

# HOMOGENIZATION OF RADIOSONDE TEMPERATURE TIME SERIES USING INNOVATION STATISTICS

LEOPOLD HAIMBERGER<sup>1</sup>

submitted to J. Climate 13 March 2006

1<sup>st</sup> revision 29 June 2006

in final form 02 August 2006

---

<sup>1</sup>DEPARTMENT OF METEOROLOGY AND GEOPHYSICS, UNIVERSITY OF VIENNA, ALTHANSTRASSE 14, A-1090  
VIENNA, AUSTRIA  
*Email address:* leopold.haimberger@univie.ac.at

### Abstract

Radiosonde temperature records contain valuable information for climate change research from the 1940s onwards. Since they are affected by numerous artificial shifts, time series homogenization efforts are required. This paper introduces a new technique that uses time series of temperature differences between the original radiosonde observations (obs) and background forecasts (bg) of an atmospheric climate data assimilation system for homogenization.

These obs-bg differences, the "innovations", are a by-product of the data assimilation process. They have been saved during the ECMWF reanalysis ERA-40 and are now available for each assimilated radiosonde record back to 1958. It is demonstrated that inhomogeneities in the obs time series due to changes in instrumentation can be automatically detected and adjusted using daily time series of innovations at 00GMT and 12GMT.

The innovations not only reveal problems of the radiosonde records but also of the data assimilation system. Although ERA-40 used a frozen data assimilation system, the time series of the bg contains some breaks as well, mainly due to changes in the satellite observing system. It has been necessary to adjust the global mean bg temperatures before the radiosonde homogenization.

After this step, homogeneity adjustments, which can be added to existing raw radiosonde observations, have been calculated for 1184 radiosonde records. The spatiotemporal consistency of the global radiosonde dataset is improved by these adjustments and spuriously large day-night differences are removed. After homogenization the climatologies of the time series from certain radiosonde types have been adjusted. This step reduces temporally constant biases which are detrimental for reanalysis purposes. Therefore the adjustments applied should yield an improved radiosonde dataset that is suitable for climate analysis and particularly useful as input for future climate data assimilation efforts. The focus of this paper relays on the lower stratosphere and on the internal consistency of the homogenized radiosonde dataset. Implications for global mean upper air temperature trends are touched upon only briefly.

## 1. INTRODUCTION

Since the 1940s radiosondes are an essential component of the global atmospheric observing system. They reach farther back than satellite records and they also provide relatively high vertical resolution compared to satellite instruments such as the Microwave Sounding Unit (MSU). Therefore they are an unique source of information about the upper air climate and they are essential for climate data assimilation efforts such as ERA-40 or the NCEP/NCAR reanalysis (Uppala et al. 2005; Kistler et al. 2001). While radiosondes measure temperature, humidity and wind, this work's analysis is restricted to temperature. The quality of the radiosonde instrumentation has improved over time. Systematic errors of radiosonde temperature measurements comprised several K in the stratosphere during the 1960s and 1970s. The main but not the only reason for systematic errors were radiation effects. In the late 1980s, at many sites the radiation error was still larger than 1 K at the 50 hPa level, as is indicated in Fig. 1. With today's modern sounding equipment the radiation error is reduced to a few tenths of a K (Nash et al. 2005).

Fig. 2 shows time series of mean 12GMT-00GMT differences of composites of radiosondes located between 30W and 40E as well as between 120E and 120W at the 50 hPa level. The 50 hPa level has been chosen as a compromise between data availability and susceptibility to radiation errors. Between 30W and 40E there is darkness at 00GMT and high solar elevation at 12GMT (except at polar regions). Between 120E and 120W the opposite is the case. A larger longitude range has been chosen in Fig. 2-b to include more stations around the Pacific. In recent years the mean differences, which exceed 1K in the 1960s, are gradually reduced at this altitude to practically zero. The resulting "trend" in the 12GMT-00GMT differences is almost certainly artificial (Sherwood et al. 2005) and must be removed before the radiosonde data can be applied for climate change research. Breaks and biases also cause problems in operational data assimilation systems as well as in climate data assimilation systems: Observations with large bias but otherwise good quality tend to be rejected more frequently by a data assimilation

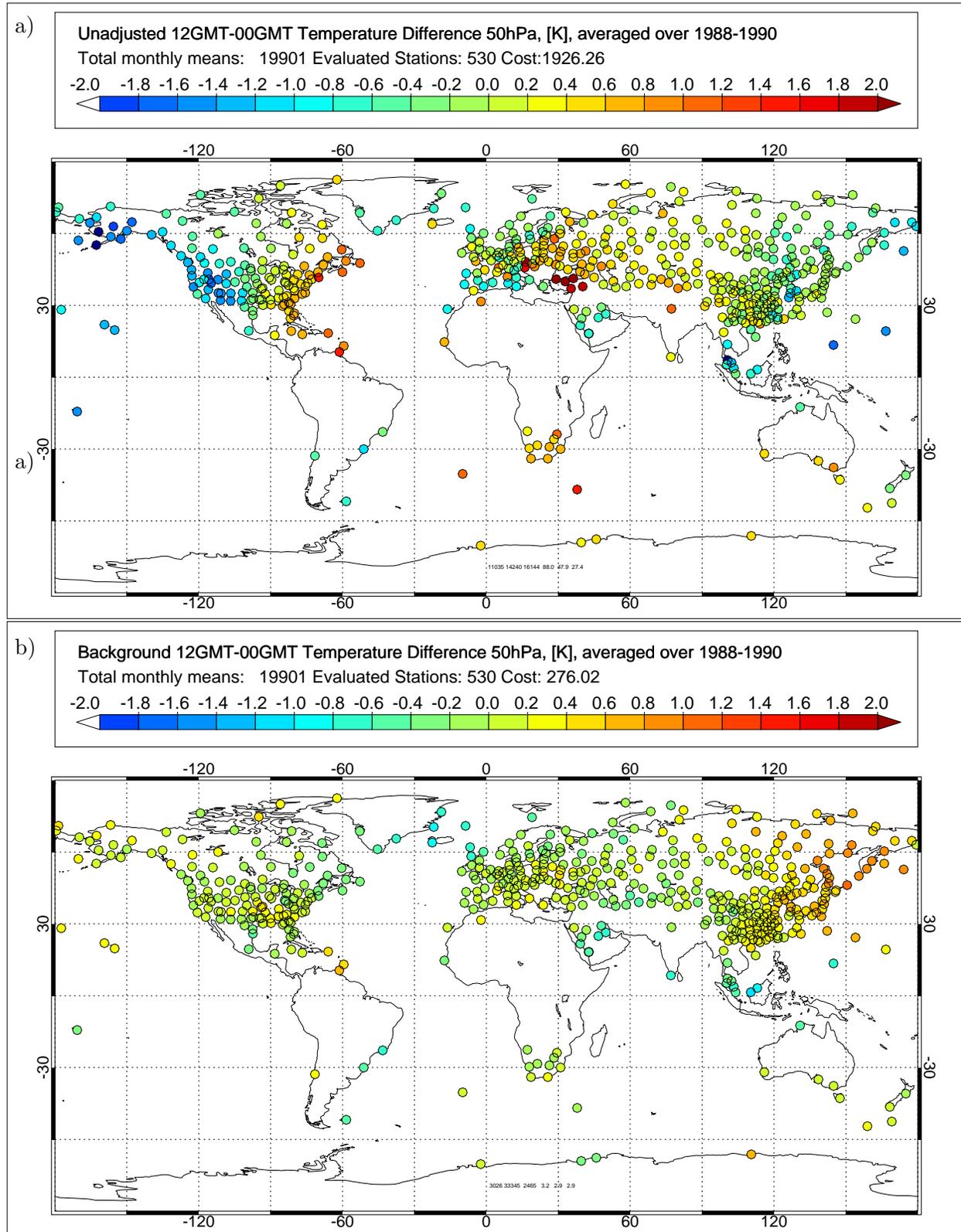
system. If they are not rejected they cause biases in the resulting analyses unless the data assimilation system is specifically designed to deal with biased observations (Dee and Da Silva 1998).

The task of removing artificial breaks from time series is referred to as *homogenization*. Some authors attempted to use the available metadata (Gaffen 1996) and detailed knowledge of the instruments to apply physically based corrections (Luers and Eskridge 1995; Eskridge et al. 2003; Redder et al. 2004). While this is potentially the most favorable approach, it can be applied only at selected sites since the required detailed information about equipment and launch times is often not available.

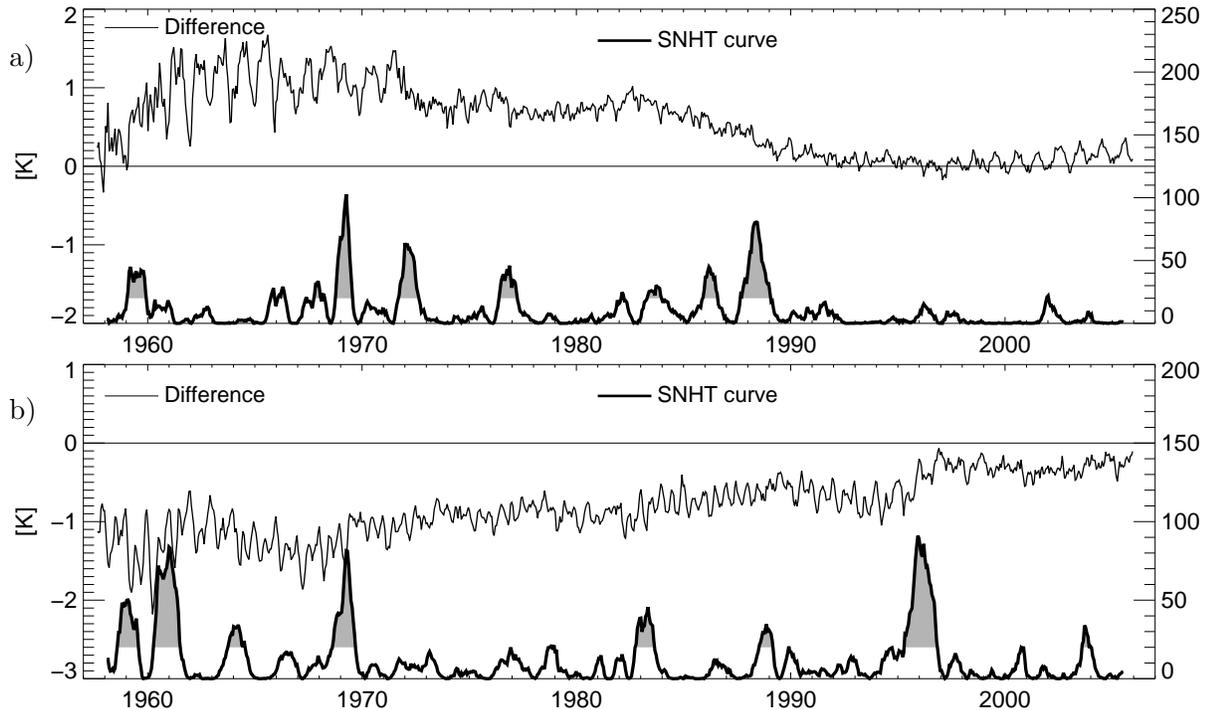
Homogeneity adjustments based on time series analysis, metadata and expert judgement have been published by Lanzante et al. (2003a); Lanzante et al. (2003b); Free et al. (2005) (referred to as LKS) and Thorne et al. (2005a). While these datasets consider only a subset of the global radiosonde network (87 stations by LKS, 678 stations by Thorne et al. 2005a), they are important achievements and valuable tools for intercomparisons with satellite data and with climate model results.

Subsampling of the data, as performed in these analyses, is acceptable if the desire is to characterise only large-scale changes. For climate data assimilation purposes it is much harder to justify omission of at least one third of the available radiosonde data in order to retain temporal homogeneity of the analyzed product. Since only anomaly time series have been adjusted, many records of these homogenized datasets may still have a constant bias and therefore may cause a bias in reanalyses if used as input for data assimilation. Both aspects limit their value as direct input for future reanalyses, although the information on breakpoint timing and magnitude that they contain may prove useful.

None of the above datasets has been created by automatic procedures. Any improvement of the datasets would be rather laborious. Further there is often lack of agreement where and how large the breaks are (see Free et al. 2002). Sizeable uncertainty therefore still remains on upper air trends and low frequency variability (Seidel



**Figure 1:** 12GMT-00GMT temperature differences at 50 hPa averaged over period 1988-1990. Each bullet denotes a radiosonde station with more than 30 out of 36 months of data; colour of bullets indicates the difference. “Cost“ refers to a spatial consistency measure defined in appendix A. Panel a) shows the differences as measured by radiosondes. Panel b) shows differences from the ERA-40 bg. These are spatially much more homogeneous.



**Figure 2:** Time series of composite mean 12GMT-00GMT differences from radiosondes, a) between 30W and 40E, b) between 120E and 120W. Only stations and days with both 00GMT and 12GMT data have been included in composite. Thin curve is 12GMT-00GMT difference, thick curve is time series of Standard Normal Homogeneity Test (SNHT) statistic as defined in eq. 4. Peaks above 20 in SNHT test statistic are shaded. They indicate abrupt changes in the difference series.

