ERA-40 Project Report Series

No. 23 Homogenization of radiosonde temperature time series using ERA-40 analysis feedback information

---

Leopold Haimberger

Series: ECMWF ERA-40 Project Report Series

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

ERA-40 Project Report Series No. 23

Homogenization of radiosonde temperature time series using ERA-40 analysis feedback information

Leopold Haimberger[1]

June 2005

[1] Current affiliation:
Institut für Meteorologie und Geophysik, Universität Wien
Althanstraße 14, A-1090 Wien, Austria

Contents

Abstract

This report describes a method developed for detecting and removing artificial breaks in radiosonde time series. These breaks affect the applicability of radiosonde time series for climate change detection. They also affect the quality of reanalyses assimilating these data, particularly in the pre-satellite era. While radiosondes measure temperature, humidity and wind, the analysis here is restricted to temperature time series.

The adjustment method, referred to as RAdiosonde Observation Bias Correction using Reanalyses (RAOBCORE), works fully automatically. It uses time series of differences between radiosonde observations (referred to as *obs*) and the 6h background forecasts from the ERA-40 assimilating model (referred to as *bg*) as principal source of information about artificial breaks. These differences are available in the ERA-40 analysis feedback files.

More than 1500 stations have been investigated. Not only has the radiosonde archive available in the ERA-40 analysis feedback been used, but also the Integrated Global Radiosonde Archive (IGRA) which is the successor of the Comprehensive Aerological Reference Dataset (CARDS). Some aspects of data quality and content of these archives are discussed.

It is demonstrated that time series of bg-obs differences allow detection of many breaks in radiosonde temperature time series that have been previously undetected. The probability of a breakpoint is primarily estimated by applying a variant of the Standard Normal Homogeneity Test (SNHT) to time series of bg-obs differences of tropospheric and stratospheric layer mean temperatures. Time series of stratospheric layer mean obs(12GMT)-obs(00GMT) differences are used as well. The probabilities from the SNHT are combined with a-priori information about changes of observation practice (metadata) to yield an objective score. If this score exceeds a threshold, a breakpoint is detected.

The calculation of adjustments to remove artificial breaks from the time series is more difficult, since breaks in the bg-obs difference series can be caused both by the bg forecasts (mainly due to changes in the satellite observing system) and by the tested radiosonde time series. It has been necessary to adjust the global mean bg temperature time series prior to using it as a reference, in order to remove the effects of suboptimal satellite radiance bias corrections from the bg time series. After the adjustment of the global mean bg the breaks at individual radiosondes are removed. The difference between the annual mean bg-obs difference before and after a diagnosed breakpoint is used as the best estimate for the adjustment.

It is demonstrated that RAOBCORE considerably improves the spatiotemporal consistency of the global radiosonde dataset. Temperature trends from neighboring radiosonde stations are much more consistent than before the adjustment. Many known biases in the radiosonde temperatures that exist even in the most recent periods are removed. There is evidence that this automatic adjustment algorithm yields physically reasonable results in a global mean sense. The effect of the adjustments on zonal mean and global mean radiosonde temperature trends is discussed in some detail and recommendations for further improvements of RAOBCORE are given.

1. Introduction

This reports described results achieved within a one-year Marie Curie Intra-European Fellowship (see http://www.cordis.lu/fp6/mobility.htm) which has been awarded to the author and was carried out at ECMWF. The aim of this fellowship was to develop software for the automatic homogenization of radiosonde time series using the ERA-40 analysis feedback (AF) dataset. Such a homogenized dataset is highly desirable for climate change analysis as well as for future reanalysis efforts. The quality of time series from reanalyses depends heavily on the quality of the observation input (Karl et al. 1995). Not only in the pre-satellite era is it important to remove possible biases from radiosonde observations since the bias correction of the early satellite observing systems is dependent on radiosonde observations.

Radiosondes have been primarily used for weather forecasting applications such as thunderstorm prediction and daily upper air analyses, not for climate purposes. The quality of the instruments has been constantly improved by reducing measurement errors, particularly at high altitudes. Radiosonde intercomparison experiments in the 1980s (Nash and Schmidlin, 1987, Richner and Philips, 1982) have shown that there were quite substantial differences between measured temperature values from different radiosondes, even when flown on the same platform. This can be seen from the coloured bullets in the upper panel of Figure 1, which indicate the temperature difference between 12GMT and 00GMT radiosonde temperature observations at the 50 hPa level, averaged over the period 1988-1990. This difference, which exceeds 2K at some stations, should yield a spatially much more coherent picture, more like that from the ERA-40 background forecast time series shown in the lower panel of Figure 1. The spuriously large differences are mainly related to radiation errors, as can be most clearly seen over the US. The diurnal cycle of the solar elevation is indicated. Changes from one radiosonde type to another therefore almost certainly introduce breaks in the corresponding time series. As will be shown below, these breaks lead to a rather heterogeneous picture of temperature trends at individual stations as well.

The task of removing these artificial breaks from a time series is referred to as *homogenization*. Many homogenized historic surface temperature records have become available in the last decade. Reviews of the methods available can be found in Peterson et al. (1998), Aguilar et al. (2004). Efforts to generate such a homogenized radiosonde temperature dataset started in the 1990s, when radiosonde observations as well as information on radiosonde types in use were collected in the Comprehensive Aerological Reference Data Set (CARDS, Eskridge et al. 1995). Soon afterwards the first homogenized temperature datasets based on CARDS were published (Parker et al. 1997; Lanzante et al. 2003a,b, referred to as LKS). These datasets contain, however, much fewer stations (87 stations by LKS, 477 stations by Parker et al.) and there are sometimes substantial differences between the modifications actually applied to these stations. LKS used only information contained in one station for detecting breaks at that station. Parker et al. (1997) calculated composites of neighboring stations in order to compare them with a station to be tested. They used also temperature records from the satellite-borne Microwave Sounding Unit (MSU) as a reference from 1979 onwards. Recently a new gridded homogenized radiosonde temperature dataset based on CARDS monthly mean data has become available (Thorne et al. 2005). While these datasets are important achievements, they are not suitable yet for future reanalyses since only a fraction of the available radiosonde records is adjusted. None of these datasets has been created by automatic procedures. Therefore any further 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). Therefore there is still large uncertainty on upper air trends and low frequency variability (Seidel et al. 2004).

With the completion of the ERA-40 project (Uppala, 2003; Simmons, 2003; Uppala et al. 2005), a 45year global time series of analyses (an) and 6h background forecasts (bg) has become available. The analyses have been used already successfully for climate change detection purposes (Santer et al. 2004). While the analyses are the most useful product for many research applications, the differences between bg forecasts and the original observations (obs) have proven most useful for the detection of systematic observation errors (Hollingsworth et al., 1986). It is the working hypothesis of this report that time series of bg-obs differences can be used also for homogenization purposes. Arguments in favor of this approach are:

1. Bg-obs differences are routinely calculated for every radiosonde station assimilated, since the bg-obs differences (also called the *innovations* in the data assimilation process) contribute directly to the cost function to be minimized (see e.g. Courtier et al., 1998 and Simmons, 2003 for a description of the 3D-Var data assimilation system used in ERA-40). Therefore much effort is put into accurate interpolation of the model state on the model grid to the observation points.
2. Since the bg forecasts are generally quite accurate, 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. The quality of the bg forecasts allows fairly stringent quality control procedures to be applied to the radiosonde observations. Quality control flags are available in the AF and can be used to eliminate suspicious data before performing the time series analysis.
3. Analysis minus observation differences are not very useful for the detection of systematic observation errors since the analyses themselves depend on the (biased observations). The bg forecasts at a particular site are much less dependent on the observations to be tested since the information for the bg forecasts comes from different places and also from earlier observations. The bg-obs difference is also more directly related to systematic observation errors than the an-obs difference (see e.g. Dee, 2003). The weak dependence of the bg forecasts on individual radiosonde observations is indicated in Fig. 1. The spatial pattern of bg(12GMT)-bg(00GMT) differences is much smoother than the pattern from the uncorrected radiosonde observations. Therefore the analysis of 6h bg-obs temperature difference time series seems well suited for homogenization purposes.

Analysis feedback data have been used for developing the ERA-40 radiosonde bias correction as documented in Andrae et al. (2004), referred to as ASO. While the adjustments applied have been conservative and mainly the radiation error has been corrected, a noticeable positive impact of the bias correction has been found. First results of using bg-obs differences for break detection over Central Europe (Haimberger, 2004) corroborate the impression that the AF can be used for correcting radiosonde time series.

The approach developed by ASO estimates radiosonde temperature biases for composites of radiosonde stations as a function of solar elevation angle and pressure. The composites are defined such that the radiosondes have the same manufacturer and are maintained by the same country. Large countries have been divided into regions, small countries have been combined if possible. This has the advantage that the method works with data from a short time period (~1 year). The adjustment amounts calculated from one year have been used for calculating the bias corrections used in the following year. This procedure has been adopted since in contrast to now no adequate time series of analysis feedback data have been available during the production of ERA-40 (Onogi, 2000). The method has the disadvantage that inhomogeneities specific to individual radiosondes are not removed since only composites are considered. Further, it is dangerous to use this method for eliminating temperature biases not related to 12GMT-00GMT observation differences. While

it can be justified in most cases that the 12GMT-00GMT difference of the background forecasts is unbiased, this is not the case for the daily mean bg temperatures. At many occasions below, the results achieved with the approach of ASO will be used for intercomparison.

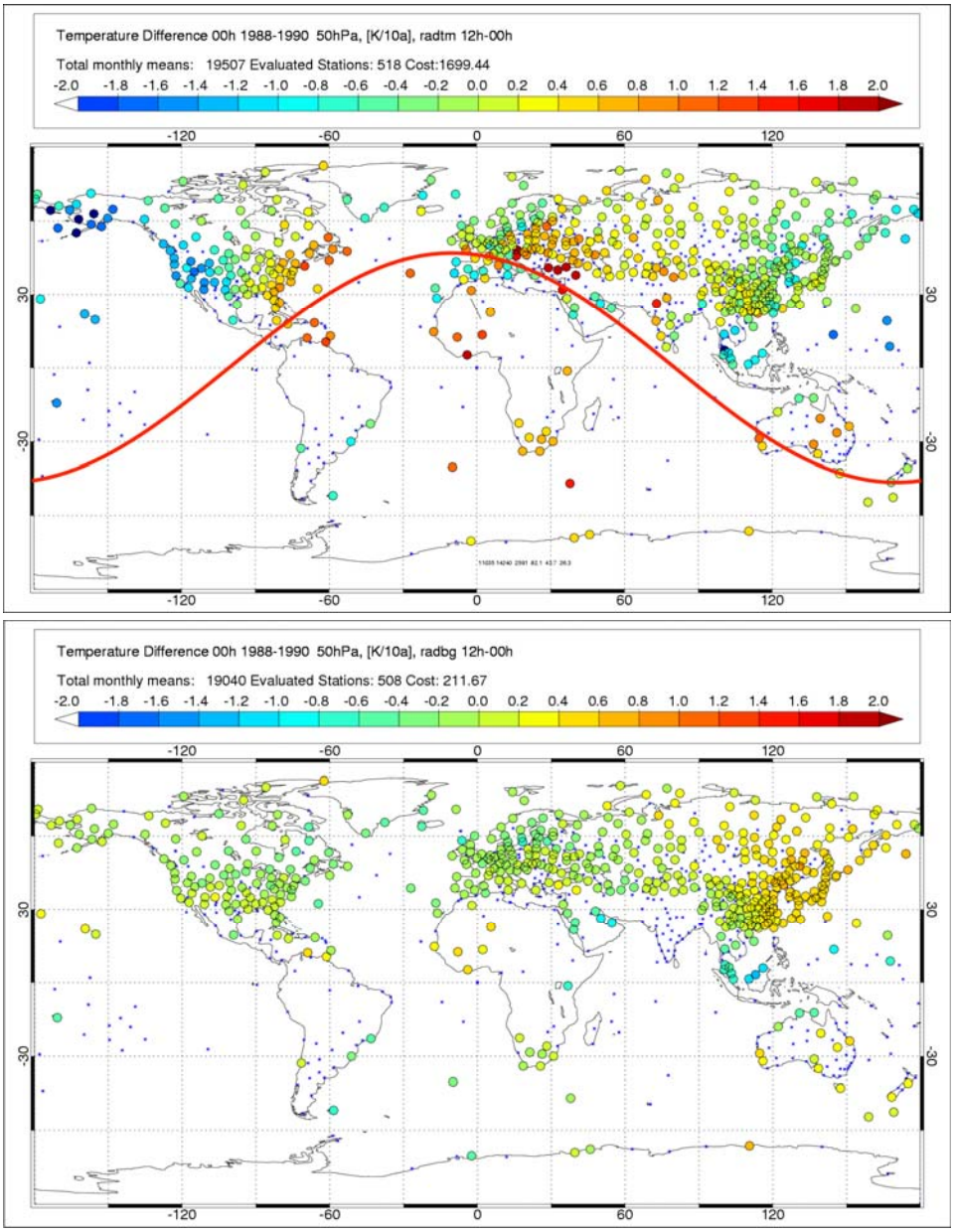

Figure 1: 12GMT-00GMT temperature differences at 50 hPa averaged over period 1988-1990. Each bullet denotes a radiosonde station; the colour of the bullet indicates the difference. The blue stars depict radiosonde stations with insufficient number of launches for trend calculation in the specified time interval. Upper panel shows the difference as measured by radiosondes. The sine wave indicates the solar elevation at 12h40' GMT. Lower panel shows the bg 12GMT-00GMT differences which are spatially much more homogeneous.

Since long time series of bg and obs are available now, a new automatic method based on time series analysis of the stations can be used. The method is referred to as *Radiosonde Observation Bias Correction using Reanalyses* (RAOBCORE). It is designed to be capable of correcting not only radiation errors but also observation biases that are not a function of solar elevation. The reasons for temperature biases other than radiation errors are (i) sensors with slow response or insufficient ventilation, (ii) too short cordless between

balloon and sensor, (iii) biases of the pressure sensors. A bias of 1 hPa of the pressure sensor corresponds to a displacement of the 200 hPa level by 50 m. Assuming a standard atmosphere lapse rate of 6.5K/km this corresponds to a temperature bias of 0.35K. Further it is often difficult to assess the radiation error from 00/12GMT ascents if there is little variation in the solar elevation angle, e.g. when (i) only one ascent per day is available, (ii) the radiosondes are launched at dawn/dusk, (iii) in the tropics.

The main assumption of RAOBCORE is that the difference between bg and obs is *stationary* in time (apart from an annual cycle) if there is no artificial break, even if the climate changes. If the bg-obs difference is not stationary, there must be an artificial break either in the bg or in the obs. This assumption is made by any method based on relative homogeneity tests. One should note that in contrast to the method described in Andrae et al. (2004) it is not necessary to assume that the bg is unbiased.

The next section describes the available input datasets in more detail. Section 3 and section 4 contain a description of break detection and break adjustment procedures used by RAOBCORE. Results from the adjustment process are shown in section 5.

2. Input data

The main input data source for this study was the BUFR-coded ERA-40 AF dataset from 1957-2002 as it is available in the MARS archive. The AF dataset contains all observations presented to the ERA-40 data assimilation system plus quality control flags plus the bg-obs and an-obs differences mentioned above. The subset of TEMP observations plus the feedback information contains about 150GB of data from more than 2000 different stations. About 900 of these stations have records containing more than 10000 ascents.

Recently the Integrated Global Radiosonde Archive (IGRA, http://www1.ncdc.noaa.gov/pub/data/igra/, Durre et al. 2005), which contains quality-controlled data from 1536 upper air stations back to the 1930s has become available. The IGRA and ERA-40 datasets are not identical, as can be seen from Figure 2. ERA-40 contains more data in the 1960s and 1970s, whereas IGRA contains more data in the 1990s. In the case of ERA-40 only the data accepted by the ERA-40 assimilation system have been counted. The relatively high rejection rate of radiosonde observations at 10 hPa in ERA-40 is the likely reason why the difference between IGRA and ERA-40 is smaller at this level.

Both the IGRA and ERA-40 radiosonde observations have been quality controlled. While in IGRA mainly a climatological check has been performed, a generally more stringent quality control algorithm has been applied to the radiosonde observations used in ERA-40. Only those observations that have passed the variational quality control in ERA-40 have been used in the present study. The differences are mostly subtle, but there are examples where data that have not been flagged in IGRA have been rejected by the ERA-40 quality control procedure (e.g. station Tamanrasset, see section 5.8)

The ERA-40 AF data are organized such that a feedback file contains all observations at one analysis time. They had to be reorganized into time series, which was not always simple because of changing or incomplete message headers. The observation identification algorithm used for calculating the corrections described by Andrae et al. (2004) has been used also for this report. Both for the ERA-40 AF dataset and the IGRA dataset, time series with launch resolution (=daily series) as well as time series of monthly means were generated. 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 in this report.

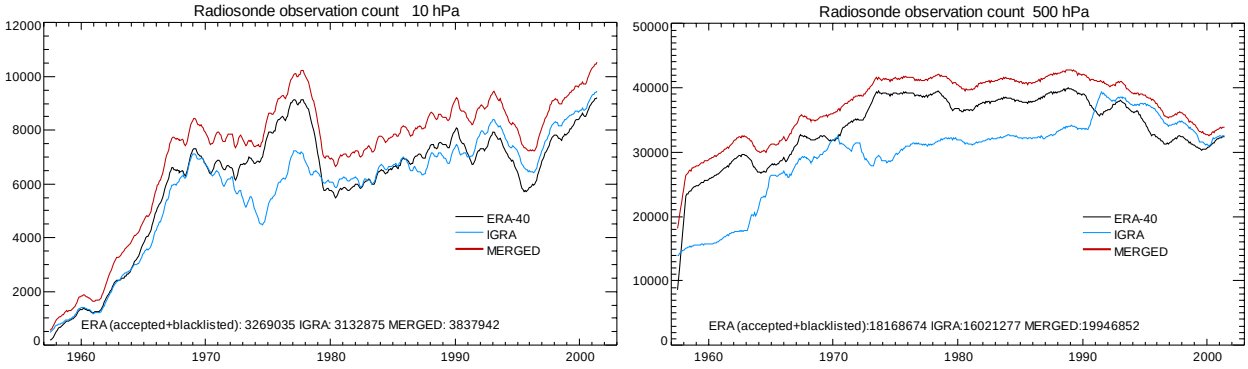

Figure 2: Time series of monthly global radiosonde temperature observation count 1958-2001 at the 10 hPa level (left panel) and at the 500 hPa level (right panel, note different scale). Only TEMP-observations from the 1536 IGRA stations have been included. TEMP-Ship observations which are also part of the ERA-40 archive, have not been included. Blue Curve: Count from IGRA dataset. Black Curve: Count in ERA-40 Analysis Feedback files. Red Curve: Count of merged ERA-40+IGRA dataset.

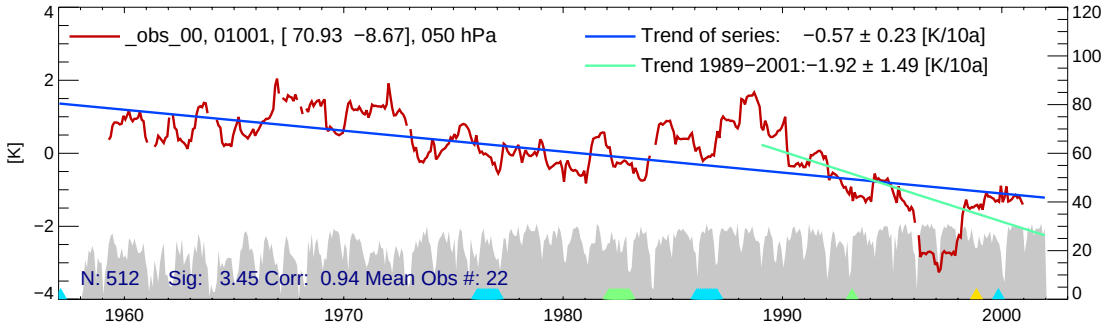

Figure 3: Temperature anomaly (red curve) and monthly observation count (grey shaded area) in ERA-40 for station Jan Mayen at 00GMT at the 50 hPa level. Anomaly is filtered with 2-yr running mean, trends as indicated. Blue triangles/trapezia indicate radiosonde changes, green triangles indicate radiation correction changes, yellow triangles indicate ground equipment changes as documented in CARDS.

