RICH-τ to moderate changes in the location of breakpoints.

The inset parameter Amp is the ratio between surface trend and the maximum trend of the ensemble mean in the troposphere. For the RAOBCORE v1.5 based profiles it is 1.4 whereas for the RAOBCORE v1.4 based profiles it is 1.3. Both values are within the uncertainty range of climate models (see Santer et al. 2005) although these often predict even stronger amplification. They are also in accord with evaluations at ECMWF (A. J. Simmons 2011, personal communication) showing that the amplification factor tropical surface temperature variability (not trends) for 300-hPa temperatures in the ERA-Interim reanalysis for this period is 2.2. Haimberger et al. (2012, manuscript submitted to *Meteor. Z.*) demonstrate that this warming amplification seen here in the interval 1979–2011 is persistent and often stronger in practically all 21-yr periods since 1960. This is new evidence that amplification of surface trends in the tropics, which has been subject of debate for 20 years (Thorne et al. 2011b; Douglass et al. 2008; Santer et al. 2008), is real. An earlier analysis of 21-yr periods only a few years ago (Thorne et al. 2007) yielded clearly smaller amplification factors. Note also, however, that this amplification factor is highly dependent also on uncertainties in the surface datasets. For example, Kennedy et al. (2011b) specify about 0.05 K decade^{-1} global mean marine temperature trend uncertainty for 30-yr trends in recent periods.

The vertical trend profiles 1979–2011 in other than tropical regions, now using again RAOBCORE v1.5 and the corresponding RICH ensembles, are shown in Fig. 12. RICH adjusted data show more warming than ERA-Interim at the lowest levels, generally fitting excellently to the Hadley Centre–Climatic Research Unit temperature dataset (HadCRUT3v; Brohan et al. 2006) surface trends. They show more cooling than ERA-Interim above 100 hPa, which is attributed to still unadjusted breaks in RICH and to the introduction of COSMIC data in ERA-Interim from 2007 onward. This event has a clear effect in the LS temperatures of ERA-Interim, as shown in Figs. 7 and 9.

Figure 13 gives an overview of trends for four MSU layers in the tropics and for the globe. The values can be compared with those in Blunden et al. (2011), although these are valid for 1979–2010. The RICH ensembles fit very well with ERA-Interim at the MT layer and show more warming than other radiosonde datasets and the reanalysis in the LT layer. In general, RICH shows more warming than UAH and fits very well to RSS in the LT and MT. Trends from STAR v2.0 show the most pronounced warming of all datasets at the MT and TS layers. Trends for the TS layer are not available over the 1979–2011 period from RSS and UAH.

In the LS the RICH estimates show more cooling than reanalysis and satellite datasets, but less cooling than HadAT data. The originally large gap between unadjusted radiosonde data and satellite data has been reduced considerably. Only RAOBCORE estimates lie within the uncertainty bounds given by RSS and fit well to the estimates by UAH. This agreement should not be overinterpreted, however, since RAOBCORE estimates are not independent of satellite data.

The spread of the RICH ensemble is rather small, particularly in the global mean. It is likely that the ensemble generated from the sensitivity experiments underestimates the true uncertainty. Little spread has been generated particularly in the Northern Hemisphere, as can be seen also in Fig. 12b). Nevertheless it is encouraging that at least three datasets now provide improved uncertainty guidance through ensemble methods.

6. Discussion and conclusions

This paper described improvements on radiosonde temperature homogenization made with the RAOBCORE and RICH homogenization methods, where RICH has been described in some detail. Both methods utilize background departure statistics available from climate data assimilation systems such as ERA-40 (Uppala et al. 2005). The method used for break detection (RAOBCORE) has already been described by Haimberger (2007). While RAOBCORE uses the background forecasts also for break size estimation, RICH estimates the breaks by comparison with reference series generated from surrounding radiosonde