In panel a) the peaks in 1969, 1972 are caused by changes in the French/Russian radiosonde network. The peaks in 1988/89 are related to the change to Vaisala RS80 radiosondes at many sites. The main peaks in panel b) are caused by changes in the Russian and Japanese radiosonde networks (1969) and the degradation of the Russian radiosonde network (ca. 1996), which had relatively large day-night differences.

et al. 2004; Free and Seidel 2005; Thorne et al. 2005b). Sherwood et al. (2005); Randel and Wu (2006) and Santer et al. (2005) have recently raised serious doubts about the validity of temperature trends from radiosondes in the tropics. Reduced uncertainty in observed upper air trends has been identified as one of the most pressing needs in climate research (Karl et al. 2006)

This article describes a new homogeneity adjustment method, called **RA**diosonde **O**Bservation **C**ORrection using **RE**analyses (**RAOBCORE**). It addresses some of the problems of the existing homogenized datasets. A preliminary version of **RAOBCORE** has been documented in Haimberger (2005). The most important characteristics of **RAOBCORE** are:

- It uses time series of innovation statistics (the difference between observation and the background forecast of the assimilating model) of a climate data assimilation system such as the European Centre for Medium-Range Weather Forecast (ECMWF) reanalysis ERA-40 (Uppala et al. 2005). The rationale for using ERA-40 innovation statistics for homogenization is discussed in section 2.
- it uses the most recent and most complete radiosonde datasets available: The Integrated Global Radiosonde Archive (IGRA, Durre et al. 2006) and ERA-40 (see section 3)
- it uses time series of individual launches, not monthly or seasonal means, for break detection and adjustment, and

it analyzes daytime and nighttime launches separately. The homogeneity adjustment method is described in section 4

- A record with no breaks may still have a large constant bias, which is problematic for reanalysis efforts. In section 7 it is outlined how RAOBCORE adjusts not only breaks of long records but also the biases of the most recent parts of the records.
- It is relatively easy to create multiple realizations of the homogenized dataset, e.g. for sensitivity experiments, because the adjustment software is fully automatic.

RAOBCORE uses innovations from a climate data assimilation system as reference instead of composites of neighbouring radiosondes (c.f. Thorne et al. 2005a) and makes no reference to the LKS dataset. Consequently it can be regarded as independent of these homogenization methods. The homogeneity properties of the ERA-40 background temperature time series are discussed in section 6 and in appendix B. Adjustment results for selected stations and for several sensitivity experiments are presented in sections 8 and 9.

## 2. INTERPRETATION OF THE INNOVATION STATISTICS OF A DATA ASSIMILATION SYSTEM

The atmospheric data assimilation process may be described as applying a filter on the multivariate time series of meteorological observations. The filtering process is optimal if the time series of differences between observations (*obs*) and the background forecast (*bg*) of the assimilating model is stationary random with zero mean and zero autocorrelation (Lewis et al. 2005). The *obs*-*bg* differences are often referred to as *innovations* and the time series of the differences is called innovation process.

In practice the filtering process is never optimal, due to systematic errors of both the observing system and the assimilating model. As a result the time series of innovations have nonzero long term means, are not stationary and not random. These nonzero means of the innovations are often referred to as *bias*. One can only

estimate the bias of one (test) dataset with respect to a second (reference) dataset. Since the mean of the true state is not known, the real bias is also unknown. However, if there is a shift in the observation time series, e.g. due to an instrument change, it should immediately result also in a shift in the innovation time series and thus of the bias. This shift can be estimated and used for homogenization purposes.

For this study mainly the innovations from the 3D-VAR data assimilation system used in ERA-40 have been used. Within the 6 hour cycle the following cost function is minimized (Courtier et al. 1998):

$$J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) + (\mathbf{y} - \mathbf{H}\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\mathbf{x}) \quad (1)$$

$\mathbf{x}$  is the state vector of the assimilating forecast model.  $\mathbf{x}_b$  is the background state, which is obtained by a 6h forecast run of the assimilating model. Then the forecast state mapped to observation space by the observation operator  $\mathbf{H}$  is compared with the available observations (the observation vector  $\mathbf{y}$ ).  $\mathbf{B}$  and  $\mathbf{R}$  are the background error and observation error covariance matrices, respectively.

At the beginning of the minimization process,  $\mathbf{x} = \mathbf{x}_b$ , thus the first term vanishes. The quantity  $(\mathbf{y} - \mathbf{H}\mathbf{x}_b)$  is called the *innovation* or *obs - bg*-difference. Its accurate calculation is essential in the data assimilation process and much effort is put into the specification of the observation operator  $\mathbf{H}$ . The difference  $(\mathbf{y} - \mathbf{H}\mathbf{x}_b)$  is available for every single observation presented to the data assimilation system and is saved in the so-called ERA-40 analysis feedback files. In the case of radiosonde temperatures,  $\mathbf{H}$  represents the transformation from the spectral space to the model grid and then simple linear interpolation from the ECMWF model grid to the observation location (ECMWF 2000).

The innovations have proven most useful for the detection and estimation of systematic observation errors (Hollingsworth et al. 1986). Monitoring of the innovations is an important task at operational forecast centres. The innovations have been used also in ERA-40 to calculate adjustments of the radiation error of radiosondes (Onogi 2000; Andrae et al. 2004).

The innovations are more useful for bias estimation than differences between observations and the analyses, since the obs-bg differences are more directly related to systematic observation errors. If the bg were perfect, the obs-bg difference would be the obs error, whereas the obs-an difference is already influenced in a complicated way by the erroneous observation (see e.g. Dee 2004).

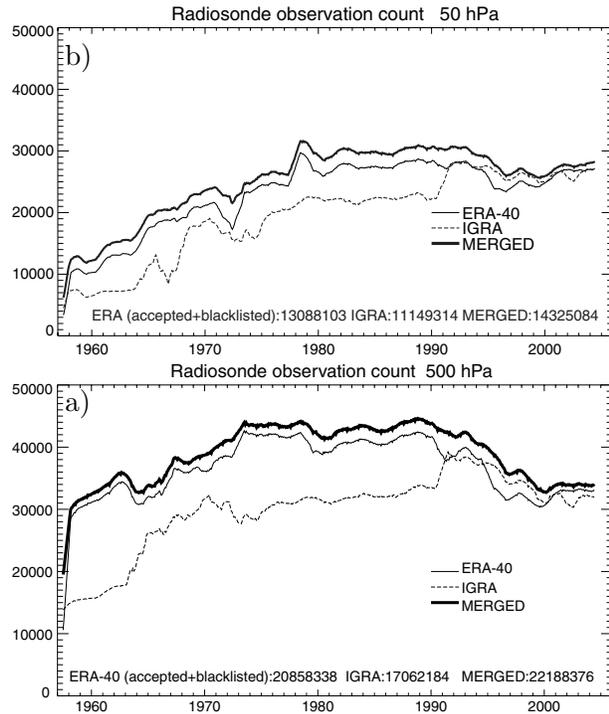
It is the working hypothesis of this paper that time series of innovations can be used to detect and remove artificial shifts in radiosonde temperature time series, i.e. to *homogenize* these time series. Since the bg error is generally quite small, the differences between the bg and the corresponding observations tend to be small and therefore the series of differences between bg and obs reacts sensitively to any change in the radiosonde temperature bias.

Small random errors are only one of the necessary ingredients for successful homogenization. The bg forecast time series must also be temporally homogeneous and independent of the radiosonde records to be tested. These aspects are discussed in more detail in sections 6 and 8.

### 3. INPUT DATA

With the completion of the ERA-40 project (Uppala et al. 2005), a 45year global time series of analyses (an) and 6h background forecasts (bg) has become available. While the analyses are the most useful product for many research applications, the observation database assembled during the reanalysis effort is equally valuable.

The BUFR-coded ERA-40 analysis feedback (AF) dataset has been the main input data source for this study. The AF dataset contains all observations from 1958-2002 presented to the ERA-40 data assimilation system, plus quality control flags and the innovations described in section 2. No other dataset holds such long and complete time series of upper air temperature innovations. The time series have daily resolution and 00GMT and 12GMT ascents were kept separate. Observations at 16 standard pressure levels (10, 20, 30, 50, 70, 100, 150, 200, 250, 300, 400, 500, 700, 850, 925, 1000 hPa) were considered and analysed for breakpoints.



**Figure 3:** Time series of monthly global radiosonde temperature observation count 1958-2004, a) at the 10 hPa level and b) at the 500 hPa level. Only TEMP-Land observations have been included. TEMP-Ship observations, which are also part of the ERA-40 archive, have been excluded in this plot but are used in RAOBCORE. Thin dashed: count from IGRA dataset. Thin solid: count in ERA-40 archive. Thick: merged ERA-40+IGRA dataset. Numbers at the bottom are total observation counts over 47 years.

Since the ERA-40 dataset ends in 2002, it has been decided to use operational analysis feedback data from 2001 onwards. Concatenating operational and ERA-40 AF has been allowed the use of radiosonde records up to December 2005. The obs-bg differences need to be adjusted since the bg of the operational version of the ECMWF data assimilation system used for 2001 and later years differs substantially from the bg of the ERA-40 assimilation system (see section 6). After 2001 no evidence for major temperature inhomogeneities in the bg temperatures due to changes in the operational ECMWF data assimilation system could be found.

The recent advent of the IGRA dataset (Durre et al. 2006) has been the second big improvement of the data basis for homogenization of the global radiosonde network. The IGRA and ERA-40/ECMWF datasets are not identical, as can be seen from Figure 3. ERA-40 contains more data in the 1960s and 1970s, whereas IGRA contains more data in the 1990s.

The IGRA data complement ERA-40 radiosonde data at some places, especially over the U.S in the 1990s. The data missing in IGRA are partly available in the older CARDS dataset (Durre et al. 2006). They have not been included in IGRA because of data quality concerns. Experience from ERA-40 indicates, however, that most of these data are of sufficient quality to be assimilated. While the original IGRA data do not contain obs-bg differences, these can be calculated quite accurately from archived ERA-40 bg fields interpolated to the IGRA observation sites (Haimberger 2005).

In this paper, the union of both datasets is used. The merging procedure has been simple: if there were duplicates, preference has been given to ERA-40 data. A total of 2881 stations (most of them on land but also a few weather ships) have been identified. These have time series containing temperature on standard pressure levels and on significant levels (TEMP-Land and TEMP-Ship). The station identification procedure which normally makes merging so complicated has been facilitated by the good documentation of IGRA data and the station tables available from ERA-40.

Apart from pure measurement information and the innovation statistics, also metadata information is available in digitized form from the CARDS archive (Gaffen 1996, with updates from Aguilar 2000). The information about the radiosonde type and the radiation corrections used is useful for finding and interpreting breaks in the radiosonde temperature time series, although the information is incomplete and sometimes inaccurate.

Quite useful additional metadata information could be extracted from the ERA-40 feedback records. From 1979 onwards, many stations have routinely transmitted the radiosonde type. This information is available in the ERA-40 feedback files as well (coded according to

WMO-BUFR table 02011; see WMO Manual on Codes, No 306, Volume I.2). From this information more than 2000 well documented and often precisely dated radiosonde type changes could be extracted.

#### 4. BREAK DETECTION METHOD

The obs-bg difference series and obs(12GMT)-obs(00GMT) difference series have been analyzed with a variant of the Standard Normal Homogeneity Test (SNHT, Alexandersson and Moberg 1997) described below. Ducre-Robetaille et al. (2003) recently compared popular homogeneity tests and the SNHT performed well in this intercomparison. However, the results of this comparison are not directly applicable for the analysis of radiosonde records. The frequency of breaks requires analysis windows of only a few (<5) years. For such short time windows it is essential to take the seasonal cycle into account and it is advisable to use daily ascents instead of monthly or annual means in order to have full control over the sampling of the available data. It is worth noting that using anomalies does not remove the annual cycle in the case of radiosonde data since the annual cycles of e.g. 12GMT-00GMT differences at the same stations are rather different for different radiosonde types (see e.g. Fig. 10 below).

##### 4.1. A variant of the Standard Normal Homogeneity Test.

The original standard normal homogeneity test (SNHT, Alexandersson and Moberg 1997) calculates the series of differences between the tested series (in this study the radiosonde time series) and a reference series (the bg). Then the means of the parts of the difference series before and after a potential breakpoint  $k$  are compared. The point dividing the sample into two parts is varied but the time interval stays fixed. For each dividing point  $k$  in the sample, a test statistic can be calculated:

$$(2) \quad T_k = ((N - k)(\mu_{1k} - \mu)^2 + k(\mu_{2k} - \mu)^2) / \sigma$$

$N$  is the sample size,  $\mu_{1k}$  is the mean of the subsample before  $k$ ,  $\mu_{2k}$  is the mean of the subsample after  $k$ ,  $\mu$  is the mean of the whole sample

and  $\sigma$  is the sample standard deviation. The difference series is considered inhomogeneous if the maximum value  $T^s$  of all  $k$  is above the threshold somewhere in the interval. In this case the point  $k^s$  where  $T^s$  occurs is the most likely location for the breakpoint. The difference of  $\mu_{2k} - \mu_{1k}$  calculated at the point  $k^s$  is the best estimate for the magnitude of the break.

The original SNHT has two drawbacks. Firstly, it tends to indicate breaks near the edges of the time interval as either  $k$  or  $N - k$  get small (Ducré-Robetaille et al. 2003). Secondly the position of a breakpoint is poorly estimated in the presence of a periodic signal. To overcome these problems the original SNHT has been modified such that  $\mu_{1k}$  and  $\mu_{2k}$  are calculated as two-year moving averages:

$$(3) \quad T_k = \left( \frac{N}{2}(\mu_{1k} - \mu)^2 + \frac{N}{2}(\mu_{2k} - \mu)^2 \right) / \sigma_k$$

The sample sizes at point  $k$  are then fixed to  $N/2$  but the interval  $[k - N/2, k + N/2]$  now depends on  $k$ . The maximum value of the test statistic changes its meaning in the sense that it is not an absolute maximum of the  $T_k$  in a fixed interval but the local maximum of the  $T_k$  in the interval  $[k - N/2, k + N/2]$ . Therefore it is denoted  $T_k^s$ .