Figure 3 shows the observation count time series together with the temperature anomaly for station Jan Mayen at the 50 hPa level. It shows a cooling trend over the whole period. All trends shown in this report are linear trends calculated by standard least squares. Jan Mayen has been used as example throughout this report, because it has the lowest station ID, because it has a long record (00GMT ascents from 1958 onwards, 12GMT ascents from 1963 onwards) and since it is fairly remote. The 50 hPa level has been chosen throughout this report since it is reached by many ascents even in the 1950s, although data availability is a challenge. Further the 50 hPa temperature measurements are already quite sensitive to radiation errors and pressure sensor errors.

Before IGRA data and the ERA-40 AF data were compared, the IGRA data had been supplemented with ERA-40 feedback information. The bg-obs differences and an-obs differences had been calculated by bilinear interpolation from the N80 Gaussian grid used in ERA-40 to the radiosonde stations. Figure 4 and Figure 5 show that feedback information can be calculated “offline” for data not used in ERA-40, at least for TEMP data (for surface and satellite data this may be more challenging or impossible). The differences between IGRA and ERA-40 are not large at this station, except before 1964 when there are no observations in the IGRA dataset. Figure 5 shows that the interpolation error is very small at this station.

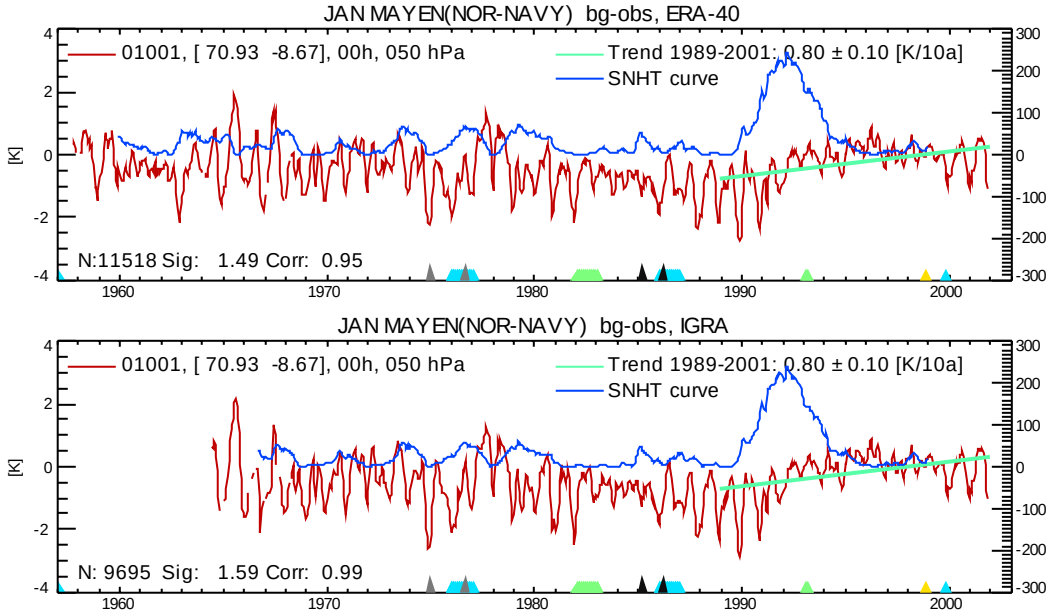

Figure 4: Daily bg-obs differences as available in the ERA-40 AF (upper panel) and as calculated for IGRA data (lower panel), again for station Jan Mayen at 00GMT. Red curves are 50day running means. Green lines are trend lines as before. Blue curves (right axes apply) are test statistics calculated with 3-year moving average SNHT (see section 3). Maxima above 20 indicate breaks in the difference series. Grey spikes at bottom of panels indicate breakpoints in the ERA-40 bg forecasts due to changes in the ERA-40 (mainly satellite) observing system. The legends in the upper left corners of the panels can be deciphered as: 01001=WMO station ID, [70.93 -8.67] = station lat/lon, 00h= nominal launch time, 050 hPa = pressure level. The numbers in the lower left corners of the panels are: N= number of values in the shown time series, Sig=standard deviation of the time series, Corr is Pearson's correlation coefficient between two time series (e.g. bg-obs) if the time series shown is the difference between the two.

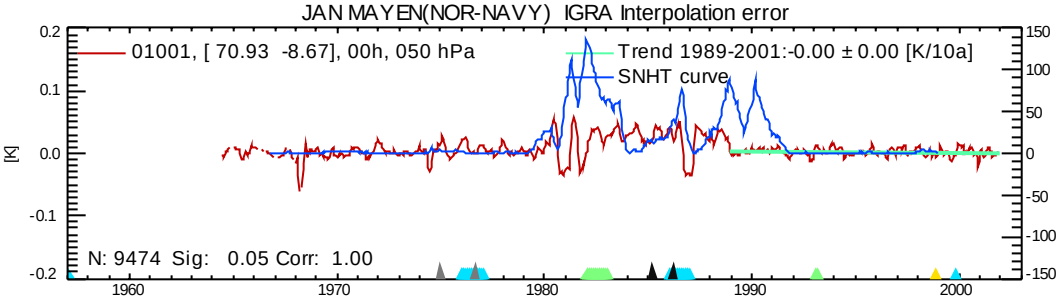

Figure 5: Difference between the ERA-40 AF bg-obs difference and the IGRA bg-obs difference at Jan Mayen. Difference may be not zero since the ERA-40 bg-obs differences are calculated using First Guess at Appropriate Time(FGAT, see Uppala et al. 2005), whereas the IGRA bg-obs differences are calculated offline from bg fields valid at 00GMT linearly interpolated to observation location. Note different scale compared to Figure 4. Increase in variability in the 1980s is due to slightly different station coordinates reported in the ERA-40 records.

Following a suggestion by P. Thorne (pers. comm.) also the 12h forecasts (*fc*) computed twice daily in ERA-40 have been used to calculate forecast-obs differences for the available IGRA stations. A comparison of the 00GMT fc-obs differences (Figure 6) with the bg-obs differences (Figure 4) shows that the standard deviation of the fc-obs differences is slightly smaller than the standard deviation of the bg-obs differences. The reason is that the 12h-forecasts started from analyses that were more strongly influenced by radiosondes than the 6h forecasts used for calculating the background, at least in the satellite era.

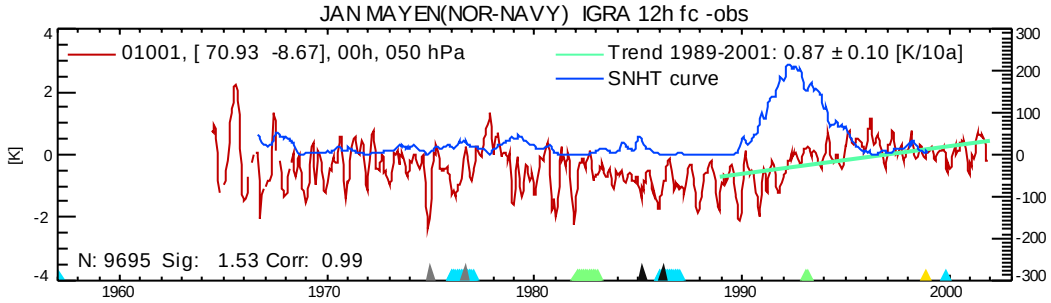

Figure 6: 12hour-forecast-obs difference at Jan Mayen calculated for IGRA data. Difference is slightly smaller than bg-obs differences shown in Figure 4 (standard deviation is 1.59 compared to 1.53). Similar difference time series for ERA-40 AF has not been calculated yet.

For the detection of breaks in radiosonde time series, so-called metadata containing information about the instrumentation and other events such as relocations or changes of observation practice are quite useful. Within the CARDS project a digitized metadata dataset has been made available (Aguilar, 2001), which can be used as information source for an automatic break detection/correction algorithm. If the date of a change is known exactly or if the month is known, a coloured spike is drawn in the time series plots throughout this report. If only the year of change is known, a trapezium covering a one year interval in the plots is drawn. There are also metadata (e.g. satellite launch dates) for the ERA-40 bg time series. The grey/black spikes in the time series plots denote locations of breaks in the global mean bg time series (see section 5.1 below).

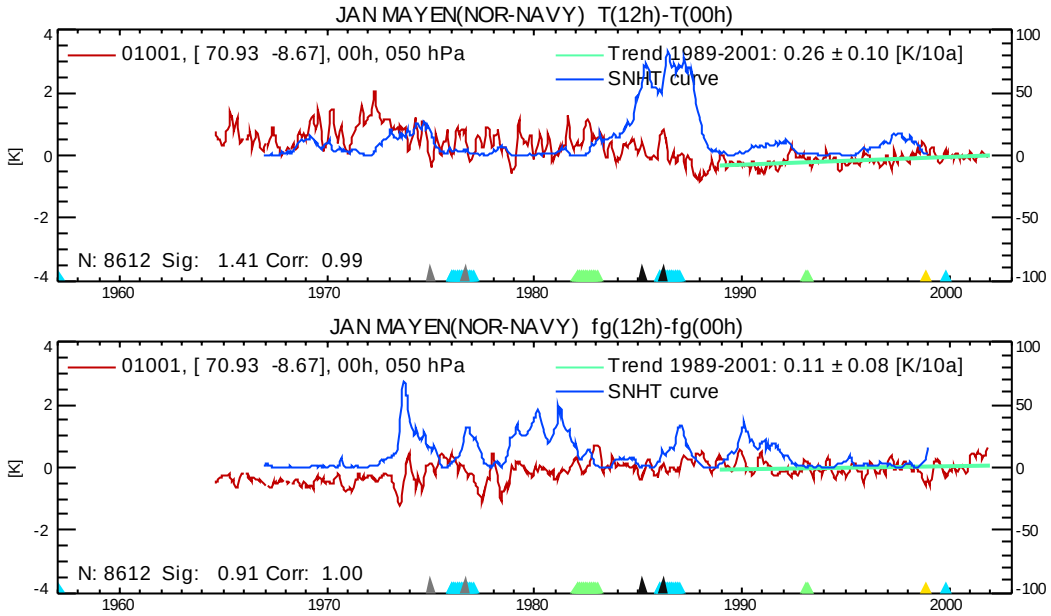

Figure 7: 12GMT-00GMT temperature difference at Jan Mayen, calculated from observations (upper panel) and from ERA-40 bg forecasts (lower panel). At 00GMT there is mostly darkness at this station, except around the summer solstice.

The bg-obs differences are not the only time series yielding useful information about breaks in the radiosonde time series. If available, time series of the differences between 00GMT and 12GMT ascents yield valuable additional information about the radiation errors of a radiosonde. Note that these differences are independent of the ERA-40 bg.

Figure 7 shows this difference for the temperature observations and for the bg temperatures at Jan Mayen. If the 12-00GMT obs difference series shows a break due to changed radiation bias correction but the 12-00 bg

difference does not show this break, this is a clear indication that (i) there is a break in the radiosonde series and that (ii) the bg is indeed nearly independent of the radiosonde to be tested. The time series shown in panel a) indicates such a break in 1987 that is only visible in the observation data but not in the bg data. The bg(12GMT)-bg(00GMT) difference is affected by breaks due to satellites in the 1970s, however. The dependence of the obs(12GMT)-obs(00GMT) differences on the solar angle is low at this station. There is little evidence of an annual cycle in Figure 7, although the solar elevation varies substantially at 12GMT at the latitude of Jan Mayen. Stations with much larger radiation errors are discussed in section 5.6.

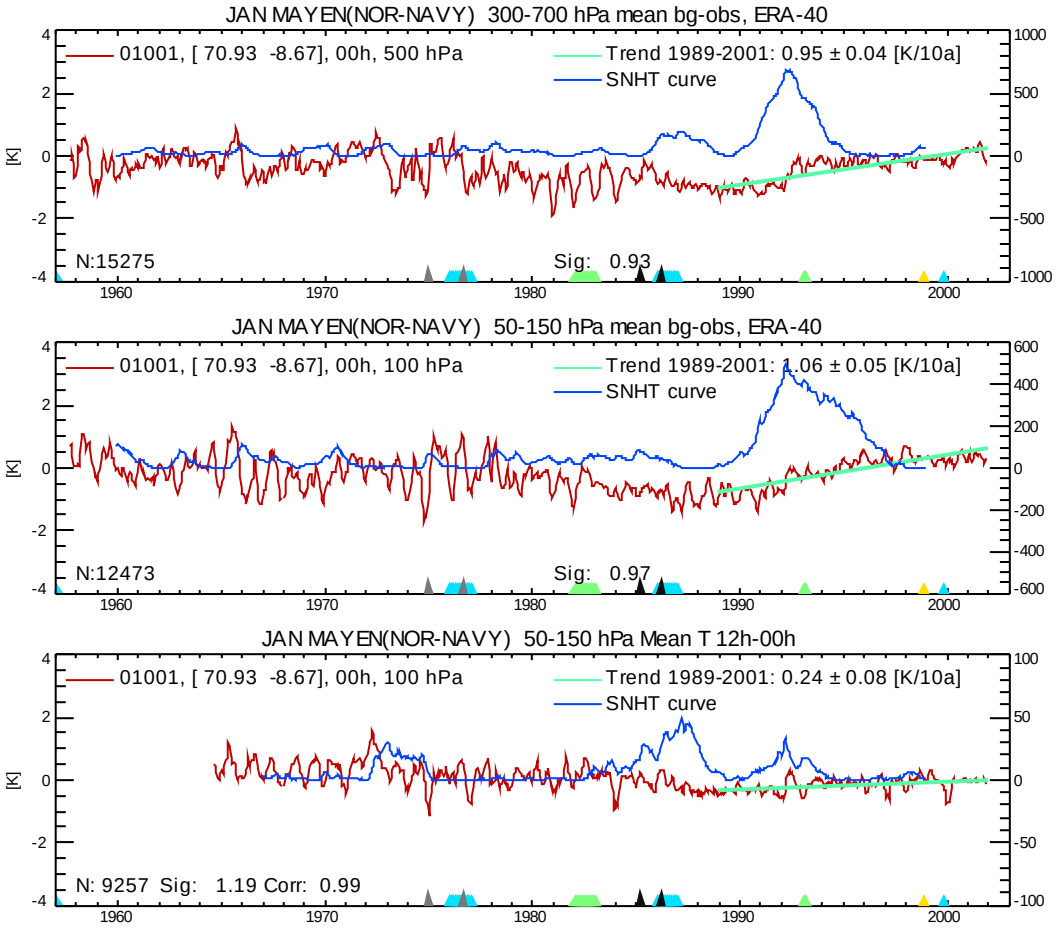

Figure 8: Layer mean series for tropospheric bg-obs (300-700 hPa, upper panel), b) stratospheric bg-obs (50-150 hPa, middle panel) and stratospheric obs(12GMT)-obs(00GMT) differences (lower panel) at station Jan Mayen. These three time series plus the two bg-obs series at 12GMT (not shown) are used for break detection at an individual station as discussed below.

Vertical averaging of the bg-obs differences or other differences is a simple but effective tool for finding vertically coherent breaks. The vertical averages are obtained here by integrating over the logarithm of pressure. Such temperature averages are practically equivalent to the geopotential difference between lowest and highest level of the averaging interval. In fact one could use the geopotential feedback information, which is also available, for calculating these temperature averages. It has been preferred to calculate the mean temperatures from the temperature feedback information, however, since at some stations breaks in the geopotential series are introduced by changes in the method for calculating the geopotential. The averaged temperature series cannot be affected by such a problem. Further the geopotential information has been blacklisted in ERA-40 and therefore outliers, which could affect the statistics, are not flagged.

Figure 8 shows vertical layer mean bg-obs differences and vertical mean obs12-obs00 differences for Jan Mayen. The standard deviations involved are smaller than for single levels and some of the breaks show up more clearly, especially those in 1987 and in 1992. The detection algorithm below is based on time series of stratospheric and tropospheric layer mean temperatures as shown in Figure 8. It has been decided to analyze layer mean temperatures of troposphere and stratosphere separately, since there are sometimes compensating errors in the troposphere and stratosphere. The obs(12GMT)-obs(00GMT) difference time series, which is useful for the detection of radiation errors, is only analyzed in the stratosphere.

3. Analysis tools

The difference series have been analyzed with a variant of the Standard Normal Homogeneity Test (SNHT, Alexandersson and Moberg, 1997) described below. Ducre-Robitaille et al. (2003) recently compared popular homogeneity tests and the SNHT performed well in this intercomparison. The results of this comparison are not simply applicable, however, for the analysis of radiosonde records. In particular they have assumed stationary random (or stochastic) series between breaks in their test procedure, which is valid for long-term annual mean surface temperature or precipitation records.

Radiosonde records pose a different problem since breaks occur much more frequently and therefore the assumption of stationary random time series before and after a break is often not valid. 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 data. Homogeneity tests relying on trend estimates (e.g. Easterling and Peterson, 1995) or autocorrelations (Vincent, 1998) do not work well in these circumstances.

The availability of the ERA-40 first guess allows avoidance of the cumbersome generation of a reference series from radiosondes surrounding the radiosonde site to be tested. Instead it is argued that the ERA-40 bg interpolated to the radiosonde site can be used as reference series. Methods that work with mutual testing of records from a sample of nearby sites against each other (Peterson et al. 1998, Caussinus and Mestre, 2004) are not considered at this time. They tend to be difficult to apply to the global radiosonde network since for remote stations the correlation between neighboring stations is often rather low.

3.1 A variant of the Standard Normal Homogeneity Test (SNHT)

The original SNHT (Alexandersson and Moberg, 1997) calculates the series of differences between the tested series and a reference series. 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 T_k can be calculated:

$$T_k = [k(\mu_1 - \mu)^2 + (N-k)(\mu_2 - \mu)^2]/\sigma$$

N is the sample size, μ_1 is the mean of the subsample before *k*, μ_2 is the mean of the subsample after *k*, μ is the mean of the whole sample and σ is the sample standard deviation. The difference series is assumed to be homogeneous if the maximum value T_k^s of all possible values *k* stays below a significance level. If it is above the threshold somewhere in the interval, the series is not considered homogeneous. The inhomogeneity may be caused either by jumps or by trends or by to some periodic signal.

If the inhomogeneity is caused by a jump, the test can be shown to be optimal if the series before and after the break is stationary random and normal distributed. The difference of μ_1 - μ_2 calculated at the point k^s

where T_k^s occurs is the best estimate for the magnitude of the break. The point k^s is the most probable location of the breakpoint.

The original SNHT has two drawbacks. Firstly, it tends to indicate breaks near the edges of the time interval, simply because either μ_1 or μ_2 tend to be poorly estimated as either *k* or *N-k* get small. Secondly the position of a breakpoint is poorly estimated in the presence of a periodic signal. To overcome these problems the SNHT has been modified such that μ_1 and μ_2 are calculated as three-year moving averages. 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*.

$$T_k = [N/2(\mu_1 - \mu)^2 + N/2(\mu_2 - \mu)^2]/\sigma$$

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 if no data are missing. The elimination of the annual cycle proved to be important since it is difficult to distinguish statistically between nonstationarities in time series caused by a break or by an annual wave; particularly if the analysed time interval needs to be short (3 years have been used in most cases in this report). The maximum value T_k^s 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 $[k+N/2, k-N/2]$.

Using three-year running means for the SNHT has not been sufficient to avoid false break detections due to the presence of an annual cycle. Quite often missing data are concentrated in the cold season, especially in high latitudes and in high altitudes, partly because the balloons burst more easily at very cold temperatures and because difficulties in launching the radiosondes are more frequent under harsh winter conditions. These data gaps can also introduce a bias if more data are missing in one of the two samples compared in the SNHT (see Figure 11 below). Therefore the data before and after the midpoint *k* have been binned into 12 bins, one for each month. If the data count of e.g. the January values in the earlier 3-year interval was less than the count of January values in the later 3-year interval, the excessive January values in the later interval were removed. This measure helped to reduce the false detection of breaks considerably. The version of the SNHT that ensures equal sampling of the annual cycle in the intervals before/after a potential breakpoint is referred to as *equal sampling SNHT*. It yields identical results as the moving average SNHT if no data are missing. In RAOBCORE the equal sampling SNHT has been used, while in the blue curves shown in the time series plots in this report the moving averaging SNHT, which is much faster, has been used. It should be noted that the periodic signal cannot be removed by using anomalies, since the amplitude of the periodic signal (e.g. of the radiation error, which often has an annual cycle) can change significantly with an instrument change.