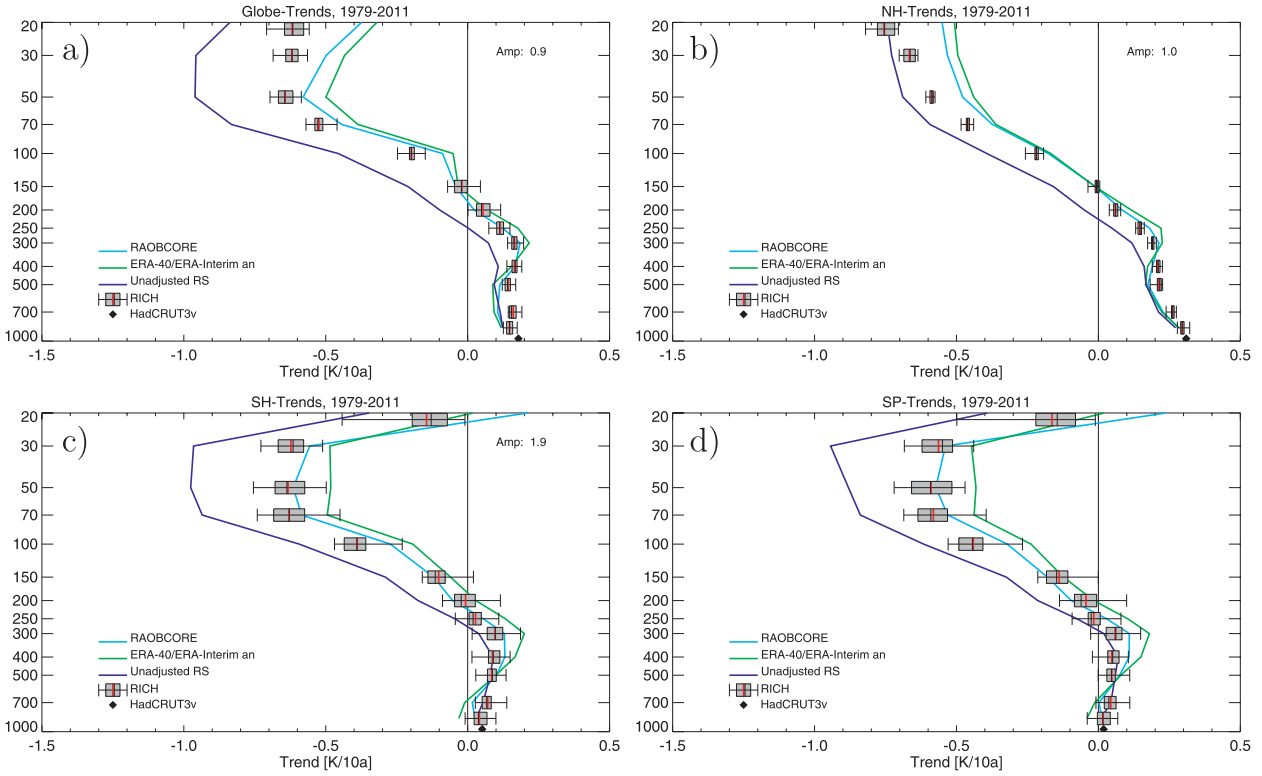

FIG. 12. Whisker plots of global belt mean trends: (a) globe, b) extratropical Northern Hemisphere (>20°N), (c) extratropical Northern Hemisphere (>20°S), and (d) southern polar region (>60°S).

stations. Reference series are generated either with (i) radiosonde observations (RICH-obs) or (ii) background departures of neighboring radiosonde stations (RICH-τ). While RICH-obs estimates are independent of the background and thus of satellite information, RICH-τ seems to have advantages for individual break size estimation. RICH-τ achieves the best overall spatiotemporal consistency of trend estimates in the satellite era. In the presatellite era, RICH-obs has the best spatiotemporal consistency. In the zonal belt means it is hard to tell at the present stage whether RICH-obs or RICH-τ yield more accurate results. Thus, using RICH-obs seems advantageous for now since the break size estimation process is independent of satellite data.