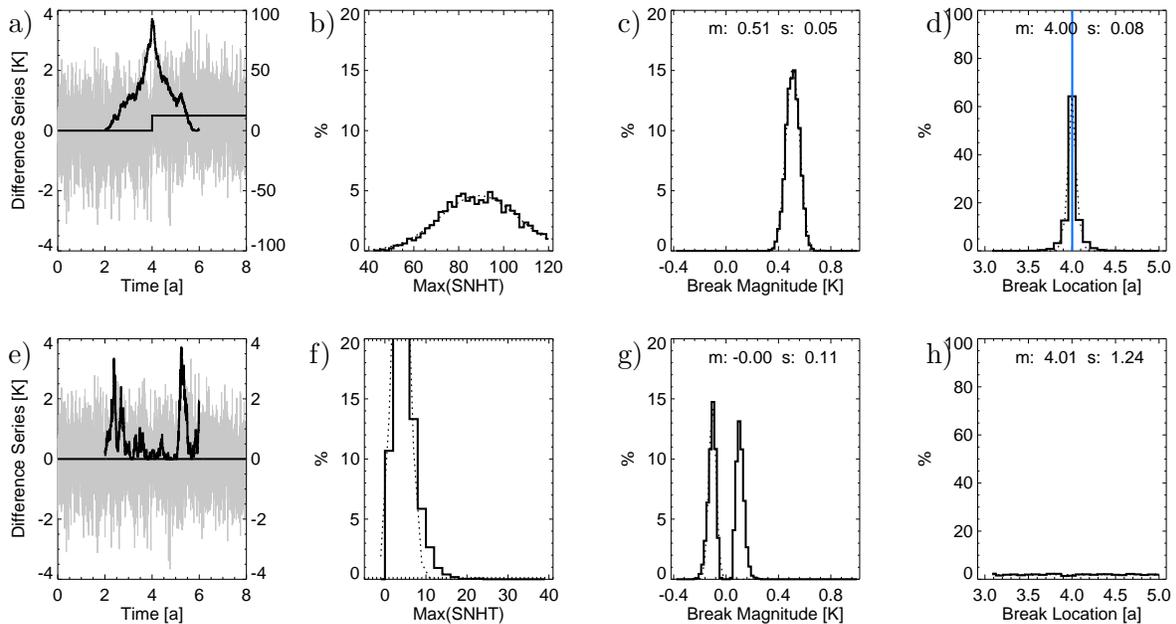
This way the means are more reliably estimated and the annual wave is averaged out exactly if  $N/2$  corresponds to an integer multiple of 365 if daily data are used and none are missing. In order to make the test more robust against combined effects of an annual cycle and missing data, data before and after the midpoint  $k$  of the analysed time interval have been placed into 12 bins, one for each month. If the data count of e.g. the January values in the earlier 2-year interval was less than the count of January values in the later 2-year interval, the excessive January values in the later interval were removed (starting with the values with the largest time distance from the middle of the investigated interval). This version of the SNHT, which ensures equal sampling of the annual cycle in the intervals before and after a potential breakpoint, is referred to as *equal sampling SNHT*.

Figure 4 shows examples of the behaviour of the moving average SNHT test statistic and related quantities with simulated data. The synthetic time series has 2920 data points, corresponding to 8 years if daily data are considered. The interval for the moving averages before and after a suspected breakpoint is 2 years, the same as used for the real data below. The grey curve in panel a) depicts one of 5000 realization of a  $N(0, 1)$  normal distributed random series with a break of size  $0.5\sigma$  in the middle of the interval. The black  $T^k$  curve in a) shows a distinct maximum at the right location. Panel b) shows the distribution of  $T_k^s$  gained from the 5000 realizations. Values are practically always above 50. Panel c) shows the distribution of the break estimates  $\mu_{1k^s} - \mu_{2k^s}$  at the locations  $k_s$  where  $T_k$  is largest. It is normally distributed around 0.5K with a standard deviation of 0.05K. Panel d) shows the distribution of the detected break locations  $k^s$ . The location of a break with size 0.5K can be detected with a standard deviation of 0.08 years (one month).

The significance levels for rejecting the null hypothesis can be gained from applying the SNHT to a sample of 5000 random series with no break. Fig. 4-e) shows a realization of such a series. The  $T_k$  vary erratically and  $T_k^s$  is much smaller. The significance levels of the moving average SNHT can be estimated from the distribution of the  $T_k^s$  in Fig. 4-f. For the 95% level it is 9.6, for the 99% level it is 12.5. Panel g) shows again the distribution of the difference between the moving averages at the point  $k^s$ . Panel h) shows the distribution of the break location  $k^s$  for the case of no breaks. It is close to uniform.

From this result, it becomes clear that breaks with a size of  $0.5\sigma$  are practically always detected since in this case more than 99% of the  $T_k^s$  reach values above 20. Panel g) indicates the problems to be expected if a break is falsely detected: the estimates break size and therefore the false correction would be of magnitude  $0.25\sigma$ .

The significance levels derived above are valid only for purely random processes (apart from an annual cycle) and for complete time series (1460 days). If data are missing, the  $T_k^s$  value necessary to reach the 95% significance



**Figure 4:** Moving average SNHT applied to random series with shifts of 0.5 K (panels a-d), 0 K (panels e-h) and a residual variance of  $1 \text{ K}^2$ . Panels a), e) show one realization of a random time series with break and the corresponding SNHT test statistic  $T_k$  (right axis). A time period of 8 years is simulated with a break right in the middle. The SNHT uses 2 year moving averages. Panels b), f) show histograms of the  $T_k^s$ , generated with 5000 different random series. Note different  $x$ -scales. Panels c), g) show distribution of diagnosed size  $\Delta T$  of break, panels d), h) show distribution of diagnosed break location  $k^s$ . The smooth curves are fitted Gauss functions. Inset numbers indicate sample means (m) and standard deviations (s).

level decreases, no matter whether the data loss occurs as one big gap or intermittently. This can be verified with Monte Carlo type experiments similar to those presented above (not shown) and has been documented by (Alexandersson and Moberg 1997) as well. If there is an annual cycle in the obs-bg, the significance threshold also decreases, since it generates additional variance. In these cases smaller thresholds for the maximum SNHT should be used.

Only if the time series are generated by stochastic (autoregressive) processes, the significance levels become higher since the degrees of freedom are reduced (von Storch and Zwiers 1999). A check of the obs-bg difference time series (with annual cycle removed) showed weak autocorrelation for lags of up to 4 days at some remote areas and practically zero autocorrelation for areas with good data coverage and for obs(12GMT)-obs(00GMT) time series. Data show generally much less autocorrelation than a

1st order autoregressive model with coefficient  $\alpha = 0.3$ , and even if such a model would apply, the 95% threshold is still below 20. An exception are remote island stations before the satellite era where the autocorrelation is similar to a 1st order autoregressive model with  $\alpha = 0.5$ , for which a Monte Carlo experiment similar to those in Fig. 4 yields a 95% significance threshold of 25. Additional sensitivity experiments with the equal sampling SNHT are documented in Haimberger (2005). It performs well in general as long as there is only one break within the tested interval. If there are more breaks only the largest break is detected.

To be safe, larger thresholds than  $T_k^s > 12.5$  suggested from the above analysis have been used for break detection. For obs-bg time series  $T_k^s$ -values above 50 are considered significant, if no metadata are available. For obs(12GMT)-obs(00GMT) time series, which show no autocorrelation and cannot be influenced by bg

errors, values above 20 are considered significant if there are no metadata available. An SNHT value above 50 is almost always reached by breaks with size  $0.5\sigma$  (see Fig. 4-b). SNHT values above 20 are reached by about half of the series with breaks of size  $0.25\sigma$  but are still not attained by homogeneous time series.

The equal sampling SNHT is now applied to the following time series:

- log-pressure weighted 12GMT-00GMT temperature difference in the stratosphere: Depending on data availability analysis starts with the 20-30 hPa layer. Then all the means from two pressure levels down to the 200-300 hPa layer are calculated. Each layer mean time series is analysed with the equal sampling SNHT. For each point in time the maximum SNHT value from the analysed pressure layers is kept. The resulting time series of maximum SNHT values is then examined for significant values. Note that the 12GMT-00GMT time series is independent of the ERA-40 bg. A break detected in this time series has high credibility since it can be safely attributed to the obs series.
- log-pressure weighted layer-mean obs-bg temperature difference at 00GMT and 12GMT in the stratosphere. These time series, if available, are analysed in the same manner as the 12GMT-00GMT temperatures.
- 300-850 hPa log-pressure weighted layer mean obs-bg temperature difference at 00GMT and 12GMT. These time series are sensitive to station relocations and are more complete than the stratospheric time series. Temperature means from this relatively thick layer have quite small variance and therefore relatively subtle breaks having the same sign throughout this layer can be detected.

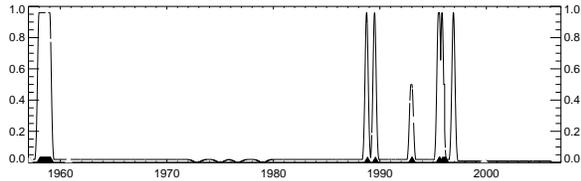
**4.2. Decision algorithm for break detection.** The probability for a break depends not only on the SNHT test statistics but also on the occurrence of events such as instrument changes. These information sources may be combined with the Bayesian rule (DeGroot, 1986). Let  $A_1$

be the event that a break with a given size occurs within a time interval of  $\pm 2$  years around a point in time and  $A_o$  be the event that no break occurs. Let B be the event that  $T_k^s$  reaches a value above a certain threshold  $x$ . We now want to know the probability of a break given that the event B occurs. Bayes' theorem states that this probability can be calculated as

$$(4) \quad Pr(A_1/B) = \frac{Pr(A_1)Pr(B/A_1)}{\sum_{j=0}^1 Pr(A_j)Pr(B/A_j)}$$

The challenge is now to specify the probabilities on the rhs of this equation. The probabilities  $Pr(B/A_1)$  and  $Pr(B/A_o)$  can be found from the histograms shown in the second column of Fig. 4. It is the area under the histograms between  $x$  and  $\infty$ . Metadata information can be included by specifying prior probabilities  $Pr(A_1)$ . If, for example the date of a radiosonde change is known, one can apply a higher prior probability  $Pr(A_1)$  to this particular date. In the examples presented in this paper, prior probabilities of 0.96, 0.5 and 0.02 have been chosen for radiosonde changes, radiation correction changes and no documented changes, respectively. This choice may appear extreme given the limited reliability of metadata but is necessary to make the detection system sensitive enough for metadata. With this choice, SNHT values of 31,43,50 (14,18,20 for 12GMT-00GMT difference series) are necessary to reach scores above 0.5 when put into the Bayes formula. A documented break with size  $0.3\sigma$  reaches similar scores as an undocumented break with size  $0.5\sigma$ . A score above 0.5 is necessary to trigger the breakpoint adjustment algorithm. This happens with breaks of size  $0.5\sigma$  without metadata.

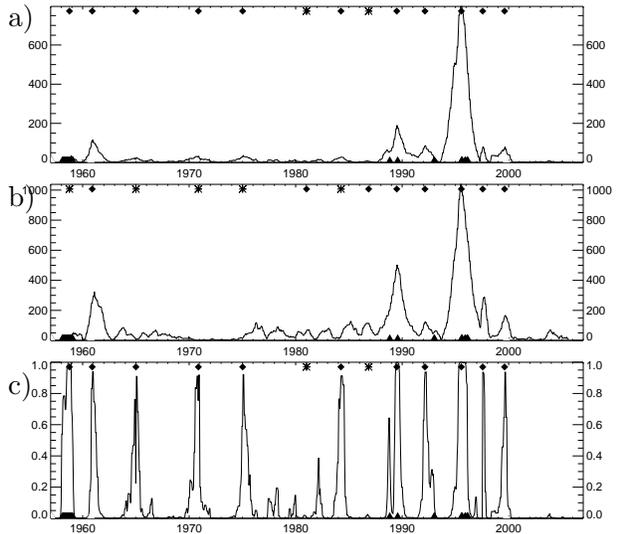
Since the metadata information may be imprecise, the prior probabilities have been modeled as Gaussians with standard deviation of 60 days. If only the year of a change has been reported the same high prior probability is specified throughout the year in order to assure that high weight is given to the metadata and to let the SNHT detect the break date. If GTS metadata have been available, the prior probability has been lowered to 0.01 between events since the sequence of RS-type reports explicitly states that there was no radiosonde type change.



**Figure 5:** Prior "probabilities" for station Bethel (70219, Alaska, 60.78N, 161.80W) derived from CARDS/BUFR metadata. Triangles/trapezoids denote metadata events from CARDS. Trapezoid in 1958 indicates that only the year of the event is known from CARDS. A constant high prior break probability is assigned throughout this year. Peaks without triangles/trapezoids are triggered by sonde type changes derived from ERA-40 feedback metadata. Gaussians are used to crudely represent metadata uncertainty. Prior probabilities in 1973,1975,1976,1979 are lowered to reduce the chance of false detections due to major changes in the satellite observing system.

Figure 5 shows the breakpoint analysis for the radiosonde station Bethel (70219, Alaska). Prior probabilities are assigned according to the CARDS/BUFR metadata. Panels a), b) of Fig. 6 show time series of the test statistic  $T_k$  of the SNHT at Bethel for the stratospheric obs(12GMT)-obs(00GMT) and for the tropospheric obs-bg time series at 00GMT. The test statistics exceed the significance levels (20 for panel a, 50 for panel b) several times at this station (see also Fig. 10 below for the corresponding difference time series). One can also see that the peaks are very sharp for the large breaks (1989, 1995), so that the break location is well determined in these cases, consistent with Fig. 4.