Figure 9 shows examples of the behaviour of the moving average SNHT test statistic and related quantities under controlled conditions with synthetic data. The upper left panel depicts an N(0,1) normal distributed random series with no break. The orange T_k curve in the upper left panel varies erratically but does not reach a high value. The upper second left panel shows the distribution of the maximum values T_k^s in the considered interval, given that no break occurs. The distribution has been gained from a sample of 5000 random series and it is close to normal. From this distribution the significance levels for rejecting the null hypothesis of a homogeneous time series (T_k^s >9.6 for the 95% and T_k^s >12.5.for the 99% level) can be derived.

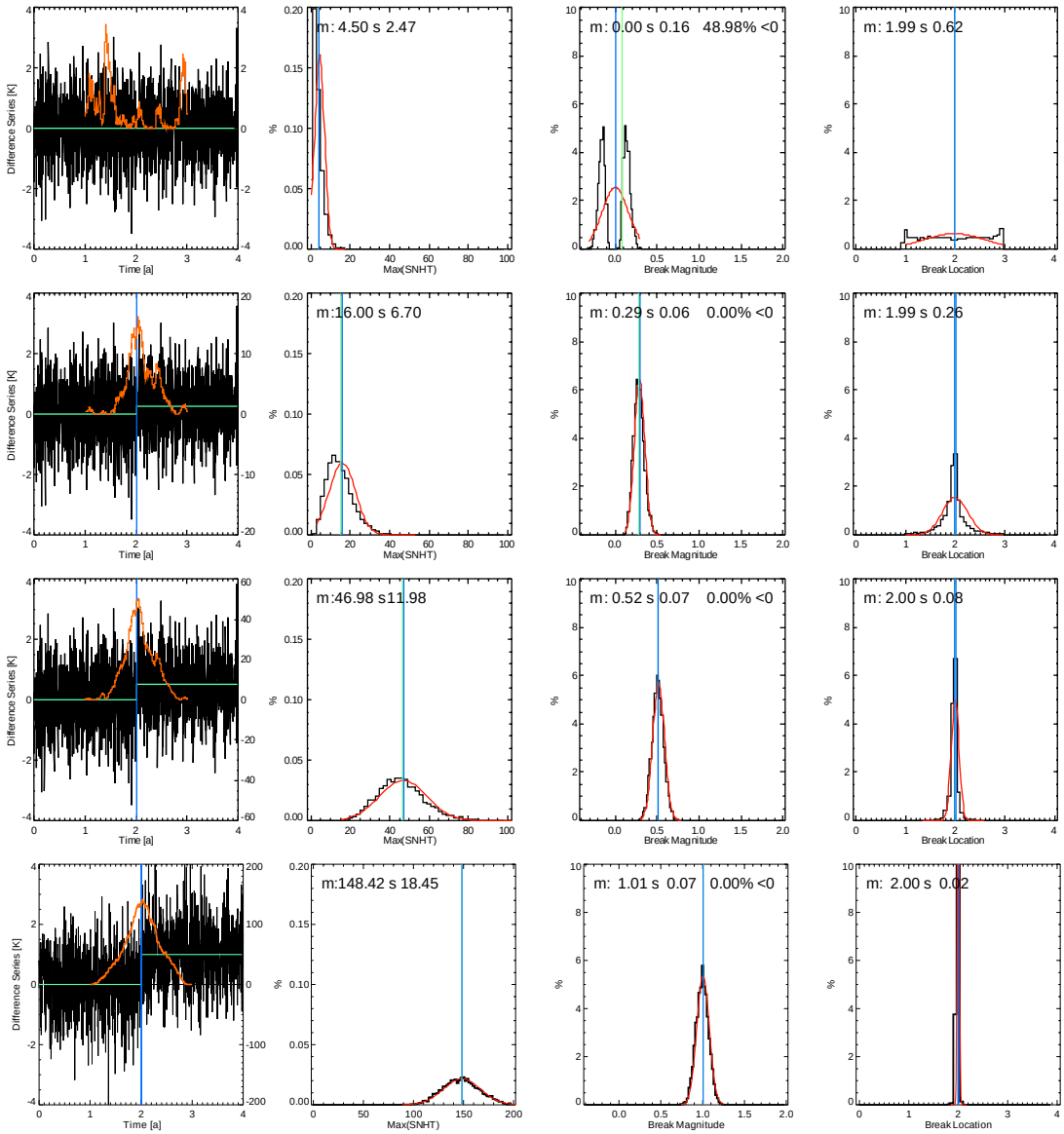

Figure 9: Moving average SNHT applied to random series with shifts of 0, 0.25, 0.5 and 1 K and a residual variance of 1 K^2. Left panels show series with break and one instance of the SNHT test parameter. A time period of 4 years is simulated with a break right in the middle. The SNHT in this example uses 1 year moving averages. Second left panels show histograms of the distribution of the T_k^s, generated with 5000 different random series. The red curves are fitted Gauss functions. Vertical lines are means and medians, respectively. Third panels show distribution of diagnosed size ΔT of break, right panels show distribution of diagnosed break location $t^s(k^s)$. Figures in panels indicate sample means and standard deviations.

The third panel shows the distribution of the difference between the moving averages at the point k^s where the maximum T_k value is achieved. It is bimodal for a homogeneous series since T_k is large if the difference between the averages is either positive or negative but large. The right panel shows the distribution of the break location k^s for the case of no breaks. It is close to uniform.

In the middle and lower panels of Figure 9, breaks of size 0.25, 0.5 and 1 have been added to the random series. It can be seen that the values of the maxima T_k^s become larger, and that the distribution of differences ΔT becomes more normal. The distribution of the detected break locations k^s becomes much narrower with

break size. Breaks with a size of 0.5σ are practically always detected since more than 99% of the T_k^s values reach values above 20 if there is a break with size 0.5σ. Breaks with 0.25σ are detected in about ¾ of the cases if T_k^s=12.5 (~99% significance level) is used as a threshold.

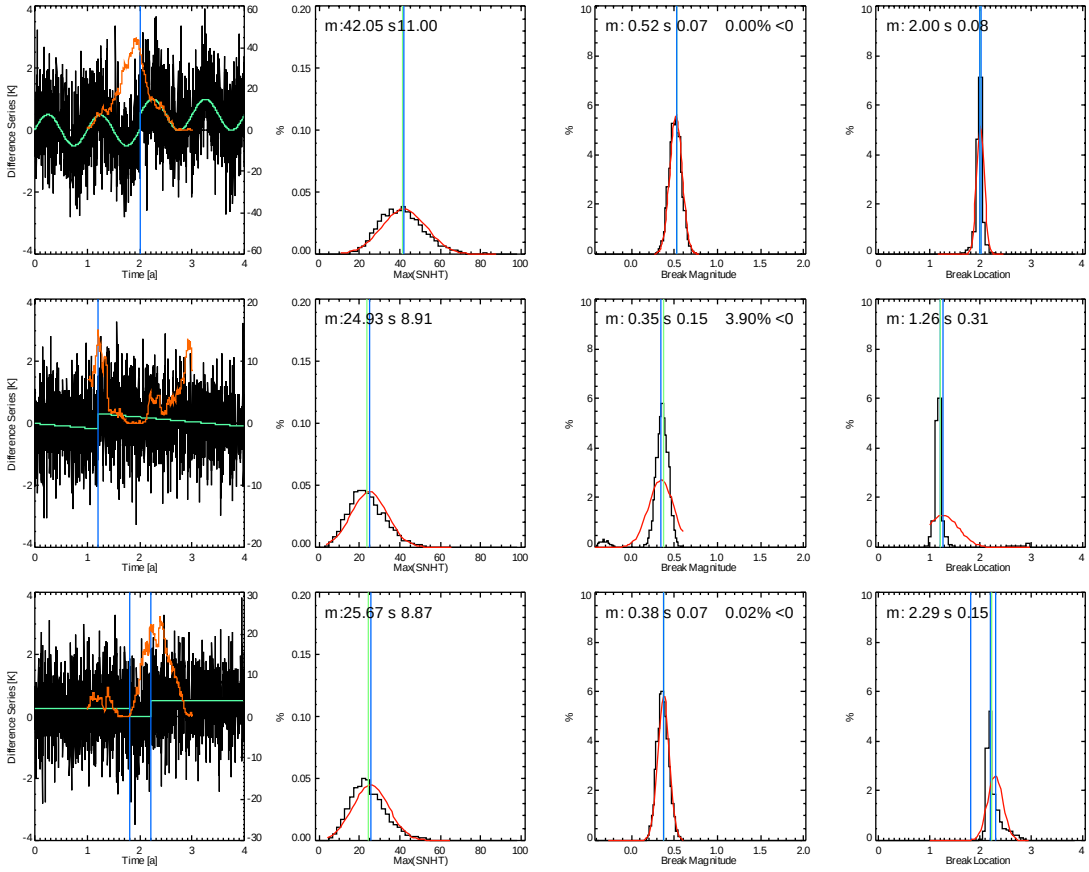

Figure 10: Some possible complications when applying the moving average SNHT to series with break of size 0.5K. Upper panels show influence of annual cycle with amplitude 0.5K. Middle panels show influence of a linear trend of 6K/decade superimposed on a break, and lower panels show case with two breakpoints separated by less than a year.

Figure 10 shows cases where the samples before/after a break are not random. The upper panel shows that the moving averaging SNHT is insensitive to an imposed annual cycle. Since the variance is increased due to the annual cycle, the SNHT becomes slightly less sensitive to breaks. If there is an annual cycle with amplitude σ, more than 99% of the T_k^s values still reach values above the 99% significance level in the simulation experiments with 5000 time series.

The middle panel shows that the test is relatively robust to a moderate imposed trend (6K/decade was assumed for this case, which is much more than expected climate trends). The break is still detected at the right place but the estimate magnitude of the break is biased (0.35K instead of 0.5K in this case). If there are multiple breaks the location and magnitude of the largest break are detected, but with biases in both break amplitude and location. In

Figure 10 above, for example, ΔT is underestimated in the middle and lower panels. The smaller break is not detected. It may be found by applying the detection/correction procedure iteratively of by using shorter analysis intervals. At present none of these measures is implemented and therefore breaks in the vicinity (+/- 2 years) of larger breaks are not detected with RAOBCORE.

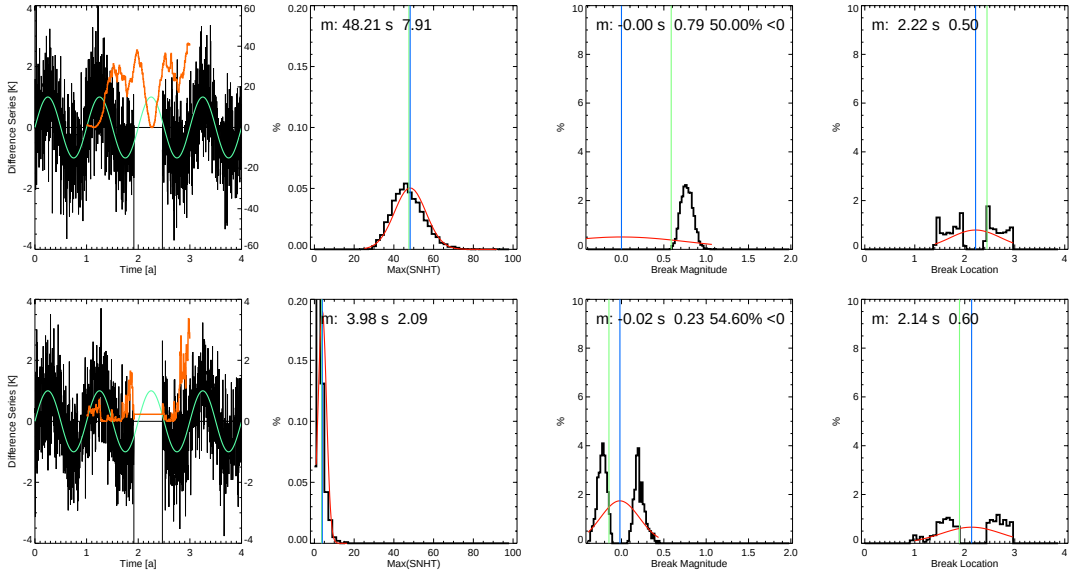

Figure 11: Influence of a data gap on break detection results using the moving average SNHT (upper panel) and the equal sampling SNHT (lower panel). Series is random with superimposed annual cycle (amplitude 1K) but without a temperature shift. Data gap has length of 200 days; analysis window of SNHT is +/- 1 year. Note large amplitude of breaks falsely detected by the moving average SNHT (third upper panel). Compare distribution of T_k^s in second left panels with uppermost second left panel of Figure 9 (homogeneous series without data gap). The significance levels change only very little despite the gap if the equal sampling SNHT is used.

Figure 11 indicates how easily data gaps can lead to false detection of breaks. It shows a random temperature series with a superimposed annual cycle but without a break. The sine wave is of comparable magnitude to the variance of the random series. In higher latitudes the sine wave may be even larger compared to the random noise (see examples in section 5.6 below). If the data gap occurs in an extreme season, the mean of one of the two samples analyzed in the SNHT is shifted compared to the other sample unless the data gap is "simulated" also in the other sample. This is done in the equal sampling SNHT. It should be noted that this problem of data gaps affects also tests based on nonparametric (robust) statistics used by LKS and Thorne et al. (2005) which also compare samples before/after a potential breakpoint.

The bg-obs time series have also been analysed for signs of autoregressive or moving average processes that could affect the significance levels for break detection. Little evidence for the existence of such signals has been found so far in the analysed time series. The equal sampling SNHT is now applied to five time series:

- 50-150 hPa log-pressure weighted layer-mean bg-obs temperature difference at 00GMT and 12GMT: This layer lies in the stratosphere at least in the extratropics and is sensitive to radiation errors. Levels above 30 hPa are not considered since the data are too sparse for reliable break detection in many regions, especially during the early years.
- 300-700 hPa log-pressure weighted layer mean bg-obs temperature difference at 00GMT and 12GMT: This tropospheric layer is less sensitive to radiation errors but is useful for correction of gross tropospheric biases or when higher levels are missing.
- 50-150 hPa log-pressure day-night temperature difference: This time series, if available, is very useful for detecting breaks independent of the ERA-40 bg. While the above time series may be affected by breaks in the ERA-40 bg, this time series is not affected. A break detected in this time series has therefore higher credibility.

At some stations different choices of the layers may be optimal and the choice may change in future versions of RAOBCORE. In principle it would be desirable to examine every pressure level of a station to be tested and to combine the break probabilities gained from the SNHT in an optimum way. It is unclear at this stage, however, how this can be done in a consistent manner since the biases at different pressure levels are not independent. Some ideas exist (e.g. Kalman filters or approaches similar to those described in Jones et al. 2001), but have not been tried out yet.

3.2 Decision algorithm for break detection

The breakpoint information from five time series has to be combined in an objective manner. Further the probability for a break depends not only on the SNHT test statistics but also on the occurrence of events such as instrument changes. Such events are connected with a higher probability for a break than times with no such changes. In order to estimate the probability for a break, we need an estimator that depends on both the a priori probabilities connected with metadata events as well as on the differences in the means. Estimators that can combine these information sources and fulfil certain optimality conditions are Bayes estimators (DeGroot, 1986). The Bayesian rule cannot be applied in a rigorous manner for the present problem since the a priori probabilities related to metadata events are not known. Nevertheless the Bayesian formula is used for calculating a score that would be a break probability if the a priori break probability were known.

Let A_1 be the event that a break with a given size ΔT occurs within a time interval of +/-3 years around a point in time and A_0 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 $\Pr(A_1 / B)$ of a break given that the event B occurs. Bayes' theorem states that this probability can be calculated as

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

The challenge is now to specify the probabilities on the rhs of this equation. A reasonable guess for the a priori probability $\Pr(A_1)$ is a uniform distribution over the time interval considered, if there are no metadata available.

The probabilities $\Pr(B / A_1)$ and $\Pr(B / A_0)$ can be found by integrating the pdfs shown in the second column of Fig. 7 from the T_k^s value found from the data to ∞. More specifically, $\Pr(B / A_0)$ can be found by integrating the pdf of the second panel of row 1 in Fig. 7. For example, let *B* now be the event that T_k^s larger than 20. If we want to detect breaks with size 0.25K one can find $\Pr(B / A_1)$ from integrating the second panel of row 2 in Figure 9.

Let us now consider T_k at every time point of the bg-obs temperature difference record in a level as an event, not only the local maxima T_k^s. This is useful since a point with value T_k which is not a local maximum of the time series of the T_k but coincides with a metadata event may reach a higher posterior probability than the point with T_k^s. Metadata information can be included in this procedure by specifying a priori probabilities $\Pr(A_1)$. If, for example the date of a radiosonde change is known, one can apply a higher a

priori probability $\Pr(A_1)$ to this particular date. For this report, the a priori probabilities listed in Table 1 have been chosen. The time series of a priori probabilities therefore contains sharp peaks at the dates of metadata events.

	No metadata	Raso. Type	Rad. Corr.	Ground
a priori Prob.	0.02	0.6	0.5	0.5

Table 1: A priori probability values used for break detection algorithm. The choices are subjective. "Raso Type" is a change of the radiosonde flown as documented in CARDS. "Rad. Corr." is a change of the radiation correction; "Ground" is a change of the ground equipment used to track the radiosonde and to process the raw observations.

After specifying the apriori probabilities the right hand side of the Bayes formula can be computed for each value T_k. Since the choice of $\Pr(A_1)$ is subjective, the results from the Bayes formula should not be called probabilities in the context of this report. Instead the output of the Bayes formula is referred to as "score" with values between 0 and 1. For this score one has to choose some threshold above which a breakpoint is detected. The value of the threshold is determined empirically.

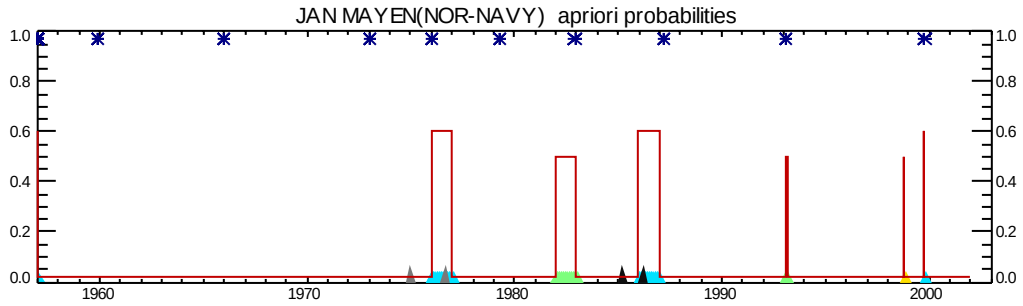

Figure 12: A priori "probabilities" specified for Jan Mayen according to CARDS metadata available.

Figures 12-14 show the breakpoint analysis for Jan Mayen. A priori probabilities are assigned in Figure 12 according to the CARDS metadata database. The values chosen assure that an adjustment is likely to be performed at metadata events. The same high value for the a priori probability is chosen over a period of a year if only the year of the change is known. The precise location of the maximum is determined from the SNHT test statistics. In practice most breaks are detected at the beginning at the end of the interval. One may argue that lower a priori probability values should be chosen if only the year of the change is known. This has not been done, however, in order to assure that high weight is given to the metadata information even if it is imprecise.

Figure 13 shows time series of the test statistic of the equal sampling SNHT. The test statistics for the bg-obs time series exceed the 99% significance level for the SNHT (~15) very frequently. Experience has shown that there are always inhomogeneities in the bg-obs difference series. The likely reason for the frequent occurrence of breaks is the contamination of the bg with small breaks due to subtle changes in the ERA-40 observing system. The time series of the obs(12GMT)-obs(00GMT) differences (lowest panel of Figure 13) tends to be much more homogeneous.