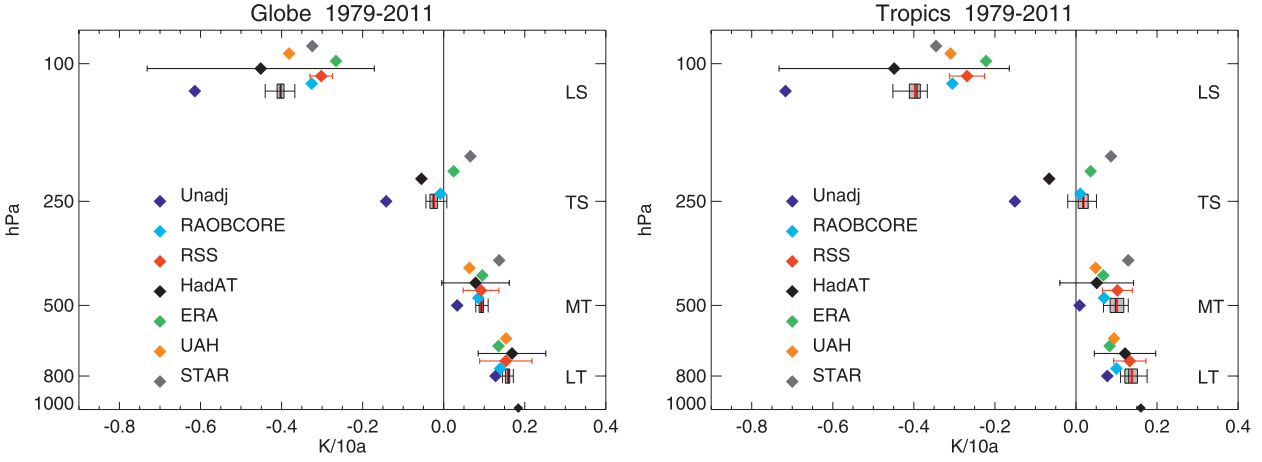

FIG. 13. Global and tropical belt mean MSU equivalent trends 1979–2011 from all datasets used in this study. HadCRUT3v surface trend is 0.12 K decade^{-1}. The gray whiskers depict the combined RICH-obs/RICH-tau v1.5 ensemble. Vertical shifts between markers are solely for better readability. RSS, UAH, and HadAT ensemble trends are not available for TS layer. MSU-STAR trends are not available for LT.

It was shown that the radiosonde temperatures adjusted by either RICH-obs or RICH-τ are more consistent with independent satellite estimates than other homogenized radiosonde temperature datasets. In sensitivity experiments and a few illustrative examples it has been shown that

1) RICH-adjusted radiosonde data reveal the upper tropospheric warming maximum projected by climate models and in most cases lead to spatially more consistent trend patterns than RAOBCORE in both the presatellite and satellite period.
2) In the satellite era adjusted trends still show less warming/more cooling than RSS and STAR satellite datasets, especially in the period 1979–99 and in the LS layer. In the period 1999–2011 they show more warming/less cooling than RSS/UAH.
3) In the presatellite era, the uncertainties become larger but all three adjustment methods do a good job in making the radiosonde series more consistent. It can be expected that rejection rates of the adjusted radiosonde data will be much smaller when assimilated with a climate data assimilation system.
4) RICH-τ is only slightly more sensitive to errors in the ERA-40 background than RICH-obs. In most cases it could be seen that RICH-τ time series and trends are between RICH-obs and RAOBCORE.
5) It is essential to have a good breakpoint database and to avoid averaging across breakpoints.

The performance of the RAOBCORE/RICH adjustment system suggests that the use of background departures from climate data assimilation systems helps to improve the original observation datasets. The next candidate is radiosonde wind, where earlier studies showed promising results (Gruber and Haimberger 2008); others are tropospheric humidity and surface parameters in remote areas. It is expected that the presented methodology will also help improving pre-1958 upper air data, as they are collected in the ongoing ERA-CLIM project (http://www.era-clim.eu). Thousands of plots and data as well as additional documentation can be found online at http://www.univie.ac.at/theoret-met/research/raobcore/.

Acknowledgments. This work has been funded by projects P18120-N10 and P21772-N22 of the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF), as well as by the EU 7th Framework Programme collaborative project ERA-CLIM (Grant 265229). The authors thank the anonymous reviewers for their constructive comments. RTTOVS is software developed within the NWP SAF. The HadCRUT3v dataset is produced by the University of East Anglia; HadAT is provided by the Met Office. MSU comparison data are produced by Remote Sensing Systems Inc., NESDIS, and University of Huntsville, Alabama, sponsored by the NOAA Climate and Global Change Programme.

REFERENCES

Alexandersson, H., and A. Moberg, 1997: Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *Int. J. Climatol.,* **17,** 25–34.

Allen, R. J., and S. C. Sherwood, 2008: Warming maximum in the tropical upper troposphere deduced from thermal winds. *Nat. Geosci.,* **1,** 399–403.

Andrae, U., N. Sokka, and K. Onogi, 2004: The radiosonde temperature bias correction in ERA-40. ERA-40 Project Report Series, Vol. 15, 34 pp.

Blunden, J., D. S. Arndt, and M. O. Baringer, 2011: State of the Climate in 2010. *Bull. Amer. Meteor. Soc.,* **92,** S1–S266.

Brohan, P., J. J. Kennedy, I. Harris, Tett S. F. B., and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new dataset from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

Caussinus, H., and O. Mestre, 2004: Detection and correction of artificial shifts in climate series. *J. Roy. Stat. Soc.,* **53C,** 405–425.