Since there are 5 time series of probability scores which may contain conflicting information (e.g. the location of the diagnosed breakpoint may differ slightly between the time series), one has to choose the location. The obs(12GMT)-obs(00GMT) score, shown in Fig. 6-c) is given the highest priority, since it is independent of possible inhomogeneities in the bg and therefore always attributable to the radiosondes. This is consistent with previous findings (Lanzante et al. 2003a; Sherwood et al. 2005) describing that most changes seem to be accompanied with changes in the obs(12GMT)-obs(00GMT). The peaks whose score is above



**Figure 6:** SNHT results and posterior probabilities for Bethel. Panels a,b show equal sampling SNHT test parameter for stratospheric obs(12GMT)-obs(00GMT) and for tropospheric obs-bg(00GMT) time series. Values above 20 (50) are considered significant. Squares denote breakpoints where SNHT test value is above significance threshold. Stars denote points where SNHT or break probability is below threshold in the respective series but is above the threshold in any other of the 5 analysed time series. All breaks are detectable in obs(12GMT)-obs(00GMT) series (see Fig. 10) except in 1981,1987 which are detected in 00GMT obs-bg difference series. c) "Probability" score for a break in the obs(12GMT)-obs(00GMT) time series, given the prior "probability" from a) and the test statistic b).

0.5 and which are the highest peaks within a  $\pm 2$  year interval are selected. Then the two probability time series from the stratospheric obs-bg time series are analyzed (the 00GMT series is shown in Fig. 6-b). The locations with the highest peaks are chosen as breakpoints unless a breakpoint has already been detected within a  $\pm 2$  year interval in the obs(12GMT)-obs(00GMT) series. Finally the probabilities from the tropospheric obs-bg time series are analyzed. Other priority choices for the breaks from obs-bg series are possible. One could give the tropospheric obs-bg time series priority to stratospheric obs-bg series or one could combine the probability scores again with the Bayesian

rule. Little overall sensitivity has been found in respect of this choice.

The diamonds and stars in Fig. 6-c above indicate the location of the finally chosen breakpoints. The diamonds indicate breaks detected already in the obs(12GMT)-obs(00GMT) time series, the stars are breaks detected in at least one of the obs-bg time series.

There are many possible ways to further improve this break detection algorithm. Perhaps the most important path for improvement is to use different time intervals (not only  $\pm 2$  years) for the SNHT since a running mean with fixed length may be incapable of detecting all breaks.

## 5. ESTIMATION OF THE BREAK PROFILES

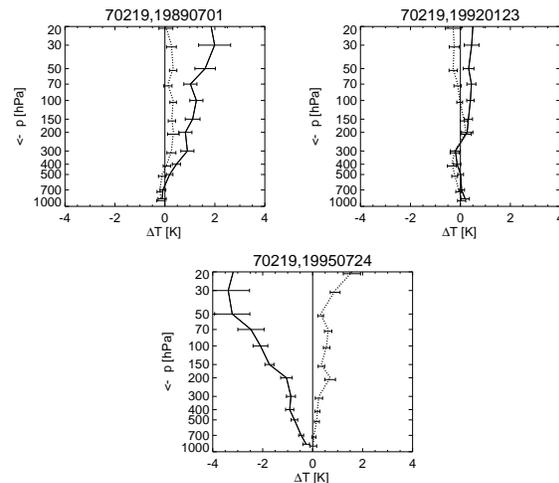
After the break detection, the mean obs-bg differences before and after the break to be adjusted are calculated at each pressure level, again ensuring that the annual cycle is sampled equally before and after the break. The difference of the means at every pressure level yields the estimated profile of the breaks (the solid and dotted curves in Figure 7). While this profile inevitably contains small-scale vertical noise it is not smoothed vertically, as in sensitivity experiments the smoothing did not improve the spatiotemporal consistency of adjusted trends or 12GMT-00GMT differences.

The time interval used for estimating the break size is varied between 1 year and 8 years. The default of 8 years is necessary for stable statistics in the presence of frequent data gaps. It is reduced only if there is another breakpoint before the breakpoint considered within 8 years. The interval after the breakpoint is always 8 years (if the time series is long enough), since the time series after the breakpoint has already been homogenized. The minimum interval of 1 year is also the minimum interval between breaks allowed by the break detection algorithm.

After estimating the break at each level, its significance is tested using Student's t-test. Only profiles where at least 2 pressure levels contain breaks which are significant at the 95% level are used for the adjustment.

Since the bg may be affected by changes in the satellite observing system, a second criterion to test the significance of a break profile has been developed. It compares obs-bg timeseries

of the tested radiosonde, with a composite of obs-bg timeseries from radiosondes surrounding the tested site. This composite was originally intended only for adjustment of the climatology of a tested time series (see section 7 below), but is quite useful for checking the break estimates gained from the obs-bg information at the tested site as well. Only if the break profile from difference series between obs-bg time series of the tested radiosonde and composite obs-bg time series of the surrounding radiosondes contains significant breaks at 2 or more pressure levels as well, the break profile is adjusted. The second criterion is more robust against jumps in the bg since the obs-bg series of the neighbouring radiosondes are likely affected by very similar bg jumps, as these have a very large-scale structure. The neighbour composite criterion used here has similarities to the methods used by Thorne et al. (2005a). They use, however, composites of obs-anomalies, not composites of bg-obs differences for break detection and adjustments. Using the above criteria, only about 70% of the detected breaks are actually adjusted.



**Figure 7:** Break profiles diagnosed for three breaks at station Bethel. Solid curves are break amplitudes at 00GMT, dashed curves are amplitudes at 12GMT. Dates in headers indicate diagnosed date of breaks. Break profiles are applied to all data preceding the breakpoint. Error bars are 90% percentiles of obs-bg difference before/after the break

If a break profile has been considered significant, adjustments have been applied at all levels, even if the adjustment amounts have been

below the significance threshold at some pressure levels. Adjustments at very high levels may be less accurate due to lack of data but the spatiotemporal consistency could be improved by the adjustments even at the 10 hPa level. No data deletions have been performed.

## 6. GLOBAL MEAN DIFFERENCES BETWEEN ERA-40 BG AND RADIOSONDE OBSERVATIONS

It is essential to be aware of any inhomogeneities of the ERA-40 bg since these reduce the applicability of the ERA-40 bg as a reference. Inhomogeneities in the bg time series may be introduced by changes in the ERA-40 observation coverage, in the observation biases correction and in the overall observation quality. Apart from radiosondes mainly the satellite data are affected by changing biases. The satellite data contributed to the realistic representation of many stratospheric features in ERA-40 analyses (Randel et al. 2004) and to a much better overall quality of the ERA-40 analyses and forecasts. However, the quality control and the adjustment of the changing biases of satellite radiances as described by Hernandez et al. (2004); Li et al. (2006); Harris and Kelly (2001) is challenging. Suboptimal radiance bias correction can easily compromise the homogeneity of the global mean ERA-40 bg forecasts (Uppala et al. 2006).

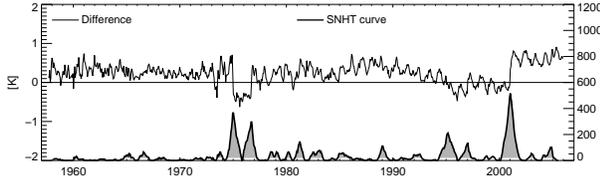
As can be seen from the global mean obs-bg difference series in Fig. 8, some changes in the ERA-40 satellite observing system did indeed lead to breaks in the global mean bg forecasts. The most prominent breaks evident in Figure 8 occurred in January 1975, September 1976 and April 1986 are related to problems with the NOAA-4 and NOAA-9 satellites. Jumps in 1995/1997 coincide with end of NOAA-11, start/end of NOAA-14 (see also Christy and Norris 2006). At high altitudes the effects of insufficient bias correction of radiances from the stratospheric sounding unit (SSU), particularly in the early 1980s, are noticeable (see Haimberger 2005; Uppala et al. 2006). Trenberth and Smith (2006) have recently diagnosed a spurious break in ERA-40 temperature analyses related to the assimilation of MSU-3 radiances at

the end of the NOAA-9 period. These problems are the likely reason for the rather weak stratospheric cooling and rather strong upper tropospheric heating trends of the ERA-40 bg compared to available radiosonde and satellite datasets (see Karl et al. 2006, their Fig. 3.4) in the ERA-40 analysis. This peculiar vertical pattern is also evident in the global mean trends of the unadjusted bg shown Fig. 9. In the tropics the upper tropospheric heating is even more pronounced.

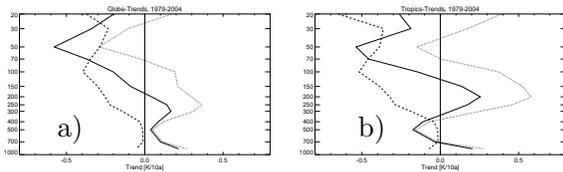
However, pervasive radiosonde temperature biases especially in the tropics (Sherwood et al. 2005; Randel and Wu 2006; Christy and Norris 2006) have most likely contributed to inhomogeneities in the global mean bg-obs series as well. From the 1990s onwards, there is also good correspondence between trends derived from MSU radiances (Santer et al. 2004) and ERA-40, which supports the bg. For this period the ERA-40 analysis looks more homogeneous also in the study of Trenberth and Smith (2006). Therefore with the present knowledge it is difficult to tell to what extent the radiosonde temperatures or the bg temperatures are responsible for the breaks and in particular for the trends in the global mean obs-bg departures of Fig. 8. Despite these uncertainties, the adjustment of the global mean bg has been tried such that the chance for spurious break detection due to breaks in the bg is reduced.

This is most straightforward for the large break in January 2001, which is certainly caused by the switch from ERA-40 to the operational ECMWF data assimilation system. It is a clear indication of the temperature uncertainties that still exist at stratospheric levels. The break can be adjusted relatively accurately at each radiosonde station from an overlap year (2001) between operational and ERA-40 innovations.

The remaining shifts in the global mean bg-obs series are much harder to attribute to either the bg or the obs. In the bg adjustment procedure applied here, it is assumed that much of the trend of the global mean obs-bg difference stems from time-varying biases in the global mean bg. Therefore the global mean obs-bg difference shown in Fig. 8 has been subtracted from the bg time series at individual stations.



**Figure 8:** Global mean 00GMT obs-bg difference averaged over 130 radiosonde stations of the GCOS Upper Air Network (GUAN, Daan 2002) at the 50 hPa level. Shifts in 1975 and 1976 caused by erroneous NOAA-4 bias correction. Shift in 1989 caused by change of assimilation streams in ERA, shift in 1995,1997 related to problems with NOAA-11/14. Shift in 2001 is caused by switch from ERA-40 bg to ECMWF operational bg.



**Figure 9:** Vertical profiles of global mean trends 1979-2004. Dotted line is trend of unmodified background forecast (bg) temperatures, solid line is modified bg trend and dashed line is trend of the bg modification. Panel a) shows global mean trends, panel b) shows tropical mean (20S-20N) trends. The shift in 2001 due to the switch from ERA-40 to ECMWF operational bg temperatures has already been removed from the unmodified bg temperature trends. Trends have been averaged from station time series within  $10^\circ \times 10^\circ$  gridboxes and then averaged over the globe to reduce the effect of the nonuniform radiosonde station distribution.

The bg adjustment has not been applied completely uniformly but has been varied with radiosonde observation density and with latitude (see appendix B for details). Other choices for the global mean bg adjustment are definitely possible. One could, for example, try to homogenize the global mean bg in a similar manner as the individual radiosonde records. Further investigation of the properties of the bg is needed to design an optimal bg adjustment.

At individual stations this bg adjustment is subtle compared to the typical sizes of breaks in radiosonde records. In the global mean, however, the bg adjustment procedure leads to stronger cooling/weaker heating trends

(Fig. 9), especially in the tropical upper troposphere/lower stratosphere. In the lower troposphere the adjustment has less impact since the bg-obs differences are smaller and show little trend below 300 hPa. It is interesting to note that both the unadjusted radiosondes and the bg are practically neutral ( $-0.01$  and  $0.02$  K/10a) in terms of 300-850 hPa trends in the tropics. The strong warming trends in the tropical lower troposphere as suggested by the RSS MSU T2 and LT products (Santer et al. 2005) are not supported either by the radiosondes or by the ERA-40 bg.

The necessity of this bg adjustment prior to the homogenization and the sensitivity of the homogenization results on the bg adjustment (see section 9) limit the accuracy of the homogenization results. It is expected that the size of adjustments necessary for the bg from future re-analyses will be much less due to more advanced treatment of satellite biases.

## 7. ADJUSTMENT OF THE CLIMATOLOGY OF RADIOSONDE TIME SERIES

Even after successful homogenization, the climatology of a radiosonde record may have a large bias if the most recent part of the time series is biased. The reasons are typically as follows:

- Some countries still operate radiosondes with nonnegligible temperature biases. Examples are the Chinese and most Russian radiosondes.
- Some radiosonde time series end well before the year 2000. Prominent examples are the weather ships in the Atlantic and Pacific oceans. The most recent parts of these time series are likely biased.
- Some time series, e.g. over Russia in the 1990s, have gaps that are wider than the averaging intervals used for calculating the break profiles. In this case the time series before the gap cannot be adjusted by the homogenization procedure.

Adjusting the climatology and the choice of "trusted" radiosonde types whose measurements are left untouched, as done here, is a delicate task since it implies preference to certain radiosonde types. However, this task cannot be avoided for climate data assimilation purposes.

The involved biases are large enough to cause noticeable biases of the resulting analyses or even to trigger rejection of a substantial part of the radiosonde measurements, particularly at high altitudes in the satellite era.

The adjustment of the climatology is performed after the time series homogenization; it is the final adjustment step. Therefore the time series should already be without any detectable breaks. The challenge is now to calculate a reference that is less biased than the most recent part of the tested time series.

Many ascents launched in recent years have relatively small biases up to 10 hPa level or the biases are well documented and can be corrected. A subset of 342 radiosonde stations has been identified where the climatology is considered unbiased after homogenization. A map of these can be found in Haimberger (2006). These are stations which have been equipped with Vaisala RS80/90/92 radiosondes or temperature sensors, with MeiseiII radiosondes and Sippican radiosondes. Their performance is well documented in radiosonde intercomparison studies such as Nash et al. (2005).

At all other stations (e.g. with VIZ radiosondes, Chinese and Russian radiosondes) the most recent part (usually the most recent eight years, if available) of the records has been regarded as biased and therefore the records have been adjusted.