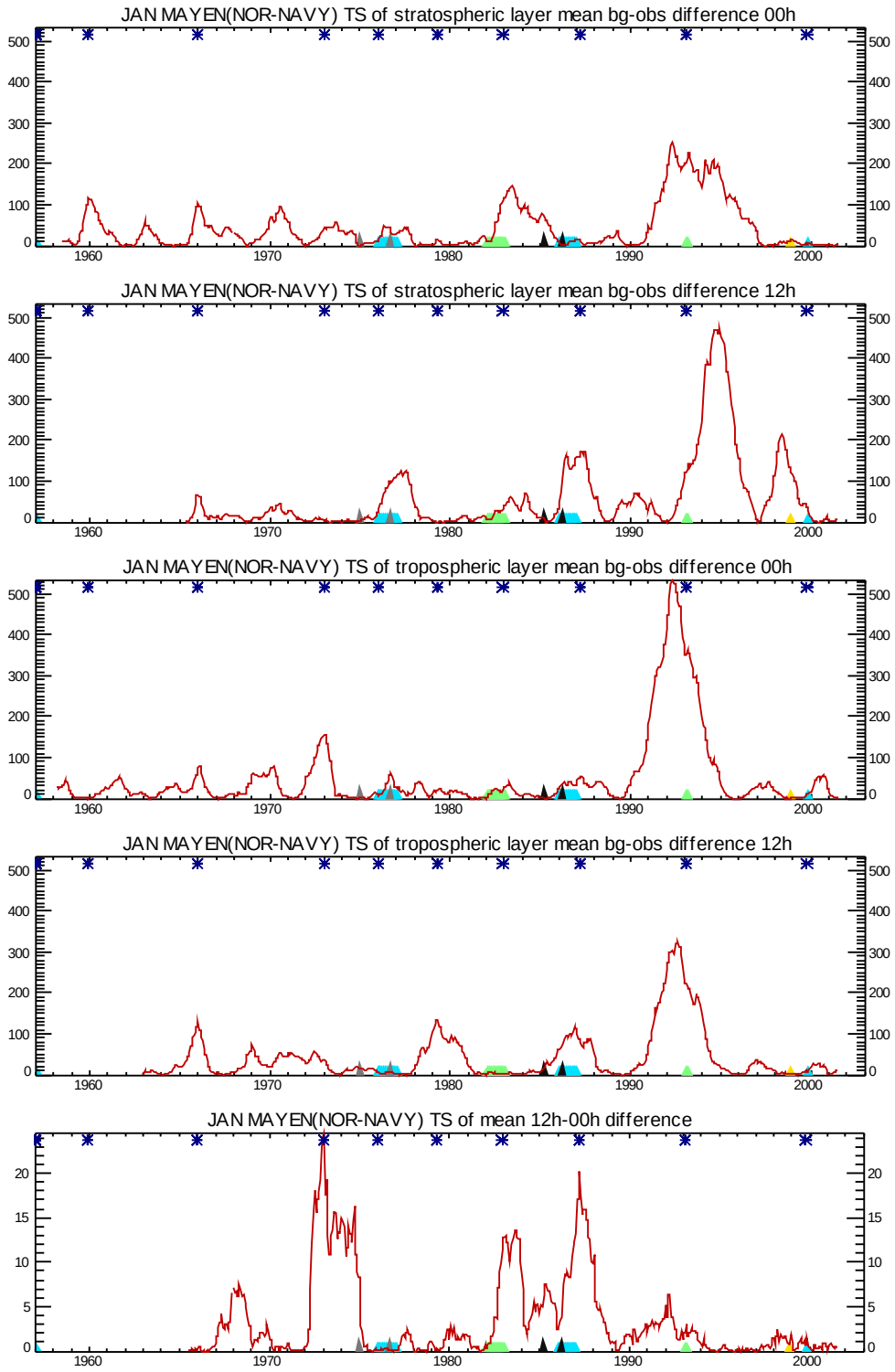

Figure 13: Equal sampling SNHT test parameter for a) 50-150 hPa bg-obs 00GMT, b) 50-150 hPa bg-obs 00GMT, c) 300-700 hPa bg-obs 00GMT, d) 300-700 hPa bg-obs 00GMT, e) 50-150 hPa obs(12GMT)-obs(00GMT). Values above 120 are considered significant for the bg-obs time series, values above 20 for the obs(12GMT)-obs(00GMT) series.

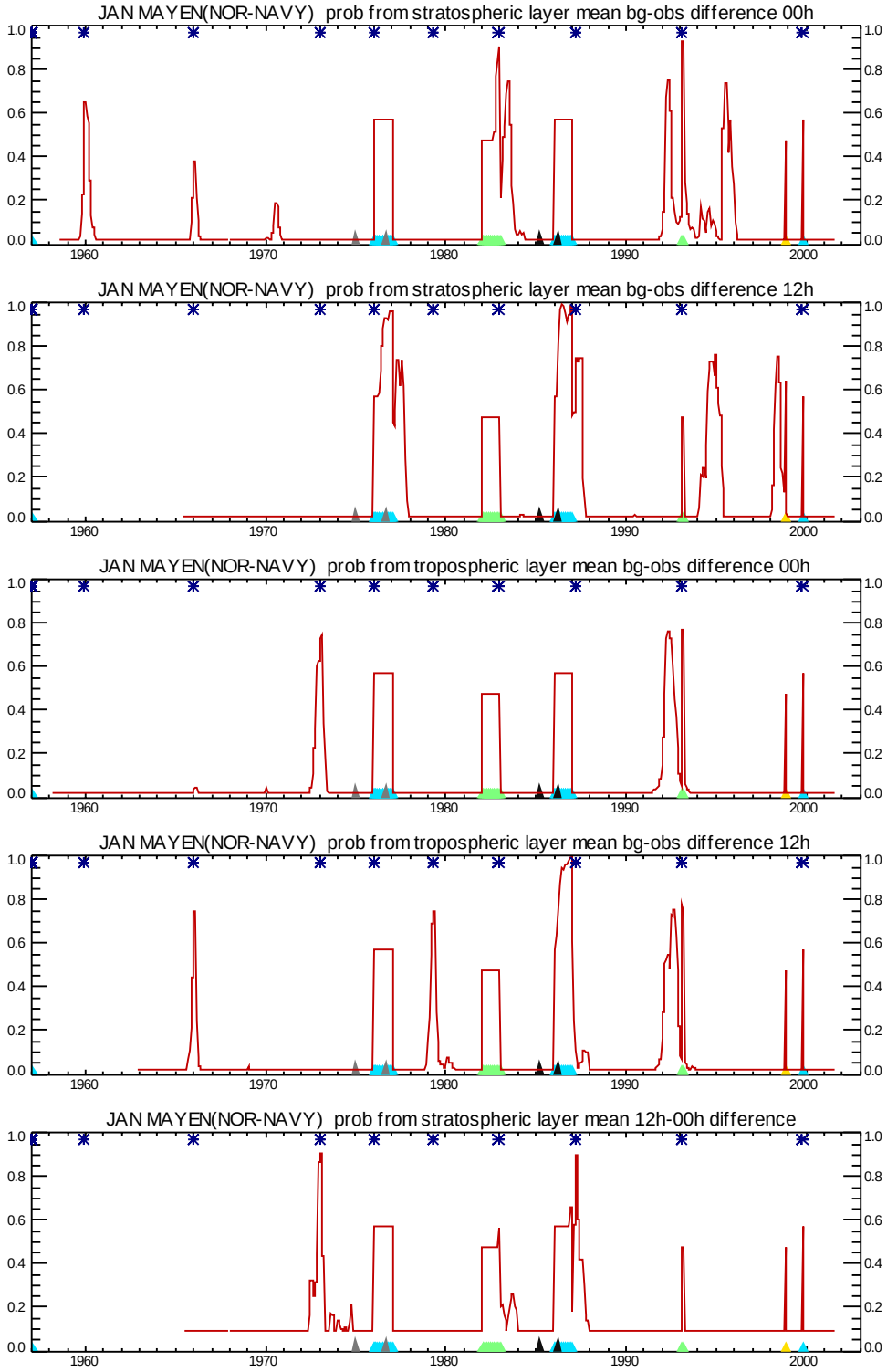

Figure 14: "Probability" score for a break given the a priori "probability" from Figure 12 and the test statistics of Figure 13. Only points with a local maximum within a +/- 2 year interval and with a minimum score of 0.5 are considered as separate breakpoints. Stars denote points where breakpoints have been detected after analyzing all five time series.

One can further see from Figure 13 that the T_k values around a large break such as that in 1992 are well above the 99% significance threshold for the SNHT, which is ca. 15, in a rather wide time interval. This reflects the fact that the SNHT only yields a probability for a break within a given *interval* (which is +/- 3 years) but not for a given point in time. The rightmost panels in Figure 9 have shown that for large breaks the break location can be determined much more accurately, however. In order to avoid wide intervals with very high break probabilities, the T_k values within +/- 2 years of the date of a peak value T_k^s are multiplied by $thresh/T_k^s$, where *thresh* is chosen empirically (120 for bg-obs time series, 20 for obs(12GMT)-obs(00GMT) series, which are much more homogeneous). This yields still a high break probability score near the date where T_k is maximum but the score drops off more quickly when moving away from the maximum. This is the reason why the peaks in Figure 14 are much sharper than those in Figure 13. The threshold of 120 for the bg-obs difference series ensures that only the largest breaks (amplitude $\geq 0.5\sigma$) are detected. T_k^s-values from bg-obs differences between 20 and 120 that would normally be considered significant are not taken into account since they are too often related to subtle changes in the satellite observing system of ERA-40.

Figure 14 shows the scores gained by combining the metadata and homogeneity test information. Due to the high a priori probability assigned to periods with metadata events, it is quite likely that breaks are detected at some date within these intervals. Besides the peaks due to metadata events there are also peaks related to maxima of the SNHT, e.g. in January 1973. At some occasions there are peaks close to but not exactly at a known metadata event (e.g. in 1987). This suggests that the dates given in some of the available CARDS metadata are not always accurate or that the break date is not estimated correctly by the statistical method, e.g. due to a second smaller break in the bg-obs time series.

Since there are 5 time series which may contain conflicting information (e.g. the location of the diagnosed breakpoint may differ slightly between the time series), one has to choose which breakpoints are actually corrected. The obs(12GMT)-obs(00GMT) score (last panel of Figure 14) is given the highest priority. The peaks whose score is above 0.5 and which are the highest peaks within a +/- 2 year interval are selected. Then the two probability time series from the tropospheric bg-obs time series are analysed. The locations with the highest peaks are chosen as breakpoints unless a breakpoint has been already detected within a +/- 2 year interval. Finally the probabilities from the stratospheric bg-obs time series (first two panels of Figure 14) are analysed. The stars in the plots above indicate the location of breakpoints which were finally chosen.

Experience with the detection algorithm was good, although it has a substantial empirical component. There are many possible ways to further improve the break detection algorithm. Perhaps the most important path for improvement is to use different time intervals (not only +/- 3 years) for the SNHT since a running mean with fixed length may be incapable of detecting all breaks. The choice of the analyzed atmospheric layers is also subjective and may be improved in future versions of the algorithm.

4. Adjustment of the breaks

After the breaks have been objectively determined, the respective time series is adjusted according to the estimated break profile from the breakpoint backwards. Therefore the adjustment of the earliest part of a time series is the sum of all adjustments at breakpoints in more recent periods.

The adjustment of the breaks is more demanding than the detection of the break and inevitably introduces errors. As can be seen from the break profiles found during the break detection at Jan Mayen in Figure 15,

the uncertainties of the amount ΔT of a break are relatively large, typically with a standard deviation of the order of 0.2K. If there are, e.g. 4 breaks in a series, the adjustment error standard deviation is then $\sqrt{\sum_4 0.2^2} = 0.4$ K. For a 40yr time series this means that the trend uncertainty due to the uncertainty in the break adjustment is on the order 0.1K/decade, depending on where the breaks actually occur (breaks at the centre of the series have most influence, breaks at the beginning/end have less influence). The time series of the adjustments as well as its errors are step functions. The adjustment errors manifest themselves as red noise in the corrected time series since the Fourier coefficients of a step function decrease with $1/k^2$ where k is the wave number. This example highlights that break adjustment is a delicate task and that only breaks should be corrected whose size is clearly larger than the errors involved in estimating it, at least at some pressure levels.

The task of break adjustment is further complicated by the fact that the breaks are largest under extreme operating conditions (typically the highest altitudes reached by the radiosondes). In these altitudes there are often missing data which make the break estimates more uncertain. In some cases it is impossible to estimate the size of the break with the data available at the station.

The default time interval used for estimating the break size is +/- 6 years. Only if there is another breakpoint before the breakpoint considered, the interval before the breakpoint is reduced accordingly but not more than to +/-2 years. The interval after the breakpoint is always 6 years (if the time series is so long) since the time series after the breakpoint is already homogenized. The minimum interval of 2 years is chosen to be consistent with the break detection algorithm, which only considers breaks that are more than 2 years apart as separate breaks. The means are calculated at each pressure level, again ensuring that the annual cycle is sampled equally before and after the break to be adjusted. The difference of the means at every pressure level yields the estimated profile of the breaks (the light blue and light red curves in Figure 15). This profile inevitably contains small-scale vertical noise. Therefore the break profile is smoothed using second order Chebychev polynomials. The coefficients are determined by a standard weighted least squares fit. A constant weighting function is used. It may be useful to use larger weights for the lowermost levels since these are usually more accurate. On the other hand they may be unrepresentative especially in mountainous terrain. In most cases the uncorrected profile is already fairly smooth. Sometimes, however, the fit with a low order Chebychev polynomial is not satisfactory. In this case the order of the smoothing polynomial (default is a quadratic polynomial in log(p)) is raised until the rms-difference between the unsmoothed and the smoothed profile is below a given threshold (0.25 K).

Figure 15 shows profiles of breaks detected and corrected at Jan Mayen. The adjustments applied are generally small for this station. At this stage all detected breaks are adjusted although it may be useful to define a threshold for the *rms* value of the smoothed break profiles to avoid adjustment of very small barely statistically significant breaks such as the one in October 1999. Other profiles such as those in 1987, 1982, 1976 and 1959 show the typical signature of changes in the radiation error, i.e. an almost linear profile increasing with height. The break profiles in 1993 and 1965 are vertically almost constant. The break in 1973 is justified by both metadata and the obs(12GMT)-obs(0GMT) time series, which is independent of the bg. The estimated break profile may have been influenced, however, by the introduction of the VTPR observing system in late 1972. Only the break in 1979 may be related to changes in the ERA-40 satellite observing system, especially since it is detected only in the bg-obs(12GMT) time series (see Figure 13).

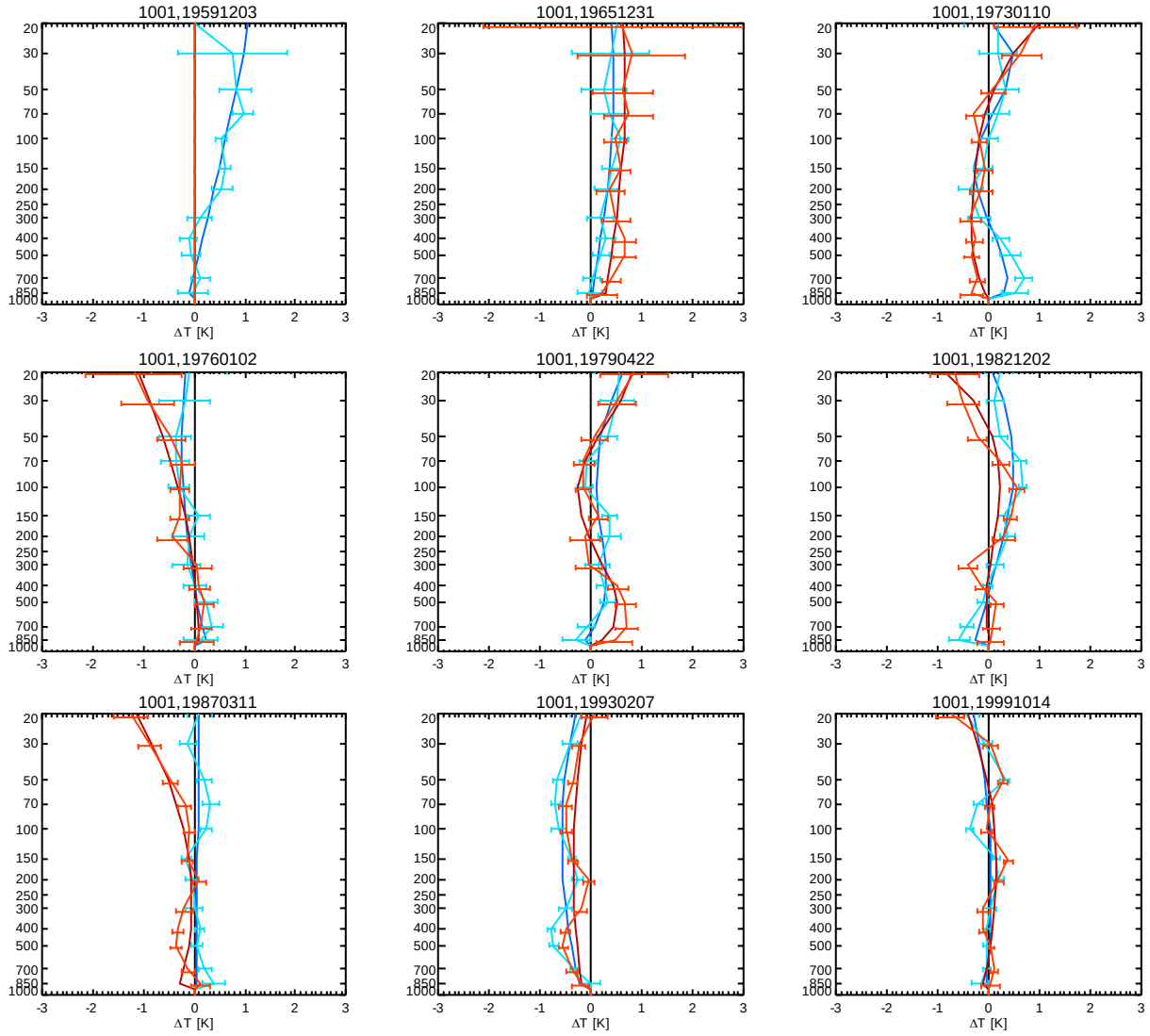

Figure 15: Unsmoothed (light blue and light red) and smoothed (dark blue and dark red) profiles of break magnitudes detected and corrected at Jan Mayen. Blue= break profiles at 00GMT, Red=break profiles at 12GMT. Titles of panels consist of station number and the date of the detected breakpoint. Error bars indicate standard deviation of sample means multiplied by 1.64 (90% percentiles). Large uncertainties at high levels are mainly due to reduced observation counts and to a lesser extent due to higher variability of the data.

The detection and correction procedure could be applied repeatedly to the bg-obs time series at a radiosonde station until no further breakpoints are found. RAOBCORE is capable of performing multiple iterations of the detection and correction step. However, only the first iteration is performed since breaks detected in later iterations are hard to justify.

4.1 Quality measures for adjustments

It is of critical importance to define quantities measuring the performance of the adjustment algorithm. We therefore define quality measures that reach minimal/maximum values if some desirable properties are fulfilled by the corrected dataset:

- One of the main reasons for undertaking homogenization efforts is the fact that trends are heavily affected by inhomogeneities. The artificial breaks often lead to very different trend estimates from nearby radiosondes which are inconsistent with the observed spatially smooth temperature autocorrelation function, especially at high altitudes (see Figure 25 below). The trends of neighbouring adjusted time series should therefore be more similar than the trends of unadjusted time series. A penalty function (referred to as *trend cost function*) can be defined for a given pressure level as weighted sum of squares of the trend differences:

$$J = \frac{1}{N(N-1)/2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(\rho_{ij} \Delta(\partial T / \partial t)_{ij} \right)^2$$

N is the number of radiosondes, *i,j* are station indices. The parameters ρ_{ij} are the weights and $\Delta(\partial T / \partial t)_{ij}$ are the difference between the temperature trends at two stations in a given time interval. Stations not far apart but with strongly differing trends contribute most to this cost function. For this report ρ_{ij} was chosen as $\rho_{ij} = \exp(-d_{ij} / 1000km)$ where d_{ij} is the spherical distance between two radiosonde stations. It is also interesting to look at the contribution of a single radiosonde station to the cost, which is defined here as:

$$J_i = \frac{1}{N-1} \sum_{j=1}^{N} \left(\rho_{ij} \Delta(\partial T / \partial t)_{ij} \right)^2$$

- Similar cost functions can be defined by replacing $\Delta(\partial T / \partial t)_{ij}$ with $(bg - obs)_i - (bg - obs)_j$ or with the difference $(obs(12GMT) - obs(00GMT))_i - (obs(12GMT) - obs(00GMT))_j$ between radiosondes *i,j* during a given time interval. These differences should also be spatially coherent and a cost function similar to the one above can be specified. Looking at the differences instead of trends has the advantage that shorter time windows (1-3 years) can be analyzed, whereas trends are typically calculated for periods of least a decade and require fairly complete time series.
- Zonal mean and global mean trends must be in reasonable agreement with independent trends estimated from, e.g. MSU layer mean temperatures.
- If the adjustments can be related to metadata events, it may be possible in some cases to compare the adjustments applied with the differences found e.g. in radiosonde intercomparison experiments.