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.

Compo, G. P., and Coauthors, 2011: The twentieth century reanalysis project. *Quart. J. Roy. Meteor. Soc.,* **137A,** 1–28, doi:10.1002/qj.776.

Courtier, P., and Coauthors, 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quart. J. Roy. Meteor. Soc.,* **124,** 1783–1807.

Dai, A., J. Wang, P. W. Thorne, D. E. Parker, L. Haimberger, and X. L. Wang, 2011: A new approach to homogenize daily radiosonde humidity data. *J. Climate,* **24,** 965–991.

Dee, D. P., and S. M. Uppala, 2009: Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. *Quart. J. Roy. Meteor. Soc.,* **135,** 1830–1841.

——, P. Poli, and A. J. Simmons, 2011a: Extension of the ERA-Interim reanalysis to 1979. *ECMWF Newsletter,* No. 128, ECMWF, Reading, United Kingdom 7. [Available online at http://www.ecmwf.int/publications/newsletters/pdf/128.pdf.]

——, and Coauthors, 2011b: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.,* **137,** 553–597, doi:10.1002/qj.828.

Della-Marta, P., and H. Wanner, 2006: A method for homogenizing the extremes and mean of daily temperature measurements. *J. Climate,* **19,** 4179–4197.

Douglass, D. H., J. R. Christy, B. D. Pearson, and S. F. Singer, 2008: A comparison of tropical temperature trends with model predictions. *Int. J. Climatol.,* **28,** 1693–1701, doi:10.1002/joc.1651.

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

Ebita, A., and Coauthors, 2011: The Japanese 55-year reanalysis ''JRA-55'': An interim report. *SOLA,* **7,** 149–152, doi:10.2151/sola.2011-038.

Free, M., D. J. Seidel, J. K. Angell, J. Lanzante, I. Durre, and T. C. Peterson, 2005: Radiosonde atmospheric temperature products for assessing climate (RATPAC): A new data set of large-area anomaly time series. *J. Geophys. Res.,* **110,** D22101, doi:10.1029/2005JD006169.

Gaffen, D. J., 1996: A digitized metadata set of global upper-air station histories. NOAA Tech. Memo. ERL ARL-211, 37 pp.

Grant, A., S. Brönnimann, and L. Haimberger, 2008: Recent arctic warming vertical structure contested. *Nature,* **455,** E2–E3, doi:10.1038/nature07257.

Gruber, C., and L. Haimberger, 2008: On the homogeneity of radiosonde wind time series. *Meteor. Z.,* **17,** 631–643.

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

——, and U. Andrae, 2011: Radiosonde temperature bias correction in ERA-Interim. ERA Rep. Series, Vol. 8, 17 pp.

——, 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.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011a: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res.,* **116,** D14103, doi:10.1029/2010JD015218.

——, ——, ——, ——, and ——, 2011b: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res.,* **116,** D14104, doi:10.1029/2010JD015220.

Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel, 2006: World map of the Köppen-Geiger climate classification updated. *Meteor. Z.,* **15,** 259–263.

Ladstätter, F., A. K. Steiner, U. Foelsche, L. Haimberger, C. Tavolato, and G. Kirchengast, 2011: An assessment of differences in lower stratospheric temperature records from (A)MSU, radiosondes and GPS radio occultation. *Atmos. Meas. Tech.,* **4,** 1965–1977.

Lanzante, J. R., S. A. Klein, and D. J. Seidel, 2003: Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities, and MSU comparison. *J. Climate,* **16,** 241–262.

Lewis, J. M., S. Lakshmivaran, and S. Dhall, 2005: *Dynamic Data Assimilation.* Cambridge University Press, 816 pp.

Luers, J. K., and R. E. Eskridge, 1995: Temperature corrections for the VIZ and Vaisala radiosondes. *J. Appl. Meteor.,* **34,** 1241–1253.

McCarthy, M. P., 2008: Spatial sampling requirements for monitoring upper-air climate change with radiosondes. *Int. J. Climatol.,* **28,** 985–993, doi:10.1002/joc.1611.

——, H. Titchner, P. Thorne, 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, 2009: Construction of the Remote Sensing Systems V3.2 atmospheric temperature records from the MSU and AMSU microwave sounders. *J. Oceanic Atmos. Technol.,* **26,** 1040–1056.

——, ——, P. W. Thorne, and D. Bernie, 2011: Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte Carlo estimation technique. *J. Geophys. Res.,* **116,** D08112, doi:10.1029/2010JD014954.