A reference time series has been calculated from a neighbour composite of obs-bg differences from the above 342 radiosonde stations. The size of the adjustment is calculated by comparing the obs-bg difference of the tested series with the composite obs-bg difference of the neighbouring time series. The mean obs-bg difference of the tested time series should be very close to the mean obs-bg difference of the composite, since the spatial pattern of a 8-year averaged bg temperature field is smooth and the bg temperature gradients are considered realistic. Since the obs-bg differences are such small increments, their composites are less affected by data gaps than composites of absolute temperatures.

The composite obs-bg difference is a weighted mean of neighbouring homogenized

radiosondes using weights decreasing exponentially with distance from the radiosonde station to be adjusted. A decorrelation distance of 3000 km has been used. At least 20 stations have been required to be available for the composite.

Experience with this approach has been encouraging. Many biases of the most recent parts of time series, as evident e.g. over Russia or at the weather ships in the Atlantic could be significantly reduced (see Figs. 17 and 18 below).

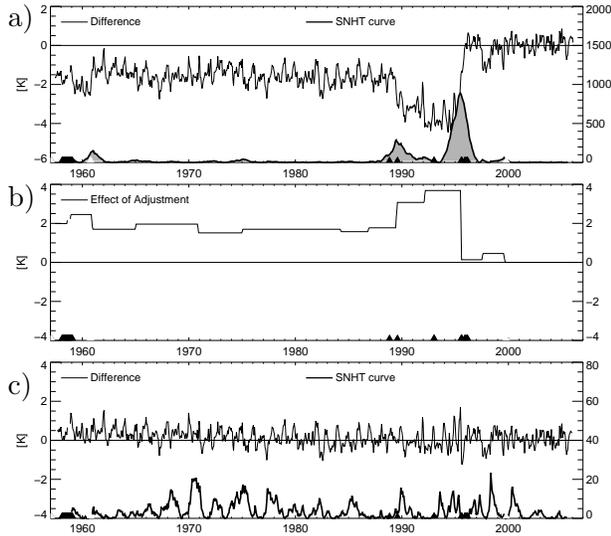
## 8. ADJUSTMENT RESULTS FOR SELECTED INDIVIDUAL STATIONS AND REGIONS

A small subsample of radiosonde records is investigated to demonstrate the strengths and the limitations of RAOBCORE. For most countries similar examples can be found since practically all long radiosonde time series contain inhomogeneities. Even if the radiosonde records presented contain large breaks, they have high information content.

**8.1. Adjustments in the satellite era.** Since homogenization works backward in time we begin with the period 2005 back to November 1978. While the ERA-40 bg contains more information independent of radiosondes than in the pre-satellite era, there is also the risk of spurious breaks due to insufficient bias adjustment between satellites (see section 6). The following examples should demonstrate that adjustment with obs-bg difference series is possible despite these problems.

One is Bethel, Alaska, which had a relatively homogeneous time series between 1961 and 1989 but then three marked breaks in 1989, 1992 and 1995. Test statistics for this station have been shown already in Figs. 5-6 above. Due to space constraints the analysis in this and other examples is restricted to the 50 hPa level in this study.

Fig. 10 shows time series of unadjusted and adjusted 12GMT-00GMT temperature differences at Bethel. One can see how different the day night differences are depending on the radiosonde system in use (VIZ before 1989, then Space Data radiosondes, then Vaisala RS80 radiosondes, according to CARDS). These jumps can only come from changes of the observing system and are independent of the ERA-40 bg.



**Figure 10:** Thin curve is unadjusted 12GMT-00GMT radiosonde temperature difference at 50 hPa, for station Bethel (70219, Alaska). Thick curve is SNHT test statistic (right axis). Peaks in SNHT test statistic indicate abrupt changes in the mean difference. Triangles at the bottom indicate changes of radiosonde type and on-site radiation correction, as documented by Aguilar (2000). b) Effect of adjustments applied by RAOBCORE on 12GMT-00GMT difference, c) 12GMT-00GMT difference time series after adjustment. Note different scale of right axis compared to panel a)

The ERA-40 bg comes into play for the adjustments. The adjustments for the 00GMT and 12GMT ascents are calculated independently from obs-bg time series at 00GMT and 12GMT. It is worth emphasizing that the 12GMT-00GMT obs time series is never used for adjustment purposes in RAOBCORE; it is used for break detection purposes only.

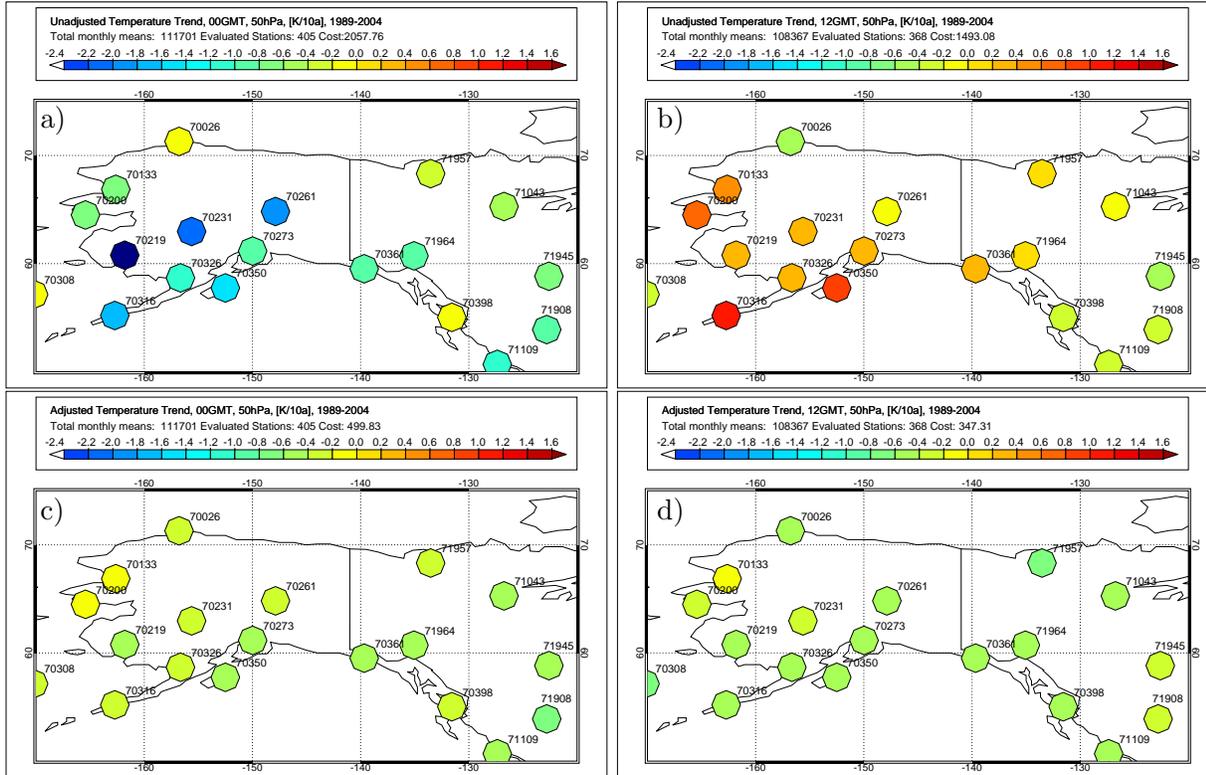
As one can see from panel b) the diagnosed breakpoints coincide well with metadata events in the 1980s/1990s. The breaks in 1976 and in the 1960s are smaller but one can see from visual examination of the unadjusted 12GMT-00GMT time series, that there are breaks indeed. The adjusted 12GMT-00GMT time series is much more homogeneous and has an almost neutral trend, as is expected for this difference series. The homogeneity of the 12GMT-00GMT time

series is a good consistency check for RAOBCORE since this time series has not been used for calculating the adjustments.

Bethel is not the only radiosonde station in Alaska affected by instrument changes. As can be seen from Fig. 11, the unadjusted radiosonde records over Alaska and Canada yield rather different temperature trends for the period 1989-2004 at 00GMT and 12GMT and the trends are also spatially rather heterogeneous due to the different station histories. After adjustment with RAOBCORE, the temperature trends at 00GMT and 12GMT are in good agreement and also the spatial homogeneity is much better. Note that there were substantial adjustments not only for the daytime (00GMT) ascents. This is consistent with findings from radiosonde intercomparisons (Nash and Schmidlin 1987) that showed substantial nighttime deviations between VIZ and Vaisala RS80 radiosondes and contradicts the impression one may get from Sherwood et al. (2005) who stress the importance of daytime biases.

The second example is Darwin, Australia, which switched from Philips MKIII radiosondes to Vaisala RS80 radiosondes in May 1987. This station has been analysed more thoroughly for example by Free et al. (2002). Darwin is difficult since only 00GMT ascents have been available before 1987 and even those had some gaps at the 50 hPa level. Fig. 12 shows the adjustments suggested by RAOBCORE for this station. The break size is estimated 2K which is in accord with Free et al. (2002). There is a clear indication of a break in 1963 as well, which is in good agreement with breaks suggested by Lanzante et al. (2003a) for this station. Additional breaks in Fig. 12 seem evident in 1976 and 1981 but these have not been adjusted since the significance criteria have not been met.

The robustness of the RAOBCORE adjustment procedure may be seen from Fig. 13. The spurious 12GMT-00GMT differences are reduced at almost all stations that launch two radiosondes per day during the period 1988-1990. The largest adjustments can be found over the US, China and Australia. The spatial consistency of the differences (as measured by the "cost"  $C$  defined in appendix A), has improved as well compared to Fig. 1 and the amplitude of



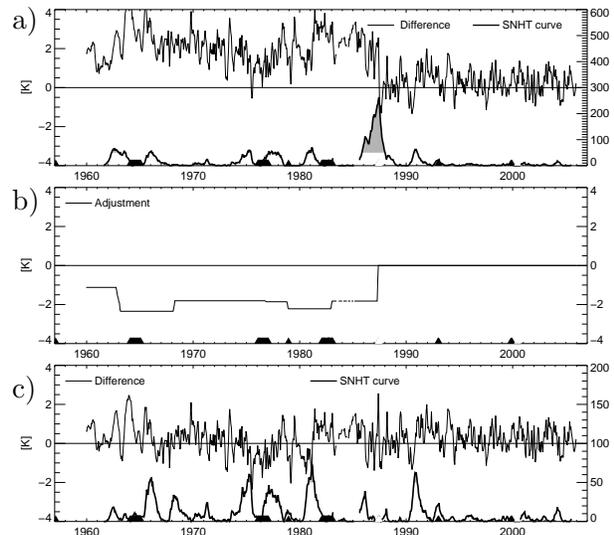
**Figure 11:** Radiosonde temperature trends during 1989-2004 in K/decade over Alaska/Northern Canada at 50 hPa. a) unadjusted obs(00GMT) time series, b) from unadjusted obs(12GMT), c) from adjusted obs(00GMT) and d) from adjusted obs(12GMT). Numbers are WMO station IDs.

the differences has been reduced to a few tenths of a K except in the tropics, in agreement with recent results from Free and Seidel (2005).

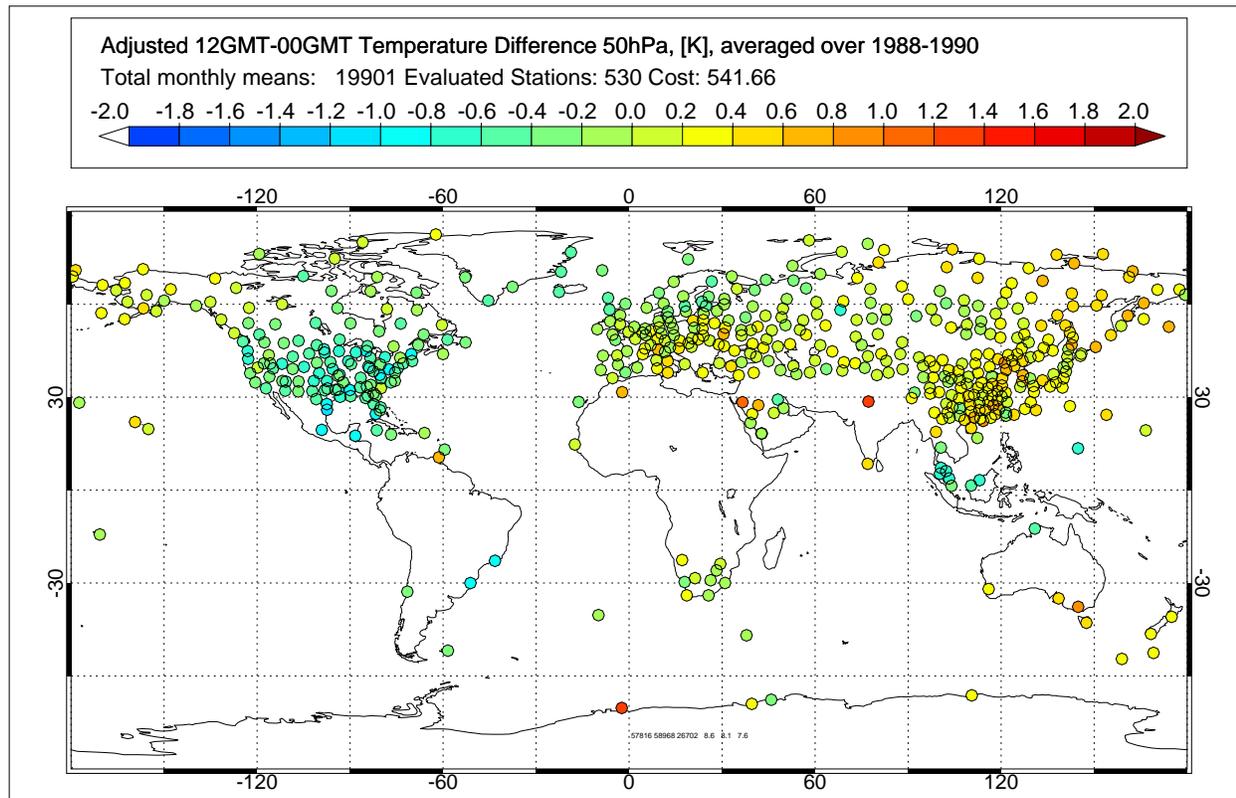
## 8.2. Adjustments in the pre-satellite era.