5. Results

This section deals first with the temporal homogeneity of the ERA-40 bg forecasts. An adjustment procedure for the global mean bg time series is described. The remaining subsections describe how the homogeneity adjustments affect the radiosonde temperature trends and biases.

5.1 Analysis of the global mean difference series

The break detection/adjustment methods described below rely heavily on the difference between ERA-40 bg forecasts and radiosonde observations. It is essential for the homogenization procedure to be aware of any inhomogeneities of the ERA-40 bg since these reduce the applicability of the ERA-40 bg as a reference. These may be introduced by changes in the ERA-40 observation coverage (mainly due to satellites). Information about these changes is available from the ERA-40 blacklist files and other documentation (e.g. Hernandez et al. 2004, Kelly and Li, 2001). The satellite data certainly contributed to the realistic representation of many stratospheric features in ERA-40 analyses (Randel et al., 2003). It has turned out,

however, that the homogeneity of the ERA-40 bg forecasts crucially depends on the quality of the bias correction of the satellite radiances used in ERA-40 (Harris and Kelly, 2001).

Breaks in the ERA-40 bg can be most efficiently detected from the analysis of the global mean bg-obs difference series. These are averages of the bg-obs difference series at 900 radiosonde stations. They indicate systematic differences between the global radiosonde network and the ERA-40 bg. Breaks in individual radiosonde time series have little impact on the global bg-obs difference, whereas breaks introduced into the ERA-40 bg due to changes in the satellite observing system should show up clearly since these changes affect most regions. Figure 16 shows a history of the use of Vertical Temperature Profile Radiometer (VTPR) and (Advanced) Tiros Operational Sounder (A)TOVS satellite observing systems in ERA-40.

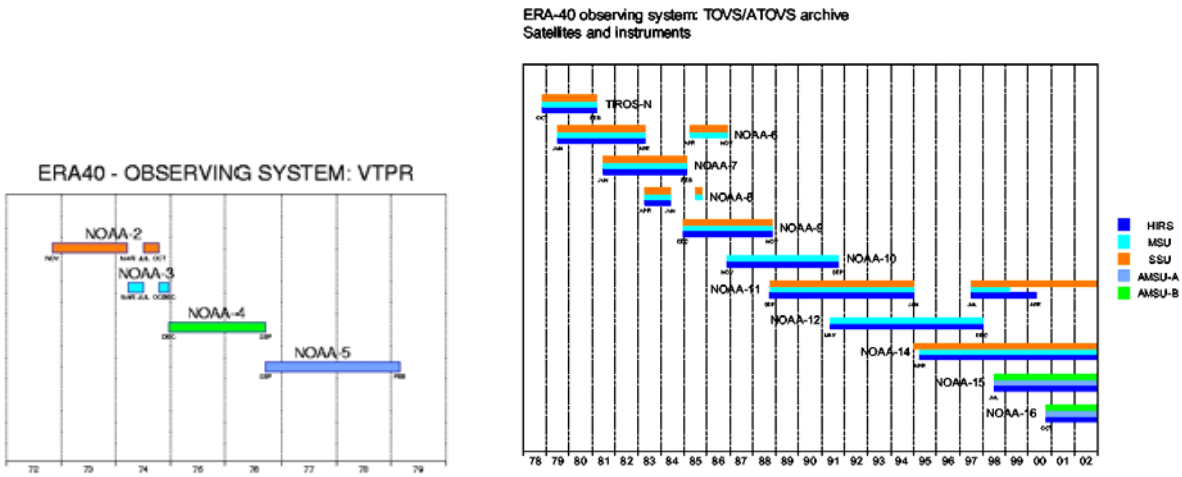

Figure 16: VTPR and TOVS/ATOVS observing system in ERA-40. See Hernandez et al. (2004) for details.

Not all changes in the observing system lead to breaks in the bg forecasts. The most prominent breaks that affect the bg down deep into the troposphere are evident in the time series of Figure 17 in January 1975, September 1976, April 1985, and April 1986. In higher altitudes additional breaks are evident. For example, in the late TOVS period in 1997 there is a significant problem in the 10 hPa level related to the use of the SSU from NOAA-11.

There are also interesting peaks of the SNHT in lower tropospheric bg-obs series in 1973, 1979 and 1997. They are related to subtle temperature changes in the ERA-40 bg probably due to the introduction of the VTPR in late 1972, the launch of NOAA-6 and the use of NOAA 11.

It is also important to note that the bg-obs difference series has a strong positive trend in more recent periods. It indicates that the time series from radiosonde measurements show much stronger stratospheric as well as tropospheric cooling from 1986 onwards than the bg. This large trend difference must be understood in detail. It is mostly related to excessive latent heating in the tropics due to problems with the analysis of satellite humidity information in the 1990s (see Uppala et al., 2005 for details). However, artificial breaks in radiosondes or by other changes in the observing system influencing the ERA-40 bg may have contributed to this discrepancy between radiosonde observations and the bg as well.

The breaks and spurious trends in the global mean bg have negative effects on the performance of RAOBCORE, which assumes a temporally homogeneous bg. They may trigger the false detection of breaks and the estimates of the size of the breaks may be biased.

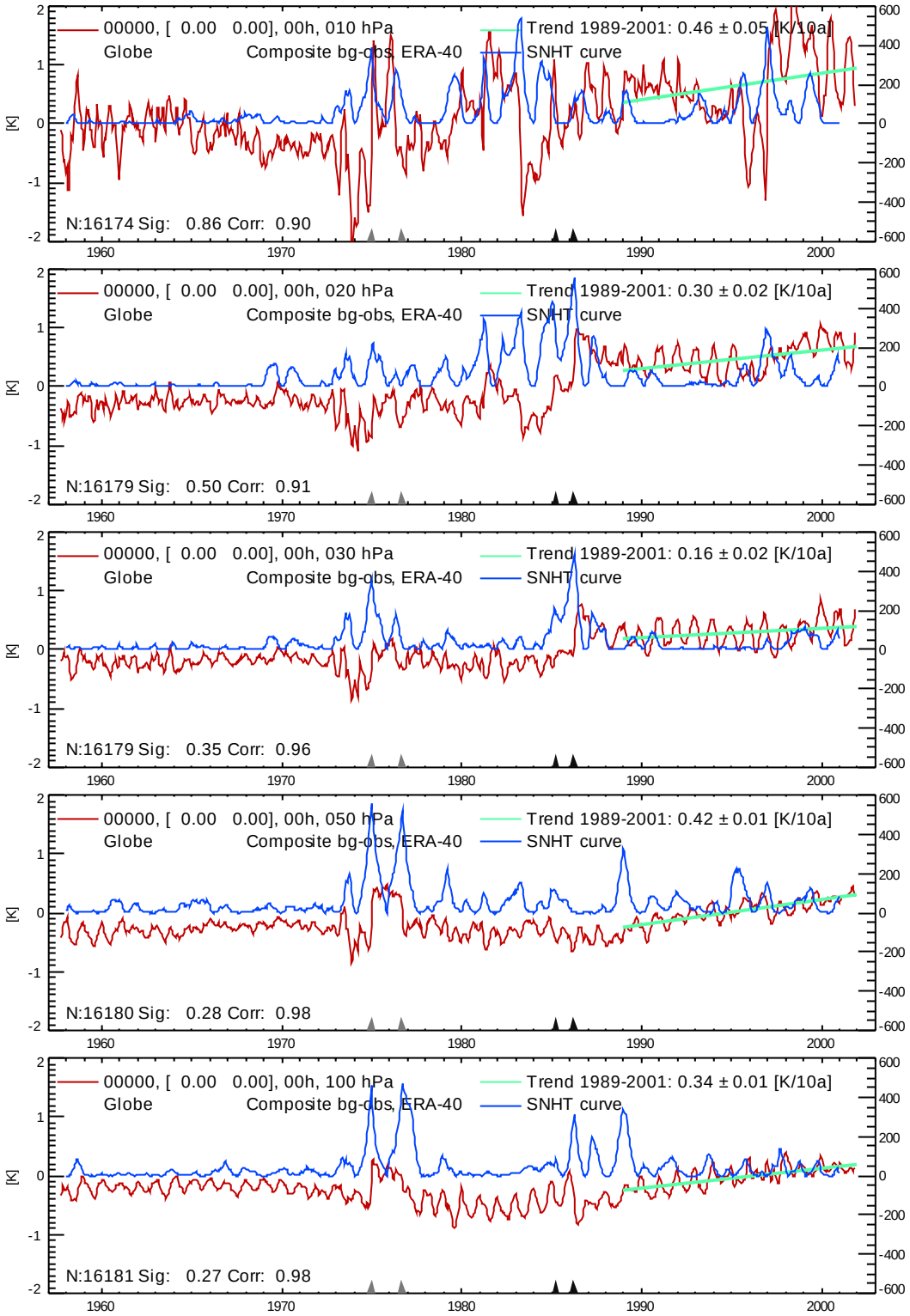

Figure 17: bg-obs at 10 pressure levels averaged over the 900 radiosonde stations. Indian stations and Antarctic stations are excluded. Red curves are 50 day running averages of difference, blue curves are 1-yr SNHT test statistics. A shorter time interval has been used here for the SNHT to get a better temporal resolution of the breaks.

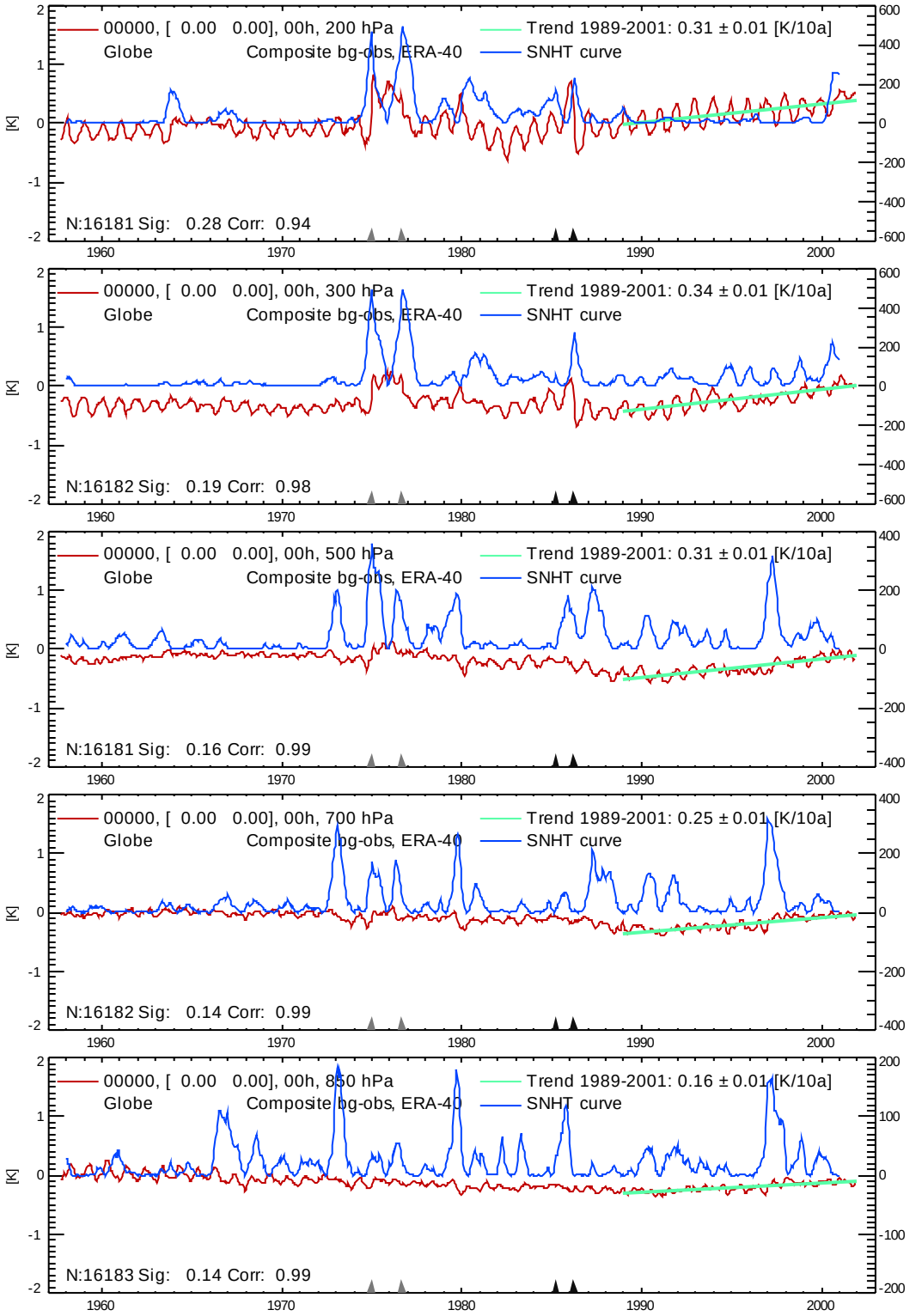

Figure 17 contd.

In order to reduce the effect of the breaks in the global mean bg, the bg time series are adjusted by the *global mean* bg-obs differences. In general the influence of the satellites is smaller in regions with dense radiosonde observation coverage. Therefore the adjustment Δbg is scaled with the radiosonde density in the following way:

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

where $\overline{bg-obs}(p,t)$ is the global mean bg-obs difference and the weighting function *w* is defined as:

$$w(\lambda,\varphi,t) = 1.2 - \rho(\lambda,\varphi,t)/\rho_{\max}(t)\,.$$

The radiosonde observation density is defined as:

$$\rho(\lambda,\varphi,t) = \sum_{i}^{N(t)} \exp(-d_i(\lambda,\varphi)/700km)\,f_i$$

where d_i is the spherical distance between location (λ,φ) and radiosonde *i*. The factor f_i takes into account whether the radiosonde reports once ($f_i = 1$) or twice daily ($f_i = 5$). $N(t)$ is the number of active radiosondes and $\rho_{\max}(t)$ is the maximum radiosonde density found at a particular time.

Finally the weights $w(\lambda,\varphi,t)$ are adjusted by a constant factor such that the global mean Δbg is exactly equal to $\overline{bg-obs}(p,t)$ when averaged over all radiosonde stations. Figure 18 shows the spatial pattern of the weight $w(\lambda,\varphi,t)$. The time variation of the weight is quite small so that this figure can be considered representative for the whole ERA-40 period.

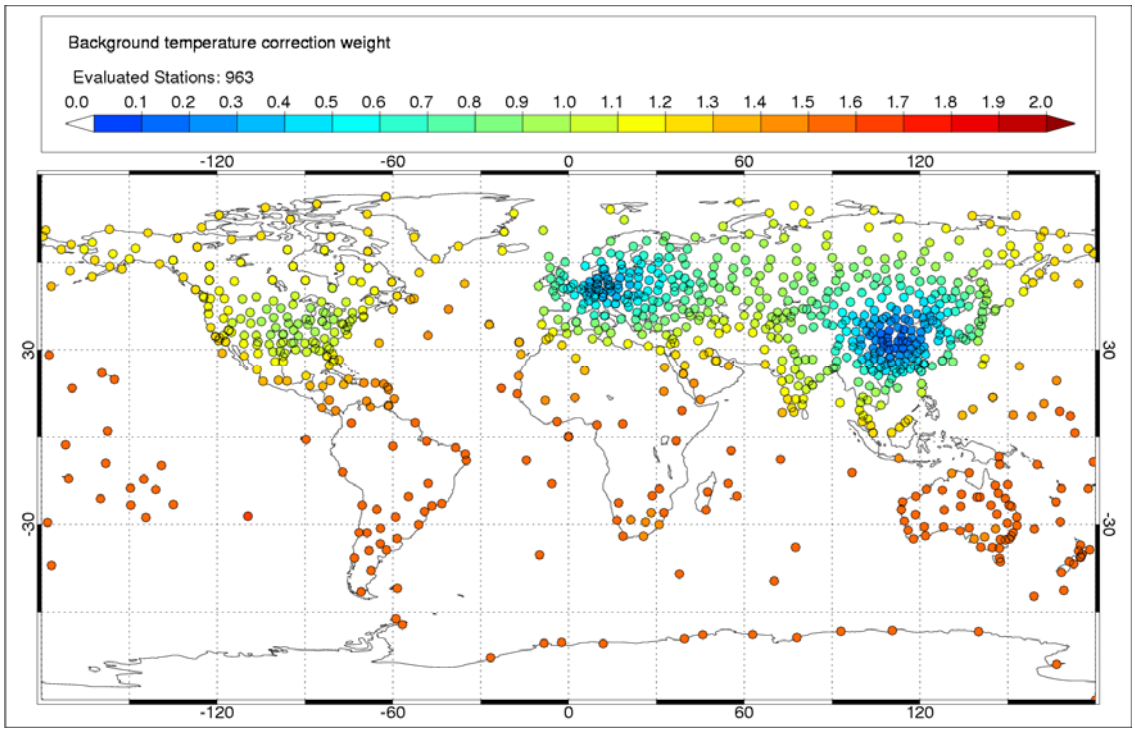

Figure 18: Weights applied to the global mean bg-obs which is subtracted from the ERA-40 bg. It is assumed that the breaks due to changes in the satellite observing system have least influence in regions with high radiosonde density and most influence in regions with sparse radiosonde data coverage. Maximum weight is limited to ca. 1.6.

One can see that the largest adjustments are made in regions where the radiosonde network is sparse. The maximum adjustment is about 1.6 times the global mean adjustment, the minimum adjustment is below 0.2 times the global mean adjustment. Figure 19 shows the bg-obs series averaged over the radiosondes south of 25N before and after adjustment with the weighted global mean bg. The large breaks due to erroneous bias correction of the NOAA-4 radiances between Jan 1975 and Sept. 1976 (see Figure 17) are almost entirely removed in the average over this data sparse region, at least at the 50 hPa level. The spurious trend of the bg in the 1990s, which is likely caused by increasingly excessive precipitation in the tropics during this period, is also removed. The bg-obs difference in the early years is slightly reduced.

The transition of the mean bg-obs difference south of 25N between 1986 and 1990 from negative values to values close to zero, which is still evident after the adjustment of the bg, is related to the introduction of the Vaisala RS80 radiosondes in Australia and many Pacific islands.

The adjustment of the bg developed here certainly cannot remove all biases in the bg which are the result of complex interaction between biases of the observing system and of the assimilating model. In future reanalyses, influence statistics (Cardinali et al. 2004) will be available that will help to design better adjustment procedures. Further it can be expected that future reanalyses will have less breaks due to changes in the satellite observing system since progress has been made recently in implementing automatic adaptive bias correction procedures (Dee, 2004). Similar procedures to cure the effect of systematic model errors are also under development.

As is shown in one of the sensitivity experiments in section 5.7 below, the effect of the bg adjustment is quite positive: The number of detected breaks is reduced, the spatiotemporal consistency of the adjusted trends is improved and the global mean adjustments are more realistic if the bg adjustment is applied.