Menne, M. J., and C. N. J. Williams, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate,* **22,** 1700–1717.

Nash, J., and F. J. Schmidlin, 1987: WMO international radiosonde intercomparison: Final report. WMO/TD 195, 103 pp.

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

Poli, P., S. B. Healy, and D. P. Dee, 2010: Assimilation of Global Positioning System radio occultation data in the ECMWF ERA-Interim reanalysis. *Quart. J. Roy. Meteor. Soc.,* **136,** 1972–1990.

Rienecker, M., and Coauthors, 2011: MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate,* **24,** 3624–3648.

Rougier, J., and D. M. H. Sexton, 2007: Inference in ensemble experiments. *Philos. Trans. Roy. Soc. London,* **365A,** 2133–2143, doi:10.1098/rsta.2007.2071.

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

——, and Coauthors, 2008: Consistency of modelled and observed temperature trends in the tropical troposphere. *Int. J. Climatol.,* **28,** 1703–1722, doi:10.1002/joc.1756.

Saunders, R., and Coauthors, 2011: RTTOV-10 science and validation report. EUMETSAT, 31 pp. [Available online at http://research.metoffice.gov.uk/research/interproj/nwpsaf/rtm/docs_rttov10/rttov10_svr_1.11.pdf.]

Screen, J. A., and I. Simmonds, 2011: Erroneous arctic temperature trends in the ERA-40 reanalysis: A closer look. *J. Climate,* **24,** 2620–2627.

Seidel, D. J., and Coauthors, 2009: Reference upper-air observations for climate: Rationale, progress, and plans. *Bull. Amer. Meteor. Soc.,* **90,** 361–369.

Sherwood, S. C., J. Lanzante, and C. Meyer, 2005: Radiosonde daytime biases and late 20th century warming. *Science,* **309,** 1556–1559, doi:10.1126/science.1115640309.

——, C. L. Meyer, R. J. Allen, and H. A. Titchner, 2008: Robust tropospheric warming as revealed by iteratively homogenized radiosonde data. *J. Climate,* **21,** 5336–5352.

Sperka, S., 2007: Homogeneity adjustments of radiosonde temperature time series using composites of innovations from ERA-40. M.S. thesis, Dept. of Meteorology and Geophysics, University of Vienna, 60 pp. [Available online at http://www.univie.ac.at/img-wien/dipldiss/dipl/DA_Sperka.pdf.]

Steiner, A. K., B. C. Lackner, F. Ladstädter, B. Scherllin-Pirscher, U. Foelsche, and G. Kirchengast, 2011: GPS radio occultation for climate monitoring and change detection. *Radio Sci.,* 46, RS0D24, doi:10.1029/2010RS004614.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears, 2005a: Uncertainties in climate trends: Lessons from upper-air temperature records. *Bull. Amer. Meteor. Soc.,* **86,** 1437–1442.

——, ——, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, and P. Brohan, 2005b: Revisiting radiosonde upper-air temperatures from 1958 to 2002. *J. Geophys. Res.,* **110,** D18105, doi:10.1029/2004JD005753.

——, and Coauthors, 2007: Tropical vertical temperature trends: A real discrepancy? *Geophys. Res. Lett.,* **34,** L16702, doi:10.1029/2007GL029875.

——, and Coauthors, 2011a: A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *J. Geophys. Res.,* **116,** D12116, doi:10.1029/2010JD015487.

——, J. R. Lanzante, and T. C. Peterson, D. J. Seidel, and K. P. Shine, 2011b: Tropospheric temperature trends: History of an ongoing controversy. *WIREs Climate Change,* **2,** 66–88, doi:10.1002/wcc.80.

Titchner, H., M. McCarthy, P. W. Thorne, S. F. B. Tett, L. Haimberger, and D. E. Parker, 2009: Critically reassessing tropospheric

temperature trends from radiosondes using realistic validation experiments. *J. Climate,* **22,** 465–485.

Trenberth, K. E., and Coauthors, 2007: Observations: Surface and atmospheric climate change. *Climate Change 2007: The Physical Science Basis,* Cambridge University Press, 235–336.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Venema, V. K. C., and Coauthors, 2012: Benchmarking monthly homogenization algorithms. *Climate Past,* **8,** 89–115.

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

Zou, C. Z., M. Gao, and M. D. Goldberg, 2009: Error structure and atmospheric temperature trends in observations from the Microwave Sounding Unit. *J. Climate,* **22,** 1661–1681.