In the pre-satellite era, the ERA-40 bg does not contain much more upper air information than is available from radiosondes. Therefore the issue of dependence of the bg on the radiosondes to be tested becomes particularly important. Further the radiosonde density in the tropics and the southern hemisphere was very sparse, making neighbour intercomparisons rather difficult. Some examples are given indicating that adjustment of radiosondes using obs-bg time series is still possible.

The dependence of the bg on the radiosondes to be tested is expected to be particularly large for (i) remote stations and (ii) large countries using the same radiosonde equipment. The degree of dependence may be assessed qualitatively by looking for example at MESURAL radiosondes



**Figure 12:** Obs-bg difference at 50 hPa, 00GMT for the radiosonde station Darwin (12.43S, 130.87E). a) unadjusted time series, b) adjustment applied, c) adjusted time series.



**Figure 13:** Adjusted 12GMT-00GMT temperature differences in 50 hPa for the period 1988-1990. Note improved spatial homogeneity compared to Fig. 1-a).

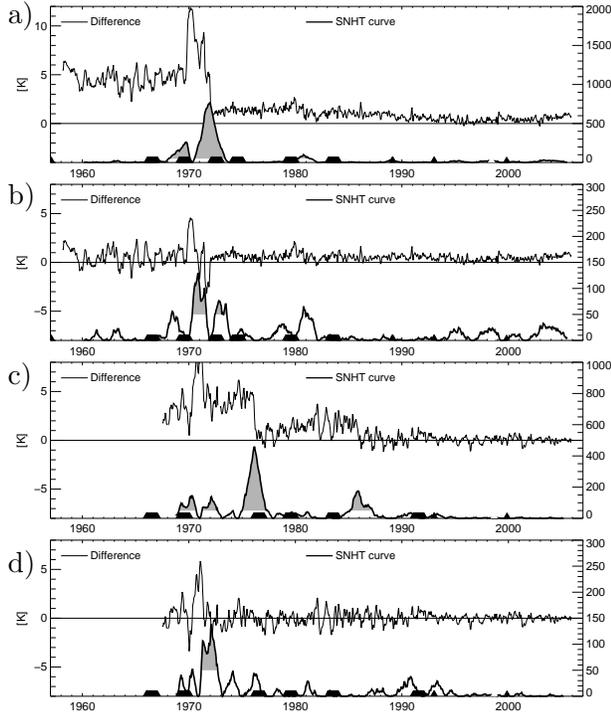
in the late 1960s/early 1970s. Particularly the MESURAL FM43 radiosondes had large radiation errors. This radiosonde type was in use in France and in former French dependencies (e.g. many Pacific islands). This gives an opportunity to see whether the same breaks are diagnosed over data rich and extremely data sparse areas. Fig. 14 shows obs-bg differences for local daytime ascents at the French radiosonde station Trappes (12GMT) and at the southeast Pacific station Rapa before and after adjustment with RAOBCORE.

Both time series show very large biases around 1970 when MESURAL-FMO-43B radiosondes were in use. These were replaced by MESURAL-FMO-44C radiosondes in 1972 at Trappes and in 1976 at Rapa. The suggested corrections from obs-bg time series are about 20% weaker at Rapa, probably because the bg is influenced by Rapa itself. A correlogram from the stratospheric obs-bg time series at Rapa (not shown) also indicated some dependence of the

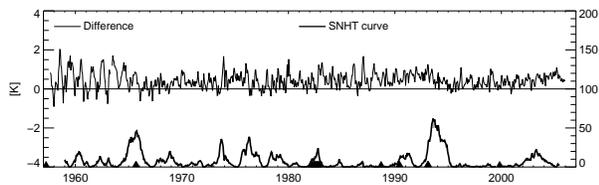
bg on the observations at Rapa: the serial correlation of the obs-bg differences decreases from values of 0.3 for 1 day lags to insignificant levels for lags of 4 days. Nevertheless it is evident that the bg contains enough information to yield sensible break estimates even in this area. The obs-bg time series at Rapa after adjustment of the observations (Fig. 14-d) looks much more homogeneous.

At Trappes the standard deviation of the obs-bg difference is much smaller since the bg forecast is more accurate in this region, and also the serial correlation of the innovation time series is practically zero (except near a breakpoint). The adjusted obs-bg time series (Fig. 14-b) is again more homogeneous although not all instationarities are removed from the strongly varying time series between 1969 and 1972.

Not only Trappes but all French radiosonde sites are affected by the MESURAL problem with daytime stratospheric temperature biases in

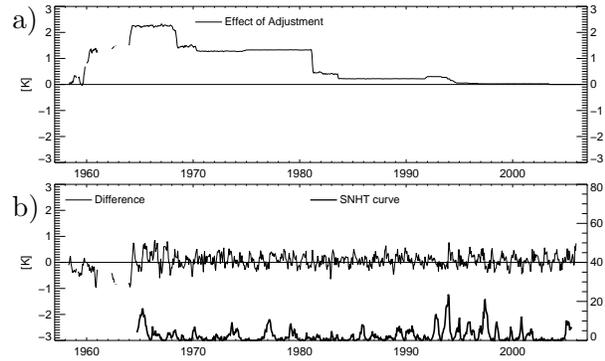


**Figure 14:** Obs-bg difference at the 50 hPa level at 12GMT for radiosonde stations Trappes (7145, France, 48.77N 2.02E, change from MESURAL-FMO-1940B to MESURAL-FMO-1943B in 1969 and to MESURAL-FMO-1944C in 1972) and Rapa (91958, 27.61S, 144.33W, change from MESURAL-FMO-1940B to MESURAL-FMO-1943B in 1969 and to MESURAL-FMO-1944C in 1976). Panels a,b: Trappes before/after adjustment with RAOBCORE. Panels c,d: Rapa before/after adjustment with RAOBCORE.

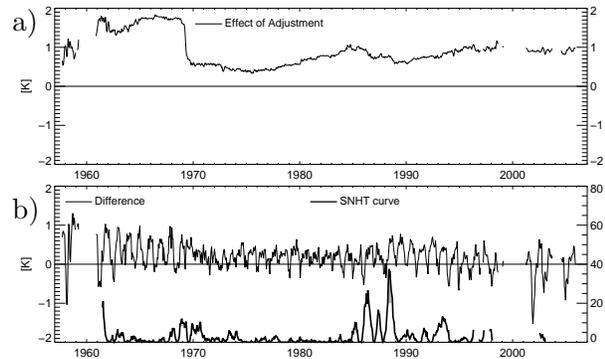


**Figure 15:** Time series of obs-bg difference at 50 hPa at 00h for station Stuttgart (10739, 48.83N, 9.2E), downstream of the French radiosondes.

the order of 10K. The bg over Europe seems almost unaffected by this problem although most of the biased MESURAL observations have been accepted by the ERA-40 quality control system. This can be concluded from the obs-bg time series at station Stuttgart (Fig. 15), which is situated just 'downstream' of France and does not



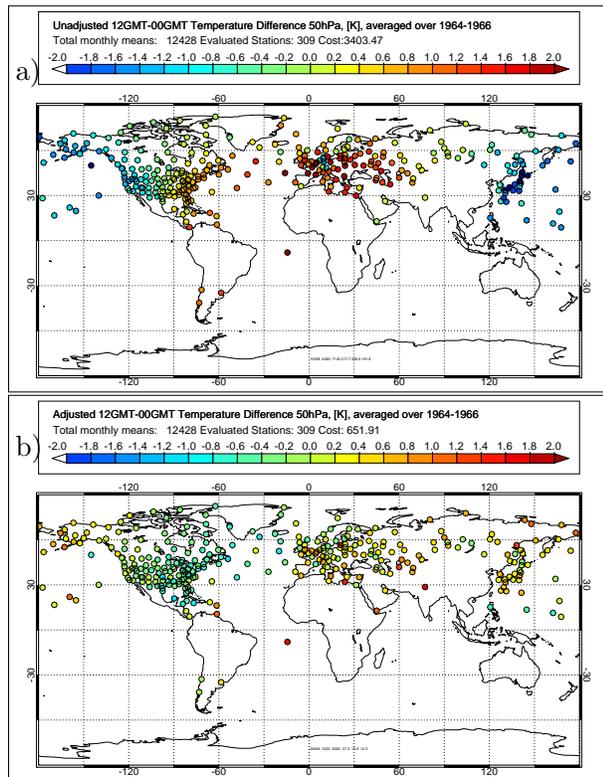
**Figure 16:** a) effect of RAOBCORE adjustments on 12GMT-00GMT series of composite of Northern Japanese radiosondes (IDs 47400-47700). b) 12GMT-00GMT series after adjustment. Documented breaks in 1967, 1981 and 1993/94.



**Figure 17:** a) Effect of RAOBCORE adjustments on obs(12GMT)-obs(00GMT) time series for far eastern Russian radiosondes (IDs 31000-33000). b) 12GMT-00GMT difference of adjusted time series. Note data gaps in recent periods. The strong annual cycle of the difference remains to be adjusted.

have documented changes around 1970. The obs-bg temperature time series of this station was homogeneous in 1969 and 1972 when France changed the radiosonde instrumentation. If the bg were influenced by the breaks of the French temperature time series, a shift would be visible in the obs-bg time series at Stuttgart.

While over Europe this result may be plausible since there is much independent information available to make the bg apparently immune against even large breaks, the situation may be different when a large country changes its radiosonde equipment simultaneously. Figures



**Figure 18:** 12GMT-00GMT temperature differences in 50 hPa for the period 1964-1966 a) before and b) after adjustment with RAOBCORE. Note weather ship stations over the Atlantic and Pacific whose time series end in the 1970s; their climatologies have been adjusted as described in section 7.

16 and 17 provide two examples that even in those situations, the bg is independent enough to be used for adjustment. In the Japanese radiosonde time series (Fig. 16), major breaks are adjusted in 1968, 1981 and 1993/94. The Japanese composite of adjusted 12GMT-00GMT time series does no longer show spurious breaks (Fig. 16 b).

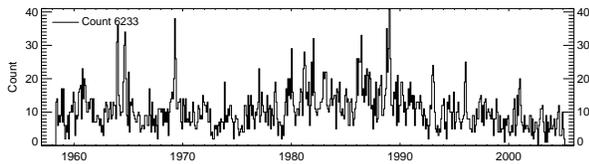
A similar example is far eastern Russia (stationIDs 31000-33000; Fig. 17). The most prominent shift in 1969 is caused by a change from MARS to RKZ5 radiosondes and is adjusted well by RAOBCORE. This composite also shows the difficulties in data availability over Russia in the 1990s. The most recent parts of most stations before the data gaps had to be adjusted by about 1 K since their obs-bg time series showed a distinct bias compared to a composite of mainly Japanese and Alaskan stations.

The resulting adjusted 12GMT-00GMT time series is nevertheless relatively stationary (apart from a remaining annual cycle), as it should be. The daily mean adjustments in 1969 are about  $-0.7$  K at 50 hPa over Eastern Russia (31000-33000). This compares well with daily mean adjustments of  $-0.6$  K over Western Russia (west of 60E, south of 60N), which have been taken at similar solar angles. This is remarkable since the bg over Western Russia is potentially influenced by European radiosondes whereas over Eastern Russia it is more likely influenced by Alaskan and Japanese radiosondes (the latter have large breaks in 1968). Only the positive 12GMT-00GMT differences after adjustment in the 1960s is an indication of a slight overcorrection, due to a too warm nighttime bg or a too cool daytime bg caused by biases in the radiosondes twelve hours earlier.

Fig. 18 shows that the RAOBCORE adjustments substantially improve the spatial consistency of 12GMT-00GMT time series also in the early period 1964-1966, although a few stations remain where the adjustments by RAOBCORE seem insufficient. Note also that many weather ships were operational in the 1960s until the early 1970s. For these ships, which operated up to the early 1970s, the adjustment of the most recent period remove the large biases visible in Fig. 18-a.

Fig. 19 shows the time series of the adjustments per month applied to the global radiosonde network. The combined length of all time series divided by the total number of adjustments applied (6233) yields about one adjustment every seven years. While this number may seem large, it is in good agreement with Thorne et al. (2005a) and is also supported by the number of radiosonde type changes derived from GTS (about 700 from 1990 onwards) and documented in CARDS (about 6700). About 30 percent of the adjustments coincide with metadata events within  $\pm 14$  days of the detected breakpoint. Most peaks are related to changes in large countries (e.g. Russia, France in 1969). There are no peaks evident in 1973, 1975, 1976, 1978/79 where major changes in the ERA-40 observing system occurred.

The adjustments for all 1184 stations are available from



**Figure 19:** Time series of number of adjustments per month. Peaks are related to metadata events in large countries, e.g. radiation correction changes in 1982 and 1993. Break count includes climatology adjustments as explained in section 7.

<http://www.univie.ac.at/theoret-met/research/RAOBCORE/> as plots and as ASCII-formatted file. The dataset will be updated annually, as soon as the input data from IGRA and ECMWF become available.

## 9. SENSITIVITY EXPERIMENTS

The homogenization process has several sources of uncertainties, as outlined e.g. by Thorne et al. (2005a); McCarthy et al. (2006). Since RAOBCORE is an automated system, it has been possible to test the sensitivity with respect to various parameters of the adjustment system.