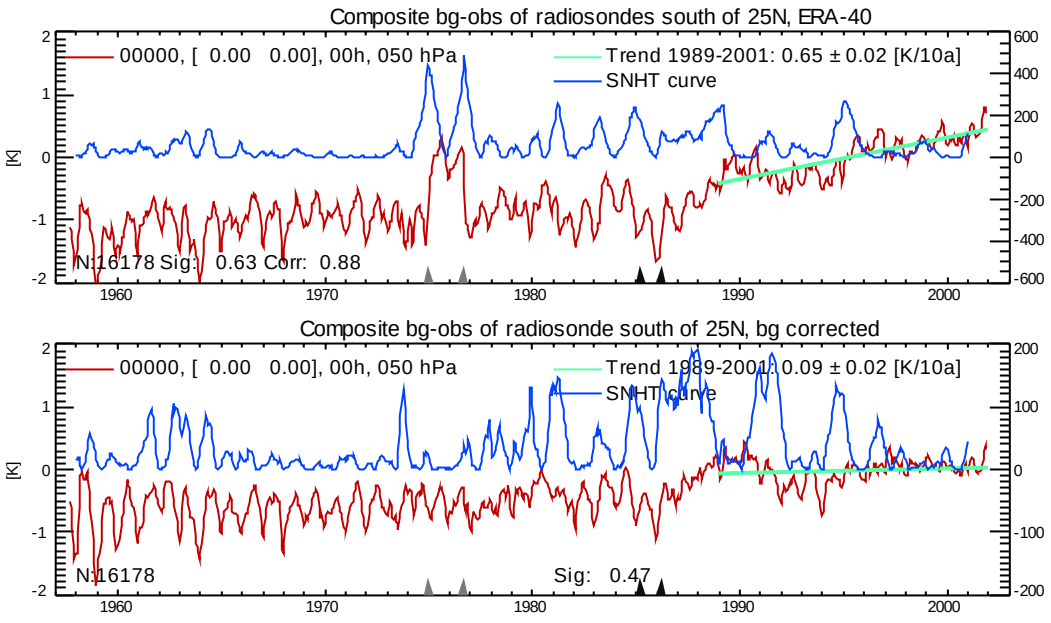

Figure 19: Average of bg-obs series for radiosondes south of 25N (India and Antarctica excluded) in 50 hPa. Upper panel: bg unadjusted, lower panel: bg adjusted by weighted global mean bg-obs. Adjustment removes important spurious features such as the NOAA-4 problem and the trend after 1989.

5.2 Adjustment results for Jan Mayen

After the global mean bg time series has been adjusted, one can adjust the individual radiosondes. In this section the analysis of radiosonde station Jan Mayen is continued. Figure 20 shows the adjustment of the ERA-40 bg that is added to the bg-obs time series shown in Figure 4. One sees the signature of the NOAA-4 problem and also the negative trend of the bg adjustment in the 1990s. Then the break detection and adjustment as described above are performed. The breaks detected for Jan Mayen have been shown already in Figure 15.

Figure 21 shows the adjusted bg-obs difference time series and the adjusted obs(12GMT)-obs(00GMT) difference time series. Both curves show a better temporal homogeneity compared to the unadjusted difference series shown in Figure 4 and Figure 7. It should be noted that the radiosonde station Jan Mayen already had a relatively homogeneous record compared to most other radiosonde stations. Therefore the effect of the adjustment is relatively small.

While the adjustments applied seem justifiable, this example alone does not tell much about the reliability of the detection and adjustment method. In the following subsections the overall performance of the adjustment algorithm is examined.

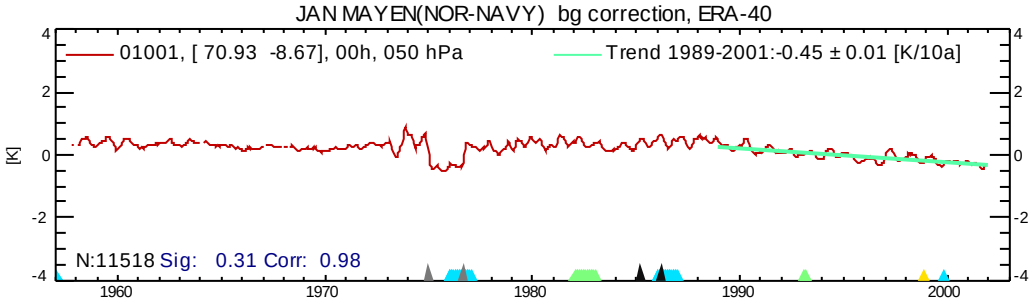

Figure 20: Adjustments of the ERA-40 bg series added to the bg-obs series of Jan Mayen at 50 hPa before applying RAOBCORE.

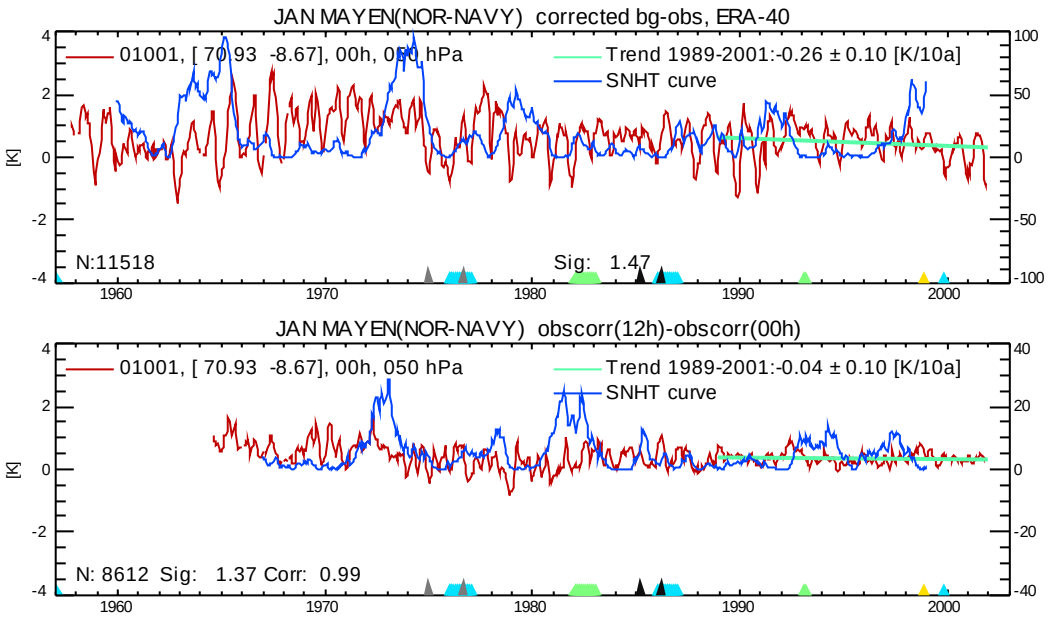

Figure 21: bg-obs time series and obs(12GMT)-obs(00GMT) at Jan Mayen at 50 hPa after adjustment of both the bg and obs series. Compare with Figure 4 and Figure 7.

5.3 Global maps of trends from adjusted radiosonde time series

This section documents how RAOBCORE changes the temperature trends at individual radiosonde stations. Before turning to the trends it is interesting to see the frequency of break detections with RAOBCORE. Figure 22 shows the monthly number of breaks detected at any radiosonde station of the global radiosonde network. The upper panel contains all detected breaks. An overall number of 4594 breaks have been detected. The second and third panels show that about 30% of the breaks coincide with changes of the radiosonde type or with radiation correction changes documented in CARDS. This percentage may seem low, but the digitized CARDS metadata database of Aguilar (2001) contains little information on breaks in the mid-1990s and very little information on changes in the Chinese radiosonde network. The peak in January 1982 is caused by many documented changes in the radiation correction. The same is true for the peak in 1993, when the radiation correction for Vaisala RS80 radiosondes has been changed at many places. If no metadata are used (see section 5.8 below) these peaks disappear.

Figure 23 and Figure 24 show one of the main results of this report. They document that due to the adjustments calculated by RAOBCORE the spatial heterogeneity *J* (defined in section 4.1) of the radiosonde temperature trends is substantially reduced. It is reduced not quite but almost to the level of spatial heterogeneity of the ERA-40 bg trends (lower panel of Figure 23). The RAOBCORE adjustment procedure is able to adjust the different trends over the US mainland and Alaska that are caused by the use of different radiosonde types. The trend heterogeneity over other regions is also reduced. 1989-2001 is a very short period for calculating trends. The beginning has been chosen because 1989 will likely be the starting date of a planned interim reanalysis at ECMWF and this report is intends to give an impression what improvements can be expected by application of RAOBCORE. Further the main aim here is to show the negative impact of breaks on the trends and not so much to interpret the found trends in the context of climate change. Figure 25 shows that RAOBCORE performs significantly better in terms of trend heterogeneity than the radiosonde bias correction used in ERA-40, even if an improved version that permits adjustment of the daily mean bias and that does not have the problem of a one year time shift.

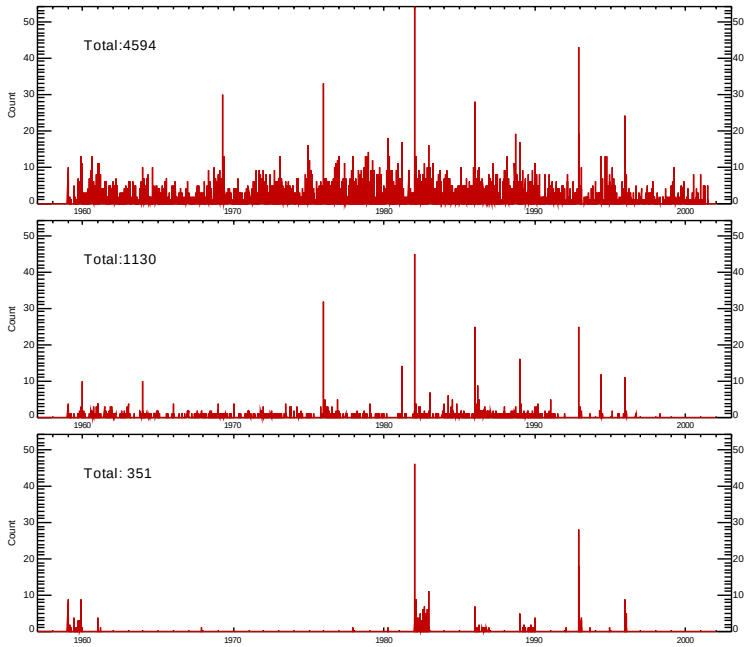

Figure 22: Number of detected breaks. Upper panel contains all detected breaks, second panel shows number of breaks coinciding with CARDS- documented radiosonde changes, third panel shows break associated with radiation corrections.

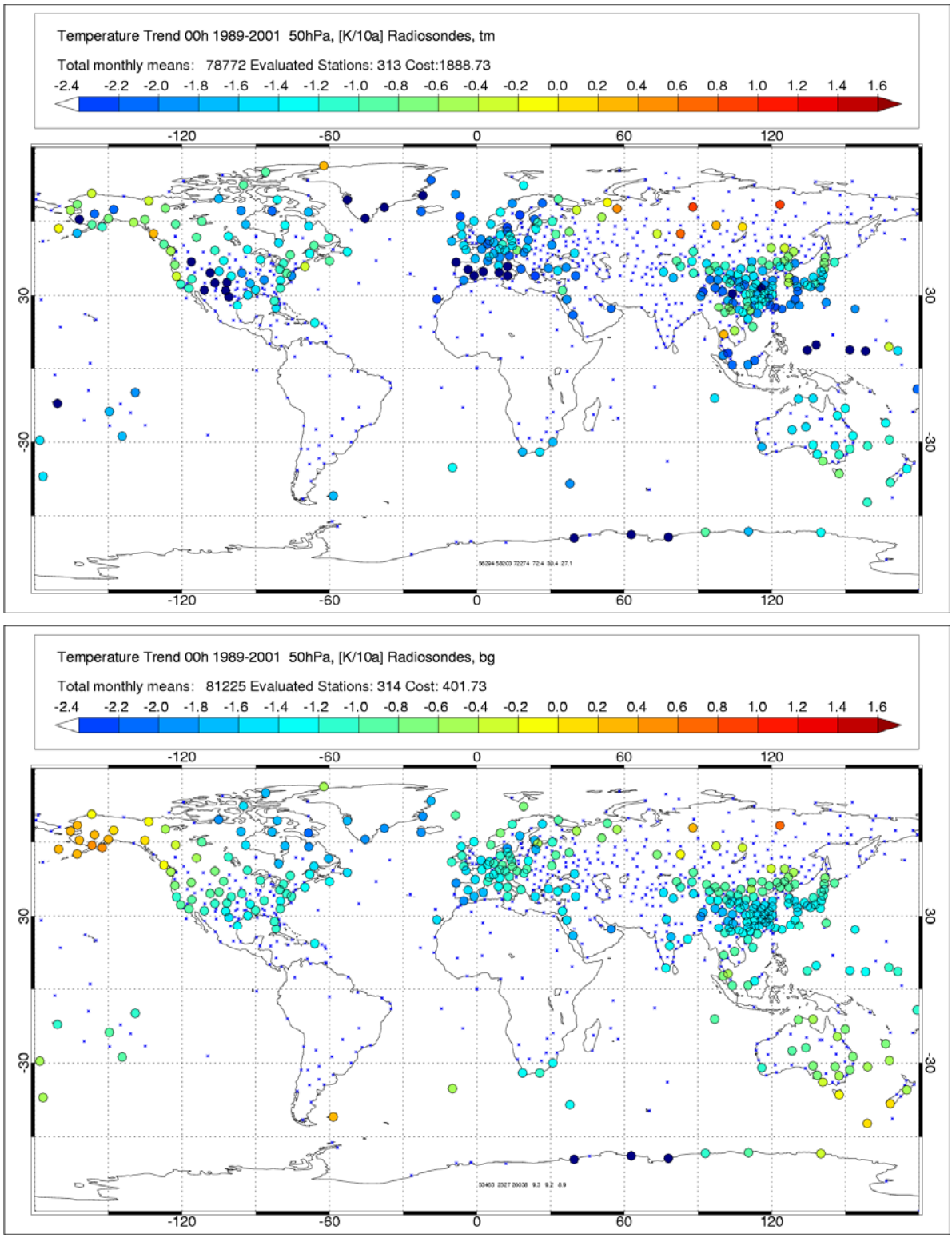

Figure 23: Temperature trends 1989-2001 at 50 hPa at 00GMT, evaluated from uncorrected radiosonde temperature time series and from the bg time series at the location of the radiosonde stations. The first number in the panel header is the total number of monthly means available for plotting this figure, including the means of stations that have not been plotted as bullets. The second number specifies how many stations had enough data to calculate reliable trends for the investigated period. The third number is the value of the trend cost function at the given pressure level, as defined in section 4.1.

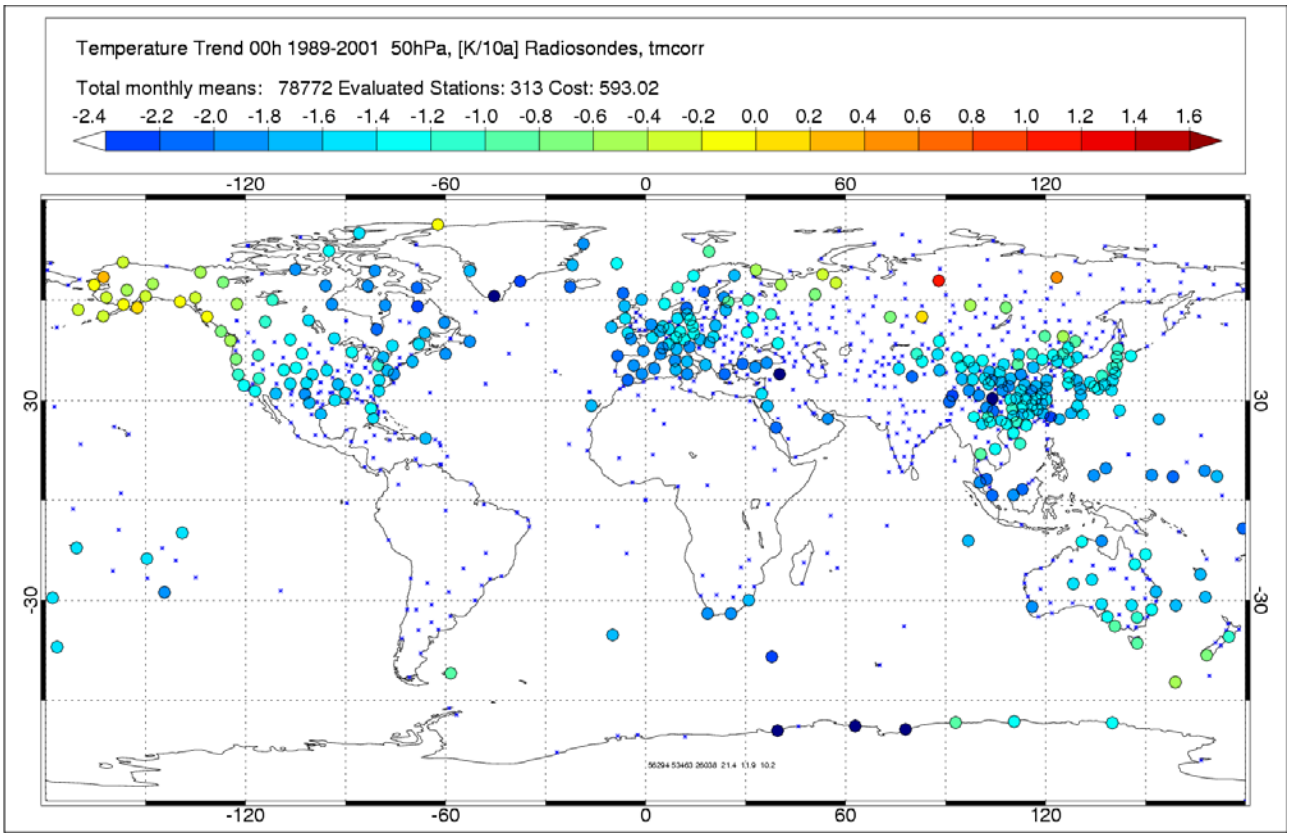

Figure 24: Radiosonde temperature trends at 50 hPa, 00GMT for the period 1989-2001 after adjustment with RAOBCORE. Note considerably better spatial consistency of trend estimates compared to the uncorrected temperatures in the upper panel of Figure 23. Note stronger cooling trend compared to bg trends in the lower panel of Figure 23 over Australia and Alaska.

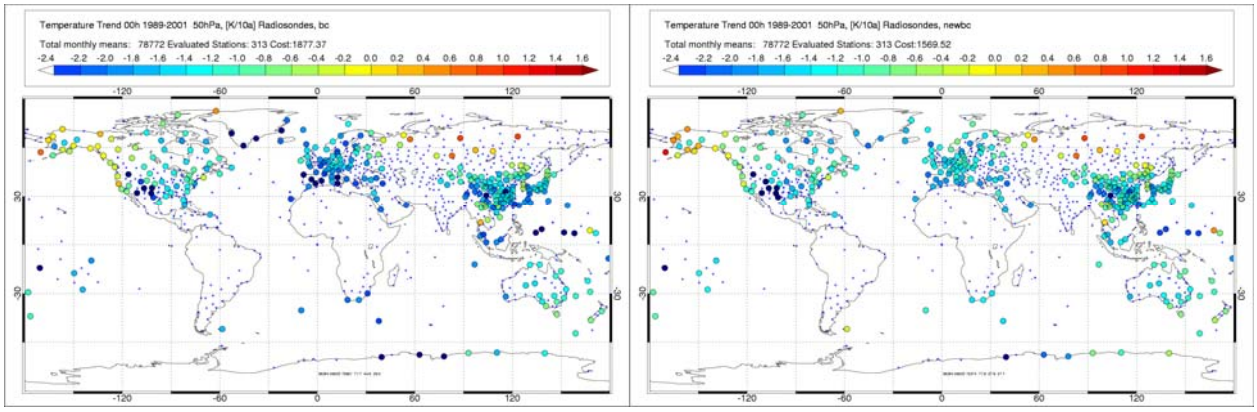

Figure 25: Maps of temperature trends at 50 hPa for the period 1989-2001. Left panel shows radiosonde trends corrected with the method described by Andrae et al. (2004), Right panel shows trends corrected with an improved version of the Andrae et al. (2004) algorithm.

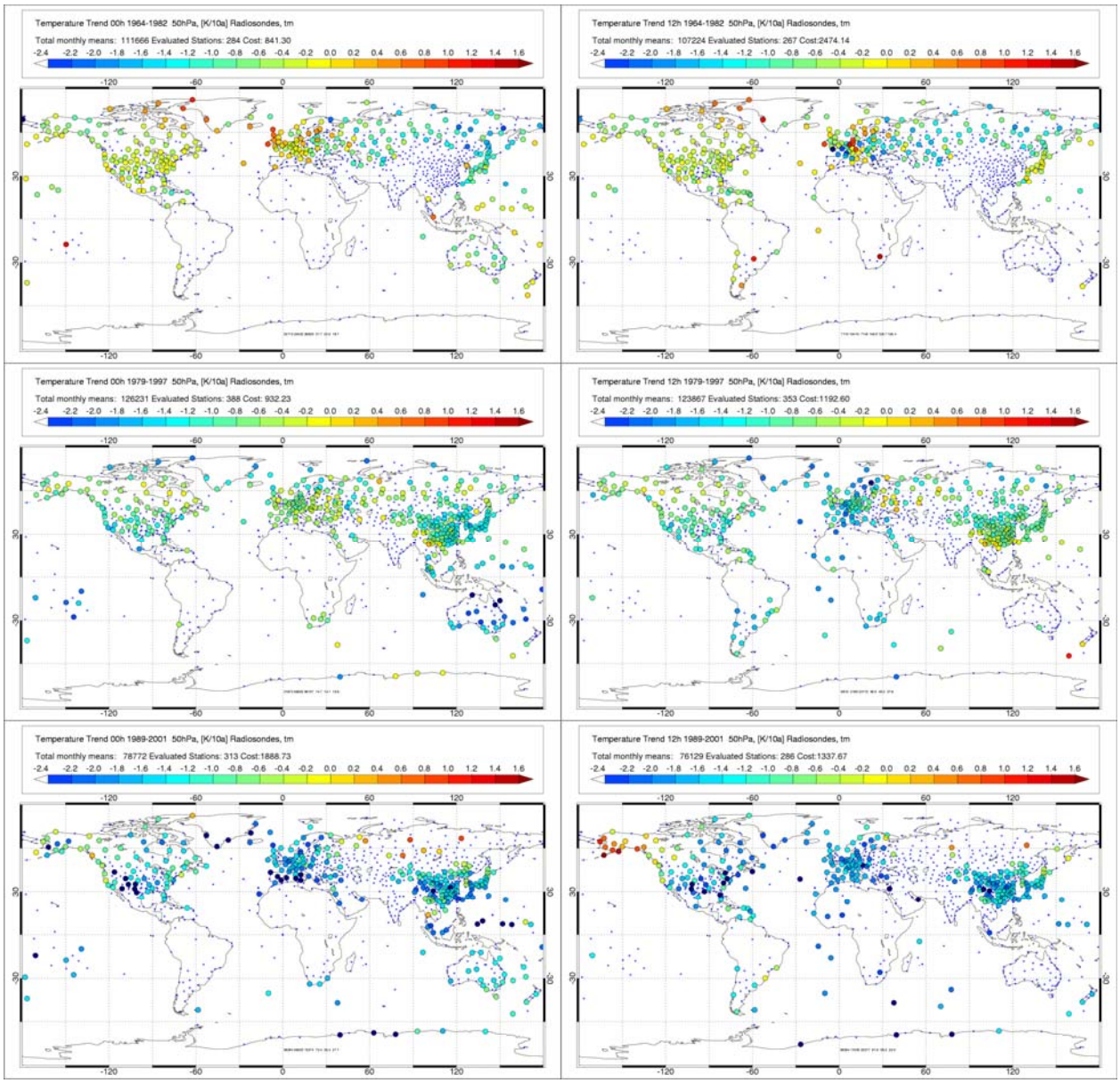

Figure 26: Trends of uncorrected temperature time series for periods 1964-1982 (upper panels), 1979-1997 (middle panels) and 1989-2001 (lower panels) at the 50 hPa level for 00GMT ascents (left panels) and 12GMT ascents (right panels).

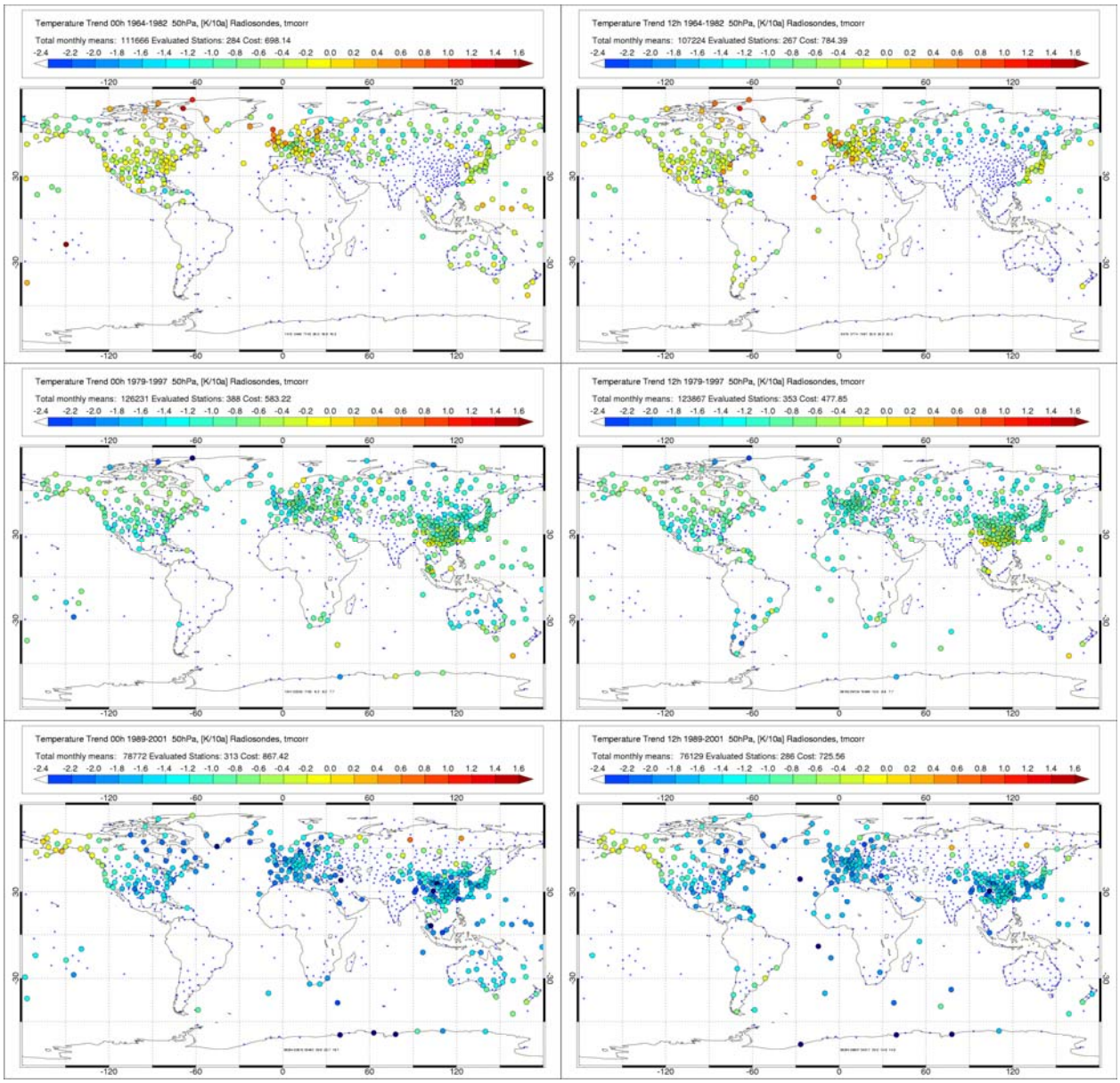

Figure 27: As Figure 26 but with corrected trends. Note marked differences compared to Figure 26 over Europe in the period 1964-1982, over Australia in the period 1979-1997 and over the US and China in the period 1989-2001.

These results are not restricted to the most recent period, as can be seen from Figure 26 and Figure 27 where trends from unadjusted and adjusted time series at the 50 hPa level from three different time periods are compared. For example the extremely different trends over Europe during the period 1964-1982 are much less heterogeneous after the adjustment with RAOBCORE. However one also gets the impression that there is still some room for improvement of the trend heterogeneity for this period.

Figure 28 is an example how the trend cost J_i for an individual radiosonde, defined in section 4.1, can be used to spot problematic radiosonde stations. It can be seen that the uncorrected time series over China and over the US, where different radiosonde types have been in use in the 1990s, contribute most to the trend cost in the interval 1989-2001. The trend cost diagnostic works best in regions with good radiosonde coverage. It is almost insensitive to strongly deviant trends at remote island stations, however.

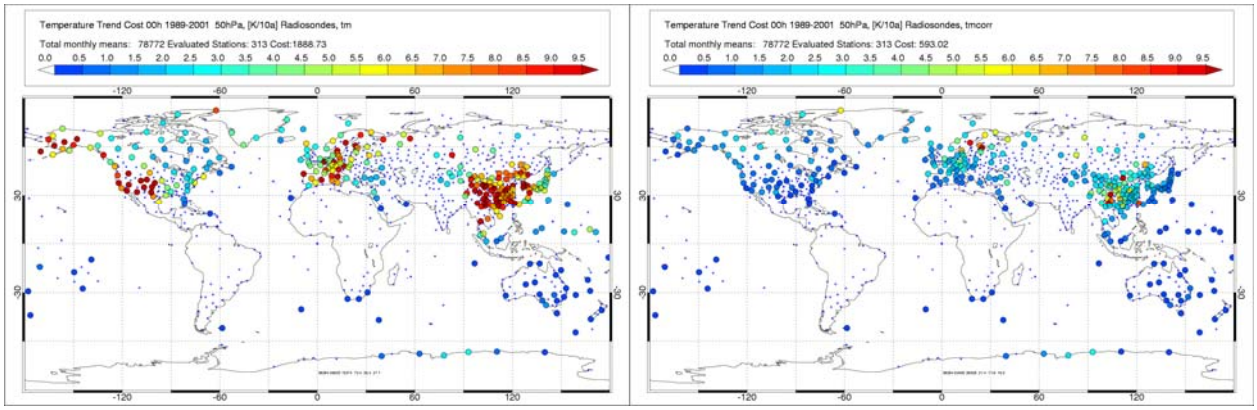

Figure 28: Trend cost contribution J_i calculated for each radiosonde station at the 50 hPa level in the period 1989-2001, before the adjustment with RAOBCORE (left panel) and after the adjustment (right panel). This diagnostic has been valuable for detecting stations where the adjustment algorithm has problems.

Figure 29 shows that RAOBCORE reduces the trend heterogeneity substantially not only at the 50 hPa level. The trend heterogeneity using RAOBCORE-adjusted time series is larger than the bg trend heterogeneity, which is not surprising since much of the small scale noise in the radiosonde observations is filtered in the data assimilation process. However, RAOBCORE consistently outperforms the adjustment algorithms used in ERA-40 in this respect.

RAOBCORE needs to improve not only the spatiotemporal consistency of the radiosonde temperature time series but also the estimates of the global mean trends. It has been argued that the many breaks in the radiosonde time series, particularly due to the reduction of the radiation error with time, may have introduced a spurious signal into the radiosonde trends. Seidel et al (2004) made a comprehensive intercomparison of uncorrected and corrected radiosonde datasets. It was found that the unadjusted radiosonde time series show stronger stratospheric cooling trends than the adjusted radiosonde time series from LKS, Parker et al. (1997), Thorne et al. (2005).

These results are corroborated in the present study. compares zonal mean trends for the period 1989-2001. The zonal means have been calculated only from time series that are complete enough to allow a trend estimate at the individual stations (i.e. 11 of 12 years of data were required). Trends from 00GMT and 12GMT have been combined for this plot. The trends of the individual stations have been binned into 10x10 degree grid boxes. In the grid boxes, cos(lat) weighted averages of the trends have been calculated. These averages have been zonally averaged. The binning procedure was used to alleviate the effect of the uneven spatial distribution of the radiosondes. In some latitude belts very few radiosondes are available. A minimum of 3 radiosonde trends per 10 degree latitude belt has been required. Regions with fewer radiosonde trends such as the inner-tropical stratosphere have been left white.

Figure 30 shows zonal mean trends for the unadjusted radiosonde time series and for different versions of adjusted time series, as well as zonal mean trends of the bg and an time series in the period 1989-2001. The uncorrected series (upper left panel) show a strong cooling trend in the stratosphere. There is cooling of the troposphere in the South Polar Region, little change in the tropics and heating in the northern hemisphere extratropics. The correction with RAOBCORE (upper right panel) changes these zonal means very little. There is a slight reduction of the stratospheric cooling during this time interval but little change elsewhere.

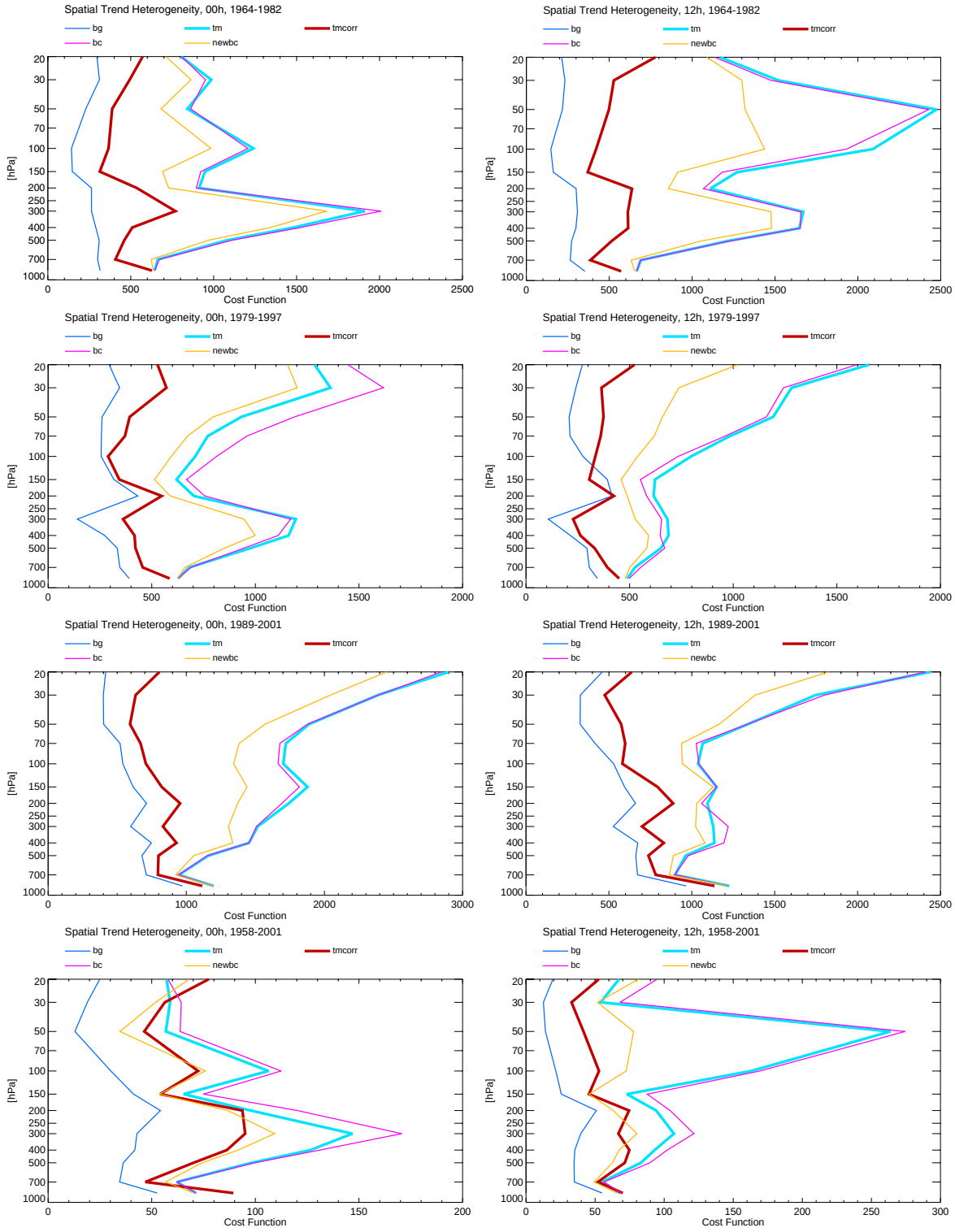

Figure 29: Vertical profiles of the cost function J calculated from series of the ERA-40 background (bg), of the original radiosonde series (tm), of the series corrected with RAOBCORE (tmcorr). J from series corrected with the ERA-40 bias correction algorithm (bc), and with new version of Andrae et al's bias correction (newbc) are also shown for reference. Profiles are shown for 00GMT and 12GMT ascents during periods 1964-1982, 1979-1997, 1989-2001 and 1958-2001.

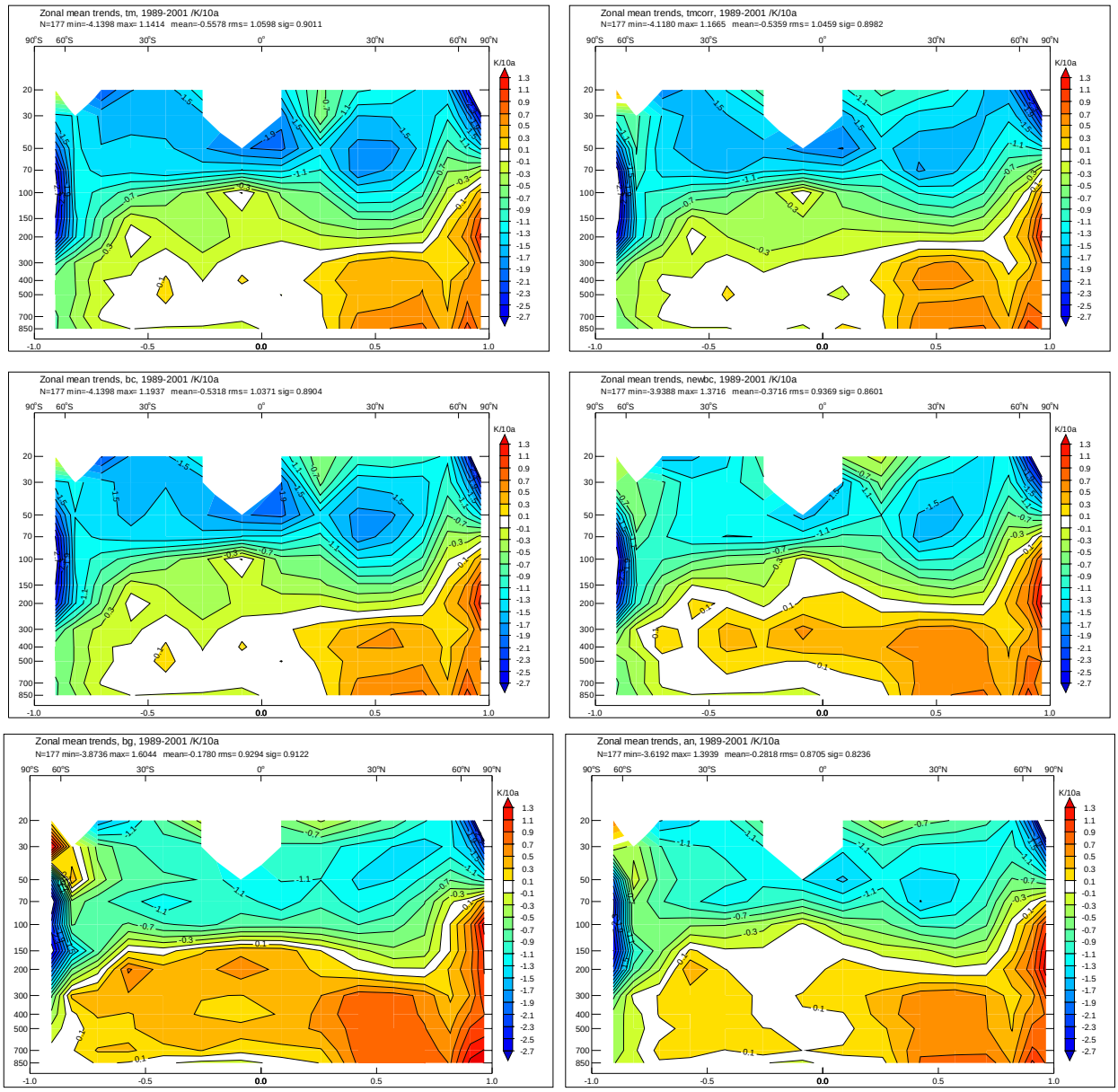

Figure 30: Zonal mean trends for the period 1989-2001. Trends in upper left panel are calculated from unadjusted time series. Upper right panel: trends from time series adjusted with RAOBCOARE. Middle left panel: trends adjusted with method of ASO, which was used in ERA-40 from 1980 onwards. Middle right panel: more aggressive version of the adjustment method of ASO. Lower panel: Trends from ERA-40 background forecast (bg) time series and ERA-40 analysis (an) time series.