The sensitivity of the following parameters has been tested: (i) 50-100 hPa and 300-850 hPa layer global mean trends which may be compared e.g. with the results collected in Fig. 3.4 of Karl et al. (2006). (ii) Tropical (20N-20S) mean trends for the same period, (iii) spatial consistency of 50 hPa 1979-2004 trends, as defined in appendix B, (iv) spatial consistency of 12GMT-00GMT differences averaged over 1964-1966, (v) breakpoint count. The main results are summarized in Table 1.

Time series have been used for the trend comparison only if less than 24 months of data out of the 26 year period 1979-2004 have been missing. If only 00GMT or 12GMT trends have been available for a station, these have been regarded as daily mean trends. If both have been available, their average has been taken. The daily mean trends have been averaged to 10x10 degree lat/lon gridboxes and the mean trends from these gridboxes have been used to calculate the tropical (20S-20N) and global mean trends. The intermediate calculation of 10x10 degree

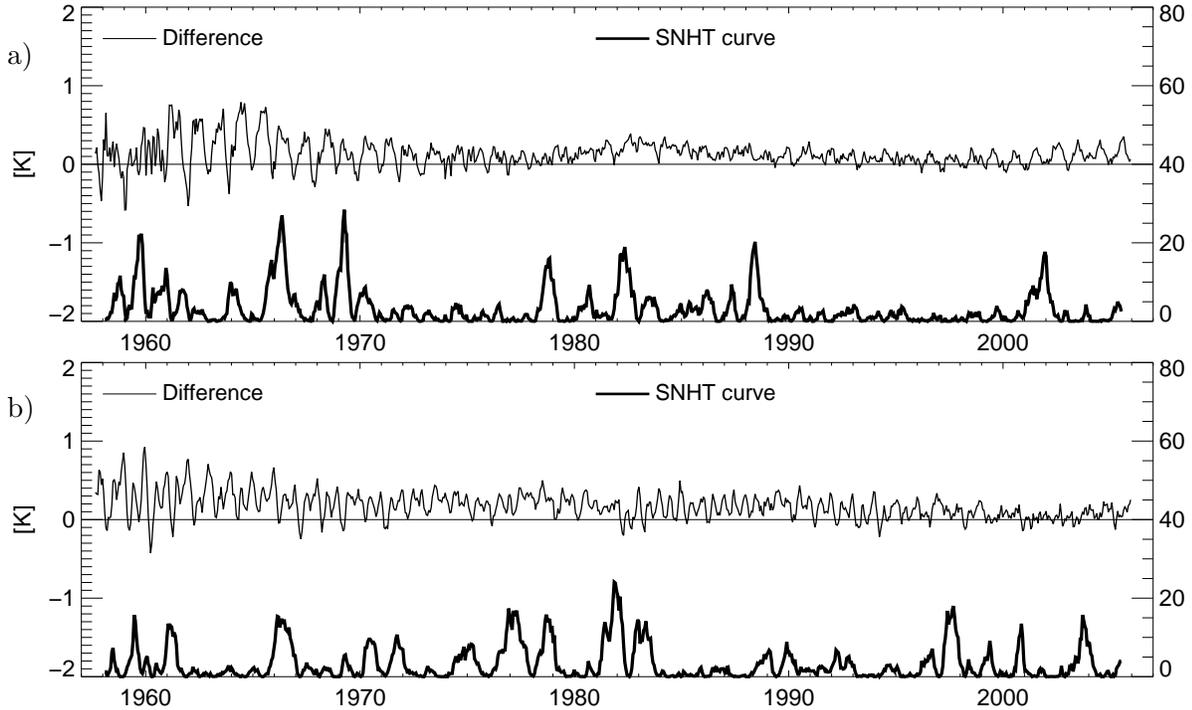
trends helped alleviate the effects of the uneven spatial distribution of radiosondes.

The first two rows BG and BGADJ describe the unadjusted and adjusted versions of the ERA-40+ECMWF background as they are used in the sensitivity experiment below. The original ERA-40+ECMWF bg (BG) has almost neutral trends in the lower stratosphere, which are not supported by the UAH and RSS satellite products and even less by the radiosondes. As one can see from row BGADJ, the adjustment described in section 6 introduces moderate stratospheric cooling (similar to that of MSU satellite products) into the bg.

Row UNADJ refers to the unadjusted radiosonde dataset used as input for RAOBCORE. Note the strong stratospheric cooling and large spatial heterogeneity compared to the bg.

Row RAOBCORE summarizes the results of RAOBCORE as used for the plots in this study, i.e. with bg adjustment, with use of metadata and with the thresholds described in section 4. The radiosonde trends adjusted with RAOBCORE show much less cooling than UNADJ and are very close to the existing homogenized radiosonde datasets (HadAT2, RATPAC, see Karl et al. 2006). Although the RAOBCORE trends are not directly compared to satellite data in this paper, it is clear that they remain more negative than those derived from satellite products. The spatial consistency of both trends and 12GMT-00GMT difference after the adjustment is much better than for the original radiosonde dataset but is still not as high as the consistency of the ERA-40 bg trends and 12GMT-00GMT differences.

The other rows in table 1 show results of sensitivity experiments. For NOMETA, RAOBCORE has been applied with constant prior probability 0.02, i.e. metadata information has been completely discarded. For ONLY-META, the prior probability has been set to zero if no metadata were available and to 0.999 when metadata events (radiosonde type changes and radiation correction changes) were documented. NOBGC refers to a RAOBCORE run where the bg was not adjusted before application of RAOBCORE. In the experiment



**Figure 20:** Time series of composite mean 12GMT-00GMT differences from radiosondes, a) between 30W and 40E, b) between 120E and 120W after adjustment with RAOBCORE. Note better homogeneity and almost neutral trends compared to Fig. 2.

NOBGC\_ONLYMETA the bg has not been adjusted and metadata have been treated as in ONLYMETA. STRICT refers to an experiment where the thresholds for break detection and adjustments have been set higher (40% higher  $T_k^s$  values required, 60% larger break size required than for the default).

The results suggest that the adjusted trends are relatively insensitive to changes in metadata treatment and to the increase of the breakpoint thresholds. However, there is sizeable sensitivity to the background adjustment applied before application of RAOBCORE. If the unadjusted bg is used as reference, the resulting RAOBCORE-adjusted trends shift towards more warming by more than 0.3 K/10a in the tropical stratosphere and by about 0.04 K in the tropical troposphere. In this case the adjusted radiosonde trends fit better to satellite-derived temperature trends but are still within the range of uncertainty of upper air trends (see again Karl et al. 2006). The bg adjustment has relatively little impact on breakpoint count and on the spatial consistency of adjusted radiosonde

time series. This shows that high spatial consistency of trends and 12GMT-00GMT differences, while a desirable property, is not sufficient to determine the quality of the diagnosed global mean upper air trends.

While the unadjusted bg trends are most likely too weak in the stratosphere (see Fig. 3.4 in Karl et al. 2006) so that the experiment NOADJ is regarded as rather extreme, it shows the importance of a homogeneous reference. In view of mounting evidence for pervasive biases of the unadjusted radiosonde dataset particularly in the tropics (Sherwood et al. 2005; Randel and Wu 2006), the bg adjustment applied in the best estimate of this paper (row RAOBCORE) may seem aggressive. There is indeed the possibility that the bg adjustment removes climate signals from the MSU and other satellite instruments. This is suggested by the good correspondence between ERA-40 MSU4 equivalent and RSS (Mears et al. 2003) datasets, at least in the stratosphere, from ca. 1989 onwards (Santer et al. 2004). Therefore the

difference in stratospheric trends between experiments RAOBCORE and NOADJ is probably larger than the true uncertainty of the radiosonde trends.

The sensitivity of the trends with respect to the bg adjustment is sizeable, even if adjustments are made only at documented breakpoints. Since the breakpoints are practically the same in the ONLYMETA and NOBGC\_ONLYMETA experiments, the trend differences must mainly come from break estimation, not from break detection. As is indicated in Fig. 7 the uncertainty in break estimation is at least 0.5 K in the stratosphere, even in recent years. Since almost every radiosonde time series contains at least one break between 1979 and 2004, trend differences in the order of 0.3 K, depending on the reference used, are not surprising. It has further been tried to use shorter intervals (4 years instead of 8 years) for break estimation but this did not reduce the sensitivity to the bg adjustment either.

The high values of  $C$  and low break count values for the NOMETA and STRICT experiment suggest that many breaks remain undetected in these experiments. One has to accept that there are thousands of breaks which need to be adjusted in order to get a spatially consistent dataset. The impact of these parameters on trends is small which indicates that the largest breaks are found even without metadata and large thresholds.

The sensitivity experiments stress the importance of the reference series used for homogenization. The uncertainties in the bg adjustment lead to uncertainties that are similar to the differences between existing homogenized radiosonde and satellite datasets.

## 10. CONCLUSIONS AND OUTLOOK

This paper has documented a method called RAOBCORE which uses innovations (observations minus background forecasts) from a frozen data assimilation system such as ERA-40 (Upala et al. 2005) for automatic homogeneity adjustments of radiosonde temperature data. Innovations back to 1958 have been used. The method has been designed to make the adjusted radiosonde dataset as suitable for reanalyses as possible. This involves not only realistic trends

but also completeness of the radiosonde dataset and adjustment of the absolute temperatures, not anomalies. It could be shown with several examples that RAOBCORE substantially reduces spurious trends in the 12GMT-00GMT differences and improves the internal spatial consistency of the radiosonde measurements.

The global trend figures from RAOBCORE show good agreement with existing global homogenized radiosonde datasets such as HadAT (Thorne et al. 2005a), LKS (Lanzante et al. 2003b). More detailed comparisons with regional radiosonde datasets, e.g. CALRAS (Haerberli 2006) as well as MSU records Mears et al. (2003); Christy et al. (2003) are in preparation.

Inhomogeneities introduced into the background forecast temperatures due to changes in the (satellite) observing system and the dependence of the background forecast error on the station to be tested and adjusted, are regarded as the largest potential sources of uncertainty when using innovations. While the results from individual stations suggest that the latter problem is relatively small, the sensitivity of the RAOBCORE-adjusted trends with respect to the global mean bg is not negligible. The uncertainty (which is estimated 0.3 K/10a for the global mean 50-100 hPa layer and 0.05 K for the 300-850 hPa layer) of trends from the RAOBCORE-adjusted dataset can be reduced below that of existing upper air datasets only if the inhomogeneities in the global mean obs-bg time series (as shown in Fig. 8 for the 50 hPa layer) can be attributed more clearly to either the bg or the obs.

The behaviour of the ERA-40 data assimilation system in the satellite period is still investigated and increasingly well understood (Upala et al. 2006). With this building knowledge the uncertainty of the ERA-40 bg and thus of the resulting RAOBCORE trends can likely be reduced. It is quite possible that the bg adjustment used in this study was too aggressive and that therefore the RAOBCORE best estimate for global trends after homogenization still shows too much cooling/little warming. A more thorough investigation of this matter is under way but is beyond the scope of this article.

Some radiosonde temperature time series have a substantial bias even in their most recent

**Table 1:** Results from sensitivity experiments with RAOBCORE. Trends in K/decade for Globe and Tropics (in brackets) valid for the period 1979-2004. Cost is value of cost function defined in eq. (5) at the 50 hPa level for 12GMT-00GMT temperature differences averaged over 1964-1966 and for 1979-2004 temperature trends. Break count is total number of breaks detected, including about 760 adjustments of climatologies (see section 7).

Acronym	Description	50-100 hPa Trend	300-850 hPa Trend	Cost 50 hPa Trend	Cost 50 hPa 12- 00	Break Count
BG	Unadjusted bg(except shift at 2001) ERA-40 + ECMWF bg temperatures	-0.05(0.11)	0.15(0.02)	139	360	
BGADJ	Adjusted bg	-0.39(-0.34)	0.11(-0.03)	134	356	
UADJ	Unadjusted radiosondes	-0.83(-0.94)	0.09(-0.01)	564	3401	
RAOBCORE	RAOBCORE best estimate (Control run)	-0.66(-0.65)	0.11(0.00)	192	652	6233
NOMETA	No Metadata, constant prior probability 0.02	-0.65(-0.65)	0.11(0.00)	193	757	5963
ONLYMETA	Adjustments only at documented changes	-0.70(-0.73)	0.12(0.03)	494	1652	2516
NOBGC	No adjustment of bg before RAOBCORE	-0.44(-0.31)	0.13(0.05)	197	638	6097
NOBGC_ ONLYMETA	No adjustment of bg before RAOBCORE, only documented breaks	-0.51(-0.46)	0.15(0.06)	503	1635	2557
STRICT	$\Delta T$ of 0.5K required at 2 places in break profile	-0.66(-0.64)	0.11(0.02)	303	895	4093

parts. For these stations the biases have been reduced by adjusting with a composite of homogenized neighbouring radiosondes in an extra step after the homogenization procedure. This step is not compulsory for climate trend analysis but adds substantial value for climate data assimilation applications which are affected by absolute biases (Dee and Da Silva 1998; Uppala et al. 2006).

Apart from the correction algorithm itself this article has documented gaps in both the ERA-40 and the IGRA (Durre et al. 2006) radiosonde datasets. The union of both datasets is about 5-10% larger than the individual datasets.

Efforts to create a comprehensive global radiosonde dataset must therefore continue.

In view of the results gained so far, it seems worthwhile to perform a pilot climate data assimilation of the period 1939-1957 in order to be able to adjust also these time series using the innovations gained from such an assimilation. Numerous radiosondes are available for this period (Durre et al. 2006; Bronnimann 2003). Innovations from a 4D-VAR assimilation system would add valuable information to these data which are otherwise be very difficult to homogenize. Preliminary results with RAOBCORE also indicate that homogenization based on analysis of innovations is feasible also

for radiosonde winds. Put into a more general context, the method of reanalysis (or climate data assimilation) together with proper archival of the innovations may become a key method for the improvement of historical observations. The recent success of assimilations using surface pressure only (Compo et al. 2006) together with potential application for homogenization makes the idea of a hundred year reanalysis attractive, despite the sparseness of the available data.

The efforts described here are part of a long term activity to provide a mature and complete homogenized radiosonde dataset, suitable as input for the next major European reanalysis planned for the end of this decade.

#### ACKNOWLEDGEMENTS

The foundation for this work has been laid during the European Commission contract MEIF-CT-2003-503976, which enabled the author to work at ECMWF in 2004. I thank Sakari Uppala and Adrian Simmons in particular for their continuous support. After 2004 the work has been funded by project P18120-N10 of the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (FWF). Christina Tavolato and Stefan Sperka were helpful in detecting bugs in the adjustment procedure. The comments of the anonymous reviewers led to substantial improvements of the original manuscript.

#### APPENDIX A. A SPATIAL CONSISTENCY MEASURE FOR TRENDS AND DAY-NIGHT DIFFERENCES

To estimate the spatial consistency of trends and 12GMT-00GMT differences, a simple cost function is defined as

$$(5) \quad C(p) = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N \left( \Delta x_{ij}(p) e^{(-d_{ij}/1000)} \right)^2$$

where  $p$  is the pressure level,  $i, j$  are station indices,  $N$  is the number of stations,  $\Delta x_{ij}$  is the difference quantity between stations and  $d_{ij}$  is the spherical distance between stations in km.  $\Delta x_{ij}$  is either the difference between trends  $(\Delta \partial T / \partial t)_{ij}$  or the difference between 12GMT-00GMT differences at stations  $i, j$ . Strongly deviant trends/differences at densely covered areas

contribute most to the cost. The cost function  $C$  for radiosondes should reach similar values as the ERA-40 bg or satellite products after homogenization.

#### APPENDIX B. ADJUSTMENT OF THE GLOBAL MEAN BG TIME SERIES

In section 6 it has been shown that the global mean bg has spurious breaks due to insufficient satellite bias corrections. The signature of these breaks can be found throughout the globe. Although it has a horizontally smooth pattern compared to the distribution of breaks in radiosonde records, it is not constant. In general the influence of the satellite data and excessive latent heating on the bg is smaller in regions with dense radiosonde observation coverage and in the extratropics. Therefore the adjustment is scaled with the radiosonde density in the following way:

$$(6) \quad \Delta bg(\lambda, \varphi, p, t) = -\overline{obs - bg}(p, t) w(\lambda, \varphi, t)$$

where  $\Delta bg(\lambda, \varphi, p, t)$  is the global mean obs-bg difference and the weighting function  $w$  is defined as:

$$(7) \quad w(\lambda, \varphi, t) = \frac{(0.5 + \cos(\varphi))}{1.5} * \left( 1.2 - \frac{\rho(\lambda, \varphi, t)}{\rho_{max}(t)} \right).$$

The stronger weight at low latitudes helps to reduce the strong heating of the bg in the tropics which is considered excessive. The radiosonde observation density  $\rho$  is defined as:

$$(8) \quad \rho(\lambda, \varphi, t) = \sum_i^{N(t)} \exp[-d_i(\lambda, \varphi)/700 \text{ km}] f_i,$$

where  $d_i$  is the spherical distance between location  $(\lambda, \phi)$  and radiosonde  $i$ . The factor  $f_i$  takes into account whether the radiosonde reports once ( $f_i=1$ ) or twice daily ( $f_i=5$ ). It was set to 5, not 2, since sites with twice daily launches tend to be better maintained.  $N(t)$  is the number of active radiosondes and  $\rho_{max}(t)$  is the maximum radiosonde density found at a particular time. Finally the weights  $w$  are adjusted by a constant factor such that the global mean  $\Delta bg$  is equal to  $-\overline{obs - bg}$  when averaged over all radiosonde stations. Figure 18 in Haimberger (2005) shows the spatial pattern of the weight  $w$ , which shows little variation in time.

With this choice of parameters the maximum adjustment for an individual radiosonde site is about 1.6 times the global mean adjustment (e.g. in the south Pacific), the minimum adjustment is below 0.2 times the global mean adjustment (over central Europe, China). It has been tuned to reduce as much as possible the conspicuous jumps in the due to the erroneous bias correction of the NOAA-4 radiances between Jan 1975 and Sept. 1976. Figure 19 in Haimberger (2005) shows the obs-bg series averaged over the radiosondes south of 25N before and after adjustment with the weighted global mean bg. Although this simple adjustment of the bg described here can by no means remove all biases, at least the breaks in Jan 1975 and Sept. 1976 are substantially reduced in the average over this data sparse region. For other breaks, e.g. due to problems with NOAA-9, other horizontal weighting functions may be optimal, but so far this has not been tested.

## REFERENCES

- Aguilar, E., 2000: The upper air station history. a brief description. unpublished manuscript, available from National Climatic Data Centre, Asheville, NC.
- Alexandersson, H., and A. Moberg, 1997: Homogenization of Swedish temperature data. part I: Homogeneity test for linear trends. *Int. J. Climatol.*, **17**, 25–34.
- Andrae, U., N. Sokka, and K. Onogi, 2004: *The radiosonde temperature bias correction in ERA-40*, Volume 15 of *ERA-40 Project Report Series*. ECMWF.
- Bronnimann, S., 2003: A historical upper air-data set for the 1939-44 period. *Int. J. Climatol.*, **23**, 769–791.
- Christy, J., and W.B. Norris, 2006: Satellite and VIZ-radiosonde intercomparisons for diagnosis of non-climatic influences. *J. Atmospheric and Oceanic Technology*. accepted.
- Christy, J.R., 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.
- Coleman, H., M. McCarthy, and P.W. Thorne, 2006: Structural uncertainties in radiosonde temperatures. Hadley centre, Exeter, UK. Defra report No. 03.09.05.
- Compo, G.P., J.S. Whitacker, and P.D. Sardeshmukh, 2006: Feasibility of a 100-year reanalysis using only surface pressure data. *Bull. Amer. Meteorol. Soc.*, **87**, 175–190.
- Courtier, P., E. Andersson, W. Heckley, J. Pailleux, D. Vasiljević, M. Hamrud, A. Hollingsworth, F. Rabier, and M. Fisher, 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quart. J. Roy. Meteor. Soc.*, **124**, 1783–1807.
- Daan, H., 2002: *GUIDE TO THE GCOS SURFACE AND UPPER AIR NETWORKS GSN AND GUAN*. WMO, Geneva. WMO-TD 1106, GCOS-73.
- Dee, D.P., 2004: Detection and correction of model bias during data assimilation. In: *Proceedings of ECMWF seminar on recent developments in data assimilation for atmosphere and ocean, 8-12 September 2003*, pp. 65–74. ECMWF.
- Dee, D.P., and A. Da Silva, 1998: Data assimilation in the presence of forecast bias. *Quart. J. Roy. Meteor. Soc.*, **124**, 269–295.
- Ducre-Robetaille, J., L. Vincent, and D. Boulet, 2003: Comparison of techniques for detection of discontinuities in temperature series. *Int. J. Climatology*, **23**, 1087–1101.
- Durre, I., R. Vose, and D.B. Wuertz, 2006: Overview of the Integrated Global Radiosonde Archive. *J. Climate*, **19**, 53–68.
- ECMWF, 2000: IFS documentation (CY23r4). ECMWF, <http://www.ecmwf.int/research/ifsdocs/CY23r4/index.html>.
- Eskridge, R.E., J.K. Luers, and C.R. Redder, 2003: Unexplained discontinuity in the U.S. radiosonde temperature data, part I: Troposphere. *J. Climate*, **16**, 2385–2395.
- Free, M., and Coauthors, 2002: Creating climate reference datasets. *Bull. Amer. Meteorol. Soc.*, **81**, 891–899.

- Free, M., and D.J. Seidel, 2005: Causes of differing temperature trends in radiosonde upper air data sets. *J. Geophys. Res.*, **110**, D07101.
- Free, M., D.J. Seidel, J.K. Angell, J. Lanzante, I. Durre, and T.C. Peterson, 2005: Radiosonde atmospheric temperature products for assessing climate (RAT-PAC): A new data set of large-area anomaly time series. *J. Geophys. Res.*, **110**, D22101.
- Gaffen, D.J., 1996: A digitized metadata set of global upper-air station histories. NOAA technical memorandum ERL ARL-211, NOAA.
- Haeberli, C., 2006: *The Comprehensive Alpine Radiosonde Dataset (CALRAS)*, Volume 4 of *Wiener Meteorologische Schriften*. Facultas, 297 pp.
- Haimberger, L., 2005: *Homogenization of radiosonde temperature time series using ERA-40 analysis feedback information*, Volume 23 of *ERA-40 Project Report Series*. ECMWF.
- Haimberger, L., 2006: Bias correction of conventional observations. In: *Proceedings of the ECMWF-NAF workshop on bias estimation and correction in data assimilation*, RG2 9AX Shinfield Park, Reading, U.K., pp. 14. ECMWF.
- Harris, B.A., and G.A. Kelly, 2001: A satellite radiance bias correction scheme for data assimilation. *Quart. J. Roy. Meteor. Soc.*, **127**, 1453–1468.
- Hernandez, A., G. Kelly, and S. Uppala, 2004: *The TOVS/ATOVS observing system in ERA-40*, Volume 16 of *ERA-40 Project Report Series*. ECMWF.
- Hollingsworth, A., D.B. Shaw, P. Lönnberg, L. Illari, and A.J. Simmons, 1986: Monitoring of observation and analysis quality by a data assimilation system. *Mon. Wea. Rev.*, **114**, 861–879.
- Karl, T.R., S.J. Hassol, C.D. Miller, and W.L. Murray (Eds.), 2006: *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*. A report by the Climate Change Science Program and the Subcommittee on Global Change Research, Washington, DC, 180 pp.
- Kistler, R., and Coauthors, 2001: The NCEP/NCAR 50-year Reanalysis: Monthly Mean CDROM and Documentation. *Bull. Amer. Meteorol. Soc.*, **82**, 247–267.
- Lanzante, J.R., S.A. Klein, and D.J. Seidel, 2003a: Temporal homogenization of monthly radiosonde temperature data. part I: Methodology. *J. Climate*, **16**, 224–240.
- Lanzante, J.R., S.A. Klein, and D.J. Seidel, 2003b: Temporal homogenization of monthly radiosonde temperature data. part II: Trends, sensitivities, and MSU comparison. *J. Climate*, **16**, 241–262.
- Lewis, J.M., S. Lakshmiwaran, and D. S., 2005: *Dynamic Data Assimilation*. Cambridge University Press, 816 pp.
- Li, X., G. Kelly, S. Uppala, R. Saunders, and J. Gibson, 2006: *The use of VTPR raw radiances in ERA-40*, Volume 21 of *ERA-40 Project Report Series*. ECMWF.
- 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., H. Coleman, and P. Thorne, 2006: Quantifying uncertainty in radiosonde climate records: An automated method. *in preparation*.
- Mears, C.A., M.C. Schabel, and F.J. Wentz, 2003: A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Climate*, **16**.
- Nash, J., and F.J. Schmidlin, 1987: WMO International Radiosonde Intercomparison: Final Report, WMO/TD No. 195. WMO, Geneva.
- Nash, J., R. Smout, T. Oakley, and S. Kurnosenko, 2005: WMO Intercomparison of high quality radiosonde systems, Mauritius, 2-25 February 2005, Final Report. WMO, Geneva.
- Onogi, K., 2000: *The long term performance of the radiosonde observing system to be used in ERA-40*, Volume 2 of *ERA-40 Project Report Series*. ECMWF, 77 pp.

- Randel, W., and Coauthors, 2004: The SPARC Intercomparison of Middle-Atmosphere Climatologies. *J. Climate*, **17**, 986–1003.
- Randel, W.J., and F. Wu, 2006: Biases in stratospheric and tropospheric temperature trends derived from historical radiosonde data. *J. Climate*, **19**, 2094–2104.
- Redder, C.R., J.K. Luers, and R.E. Eskridge, 2004: Unexplained discontinuity in the U.S. radiosonde temperature data. part II: Stratosphere. *J. Atmos. and Oceanic Technology*, **21**, 1133–1144.
- Santer, B., and Coauthors, 2004: Identification of anthropogenic climate change using a second-generation reanalysis. *J. Geophys. Res.*, **109**, D21104.
- Santer, B.D., and Coauthors, 2005: Amplification of surface temperature trends and variability in the tropical atmosphere. *Science*, **309**, 1551–1556.
- Seidel, D.J., and Coauthors, 2004: Uncertainty in Signals of Large-Scale Climate Variations in Radiosonde and Satellite Upper-Air Temperature Datasets. *J. Climate*, **17**, 2225–2240.
- Sherwood, S., J. Lanzante, and C. Meyer, 2005: Radiosonde daytime biases and late 20th century warming. *Science-express*. doi: 10.1126/science.1115640309, published online 11 August 2005.
- Thorne, P., D.E. Parker, S.F.B. Tett, P.D. Jones, M. McCarthy, H. Coleman, P. Brohan, and J.R. Knight, 2005a: Revisiting radiosonde upper-air temperatures from 1958 to 2002. *J. Geophys. Res.*, **110**, D18105.
- Thorne, P.W., and Coauthors, 2005b: Vertical profiles of temperature trends. *Bull. Amer. Meteorol. Soc.*, **86**, 1471–1476.
- Trenberth, K.E., and L. Smith, 2006: The vertical structure of temperature in the tropics: Different flavors of El Nino. *J. Climate*, **19**. in press.
- Uppala, S., G. Kelly, B.K. Park, P. Kallberg, and A. Untch, 2006: Experience in estimation of biases in ECMWF reanalyses. In: *Proceedings of the ECMWF/NWP-SAF workshop on bias estimation and correction in data assimilation*, RG2 9AX Shinfield Park, Reading, U.K., pp. 17. ECMWF.
- Uppala, S.M., and Coauthors, 2005: The ERA-40 Re-analysis. *Quart. J. Roy. Meteor. Soc.*, **131**, 2961–3012.
- von Storch, H., and F. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press.