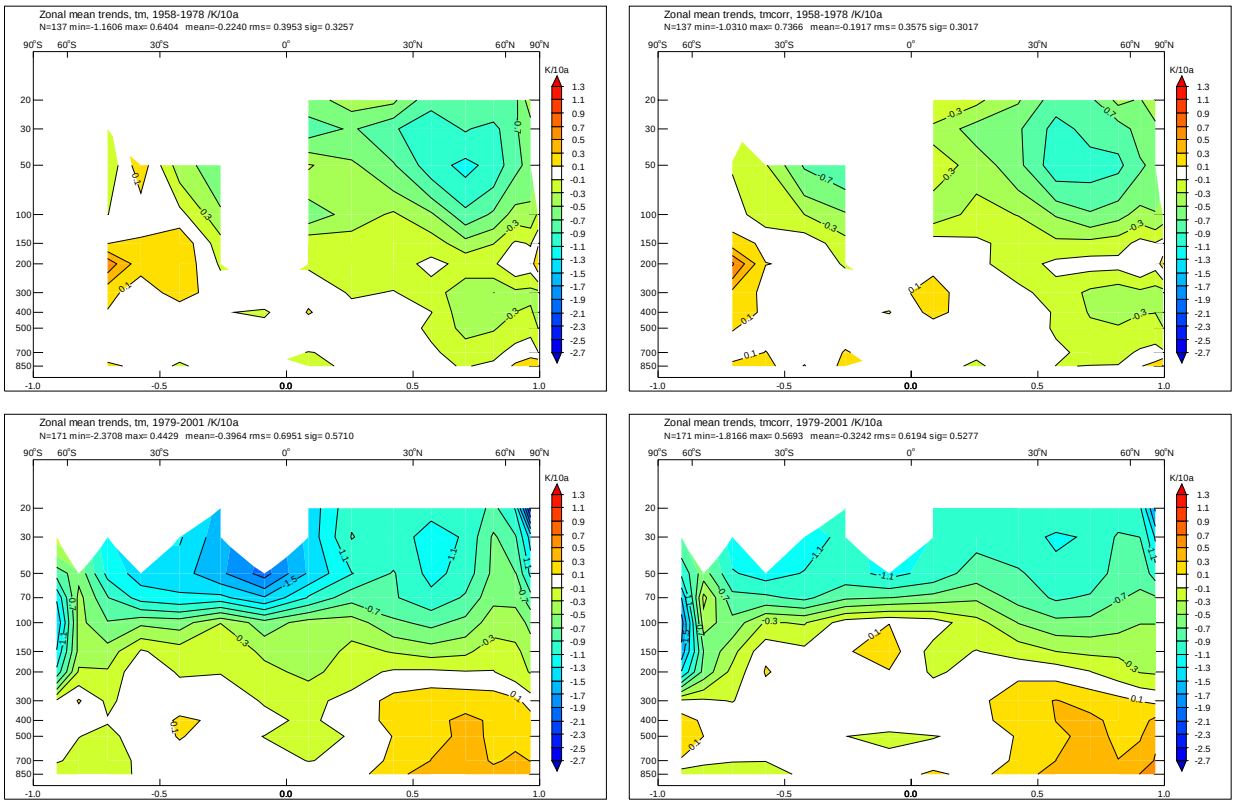

Figure 31: Zonal mean trends from unadjusted (left) and adjusted (right) radiosonde temperature time series for periods 1958-1978 and 1979-2001. These periods have been used here instead of 1964-1982 and 1979-1997 to facilitate comparison with Thorne et al. (2005).

The reduction of the stratospheric cooling is plausible since some radiosonde stations switched to more modern equipment in the 1990s. The ERA-40 bias correction described in Andrae et al. (2004) (middle left panel) had little impact on the zonal mean trends even in the stratosphere. If the more aggressively correcting new version of this algorithm is used (middle right panel), the stratospheric cooling is reduced but there is also substantial warming in the tropical and southern hemispheric troposphere. This apparent warming is introduced by using also the daily mean bg-obs for bias correction in this version of the Andrae et al. (2004) algorithm which draws the radiosonde trends more strongly to the ERA-40 bg trends.

These ERA-40 bg trends are shown in the lower left panel. One can see less cooling than the uncorrected radiosondes in the stratosphere but rather large warming throughout the troposphere. This tropospheric warming trend is likely spurious and caused by excessive latent heating over the oceans during the 1990s. This example shows how easy it is to introduce some spurious signal with the bias correction process.

The lower right panel of shows the zonal mean trends from ERA-40 analyses. The spurious tropospheric warming is much weaker than for the ERA-40 bg trends but it is not completely removed. The spurious cooling-heating pattern over Antarctica that is visible in the ERA-40 bg trends is almost absent in the ERA-40 analyses and the stratospheric cooling from the analyses is also more like the cooling diagnosed by radiosondes. One should note, however, that the ERA-40 bg trends as well as the ERA-40 an trends in this figure have been sampled only at places radiosonde records, i.e. where the influence of the radiosondes on the analyses is largest. At other places the trends from the analysed fields may be affected more strongly by the problem in the bg.

Figure 31 compares zonal mean trends for the periods 1958-1978 and 1979-2001, which have been analyzed also by Thorne et al. (2005) and Seidel et al. (2004). Data availability is a more serious problem during these periods. The RAOBCORE adjustment has little effects on trends in the period 1958-1978. Some slight warming of the tropical troposphere is evident. There is some interesting minimum of the general cooling trend in the northern hemisphere at about the tropopause level. The reason for this minimum is not fully understood. Trends for this period have been calculated separately for America, Europe, Russia and SE-Asia, respectively. The reduced cooling was evident in all regions except America.

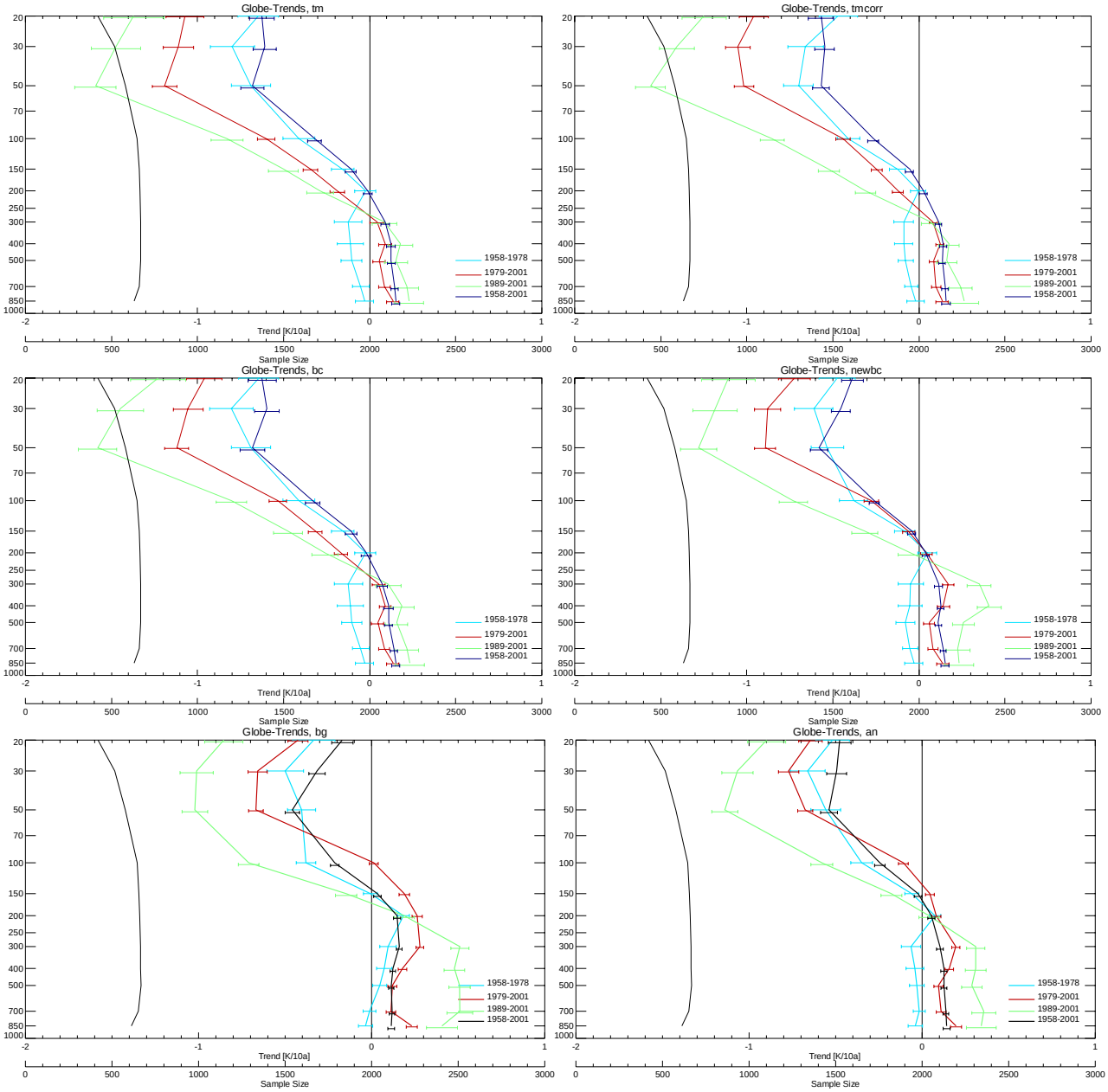

Figure 32: Global mean linear temperature trends for four different time periods. Order of panels is the same as in Figure 30. Radiosondes have been binned into 20x20 degree grid boxes with cos(lat) weighting prior to averaging. Black line indicates number of radiosondes used for calculating the means. Error bars indicate average values of 5% and 95% percentiles within the 20x20 grid boxes.

For the period 1979-2001 RAOBCORE leads to reduced cooling in the stratosphere, which can be traced back to the correction of documented changes of radiosonde types in the late 1980s and early 1990s. There is little difference between uncorrected and corrected trends over the northern hemispheric extratropical troposphere. There are, however, significant changes in the tropical troposphere. These can be explained by reduced radiation errors as well since they are evident down to the 300 hPa level. Most radiosonde sites in the tropics as well as over Australia and South America switched to more modern sounding equipment before 1990. The spurious trend of the bg in the 1990s has not contributed to the reduced cooling in the tropical troposphere. Otherwise the reduced cooling would be visible in the zonal means trends for 1989-2001 (see Figure 30, upper right panel) as well.

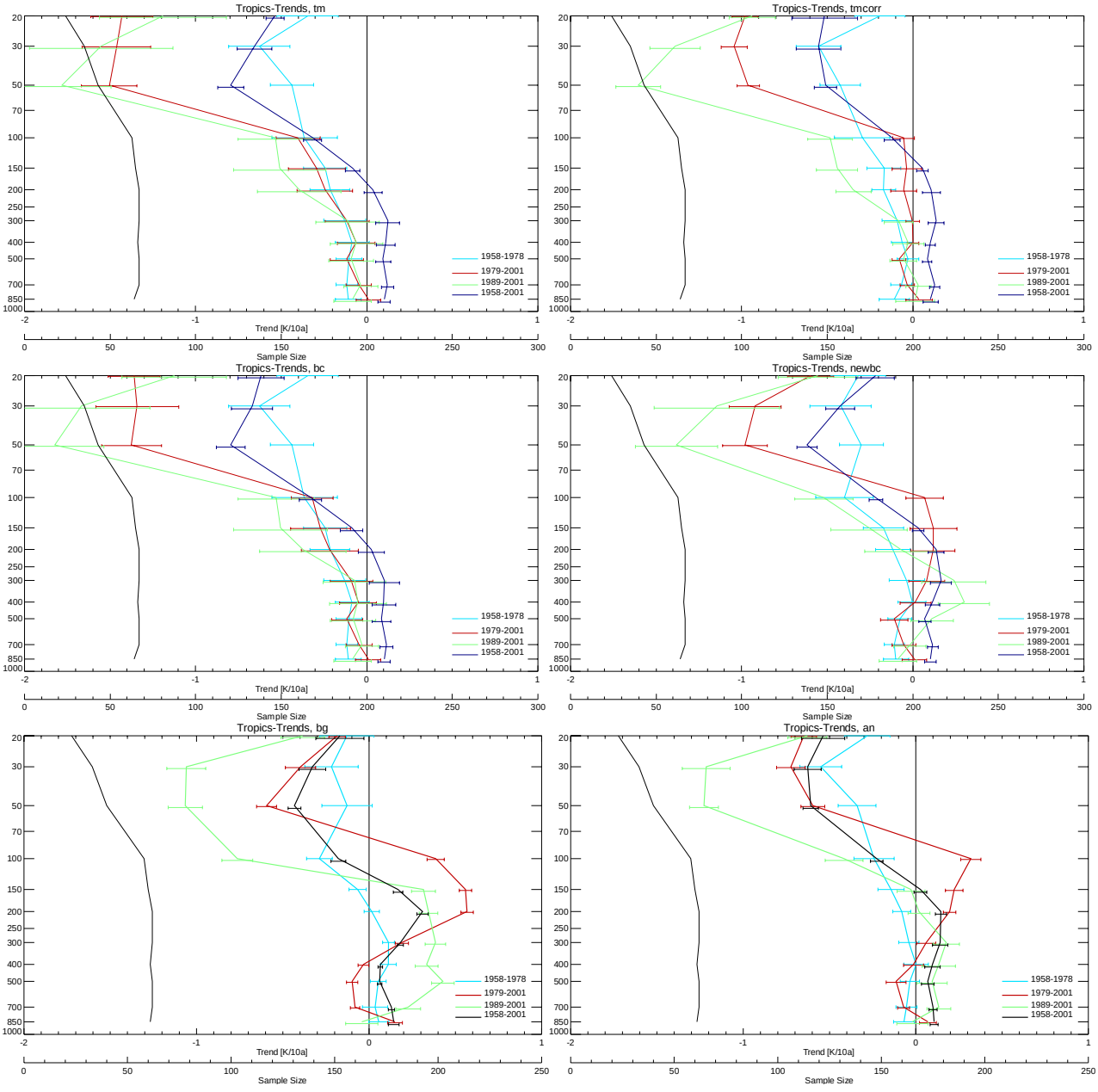

Figure 33: Same as Figure 32 but for trends in the tropics (20S-20N)

Figure 32 and Figure 33 show vertical profiles of radiosonde temperature trends, averaged over the globe and over the tropics (20S-20N), respectively. In all panels trends of four time periods are plotted, together with the number of stations used for calculating the trends. The horizontal bars indicate the average variability of the trend estimates in 20x20 grid boxes. The 5% and 95% percentiles are indicated. The variability is caused partly by artificial trends due to breaks in the time series but has also natural causes since the trend variability on a pressure level within these large boxes is not zero. The variability of the corrected trends (upper right panels) tends to be smaller than the variability of the uncorrected trends (upper left panels).

In the global mean shown in Figure 32 the uncorrected radiosonde temperatures yield strong stratospheric cooling, exceeding 1K/decade, especially in the periods 1989-2001 and 1979-2001. In the periods 1958-1978 and 1958-2001 the cooling trend is weaker (ca. 0.6K/decade). In the troposphere below 300 hPa a heating trend of 0.1-0.2 K/decade is indicated, except for the period 1958-1978 which shows cooling of up to 0.1K/decade.

The trend profiles from time series adjusted with RAOBCORE (upper left panel) are similar, but the stratospheric cooling is reduced for all four time periods. The reduction of the cooling is strongest for the period 1979-2001 where it is about 0.2K/decade. Almost no change is visible in the troposphere. The adjusted profiles are in good agreement with the profiles shown by Thorne et al. (2005).

The bias correction used in ERA-40 (middle left panel) has a similar but weaker effect. The more aggressive version of this algorithm (middle right panel) has a stronger effect. However, there is more tropospheric warming in the period 1989-2001. The global mean trend profiles from the bg time series show less cooling in the stratosphere. The trends in the troposphere differ strongly from those calculated from the radiosonde time series. The bg shows a strong spurious tropospheric warming trend in the period 1989-2001. The other periods also show stronger warming. The maximum warming occurs in the upper troposphere whereas the radiosonde time series indicate the maximum warming close to the surface. The trends from the analyses also show a tropospheric warming in the periods 1979-2001 and 1989-2001 but it is much less than the bg trend.

When looking at the tropics in Figure 33 the features observed in the global temperature time series become more pronounced. The stratospheric cooling trend in the period 1979-2001 is reduced from about 1.5K/decade to 1K/decade by the RAOBCORE algorithm. A slight reduction of the cooling is also evident for the period 1958-1978. The region of vertically almost constant trends reaches higher up in the tropics. This is plausible since the troposphere reaches higher up in the tropics. The spurious warming trend in the ERA-40 bg time series is most evident in the tropics both in 1979-2001 and 1989-2001 where it deviates by more than 0.5K/decade from the trends derived from radiosonde time series. The analyzed trends are drawn to the radiosonde trends but to a lesser extent than in the global mean, probably because of the scarcity of radiosonde observations in the tropical belt. The bg trends as well as the an trends show warming in the tropical troposphere in the later periods whereas the unadjusted as well as the adjusted radiosonde trends show cooling.

This section, which contains the main results of this report, has shown that (i) RAOBCORE substantially improves the spatiotemporal consistency of the radiosonde temperature time series. (ii) The observed radiosonde temperature trends are relatively weakly affected by the correction procedures. The main effect of the correction is a reduction of the stratospheric cooling trend by up to 0.5K/decade, especially in the tropics, in the period 1979-2001. (iii) The spurious trends evident in the bg temperature time series have the potential to introduce spurious signals also in the corrected radiosonde temperature time series. However, such spurious signal is evident neither in the bias correction procedure applied in ERA-40 nor in the RAOBCORE

correction. (iv) Trends from the ERA-40 an temperature series are affected by the spurious warming of the troposphere but are much closer to the observed radiosonde trends than are the bg trends.

5.4 Global maps of obs-bg differences

So far we have focused on stations with long-term time series that allowed calculation of trend estimates. These stations form the backbone of the radiosonde observation network and are valuable for climatological purposes. A large fraction of the radiosonde observations belong to time series, however, which are short or have big gaps. For these stations it is more useful to compare annual mean obs-bg differences or annual mean day-night differences[1]. An adjustment algorithm suitable for reanalyses must be able to sensibly correct these data as well. The differences also provide information about the consistency of actual temperatures compared to temperature trends. While it has been demonstrated above that spurious breaks can be removed, there may be still biases since in some countries even the most recent radiosonde observations are strongly biased, as is shown in Figure 34. Without adjustment of these biases in the most recent periods the whole time series remains biased.

In general the observation bias of the radiosondes is smaller than the bg bias. Therefore one must have good reasons for adjusting the mean value of the whole time series. In this report, only the time series of radiosondes with WMO IDs as specified in Table 2 have been adjusted. The adjustment procedure was simply to add the bg-obs difference of the most recent 400 launches to the obs time series. The effect of this adjustment can be seen in Figure 34. The obs-bg differences are spatially much more homogeneous after the adjustments. Some differences remain, especially in the polar regions and in the inner tropics, where model biases affect the bg temperatures. Some residual error remains after adjustment for the Indian radiosonde stations which still have significant errors in the period 1999-2001. No attempt has been made to adjust the different radiosonde types in use over the US, except station Oakland.

17000	17399	Turkey	78000	78880	Caribbean
20000	38999	Russia	78900	80001	Caribbean
42000	44000	India	82193	82193	Belem
46000	47399	Korea/Taiwan	82332	82332	Manaus
48000	48600	Thailand/Myanmar	82599	82599	Natal
48920	48999	Laos/Cambodia	83600	84000	South Brazil
50000	59999	China	91285	91285	Hilo (Hawaii)
62300	62599	Egypt	91366	91376	Majuro Kwajalei
72493	72493	Oakland			

Table 2: List of Radiosondes where mean temperature at the period 1999-2001 has been adjusted to the bg temperature. See Figure 34 for the effect of the adjustments on heterogeneity of obs-bg differences.

The second issue to be addressed is the case of time series that end before 1999. The most recent part of these radiosondes is adjusted to the bg, again using the obs-bg of the most recent 400 launches. For the radiosondes ending before 1999 no distinction with respect to station ID is made, i.e. all of these radiosonde time series are adjusted. This may not be always justified but in general these radiosonde sites did not use the most modern radiosonde equipment before they were phased out. In a future evaluation it is preferable to bias-correct the most recent part of the time series ending before 1999 with composites of already adjusted neighboring radiosondes instead of the bg.

[1] Only in this section obs-bg differences are shown instead of bg-obs differences to stress bias differences between radiosondes. The bg temperature is a quite smooth field (see Figure. 1)

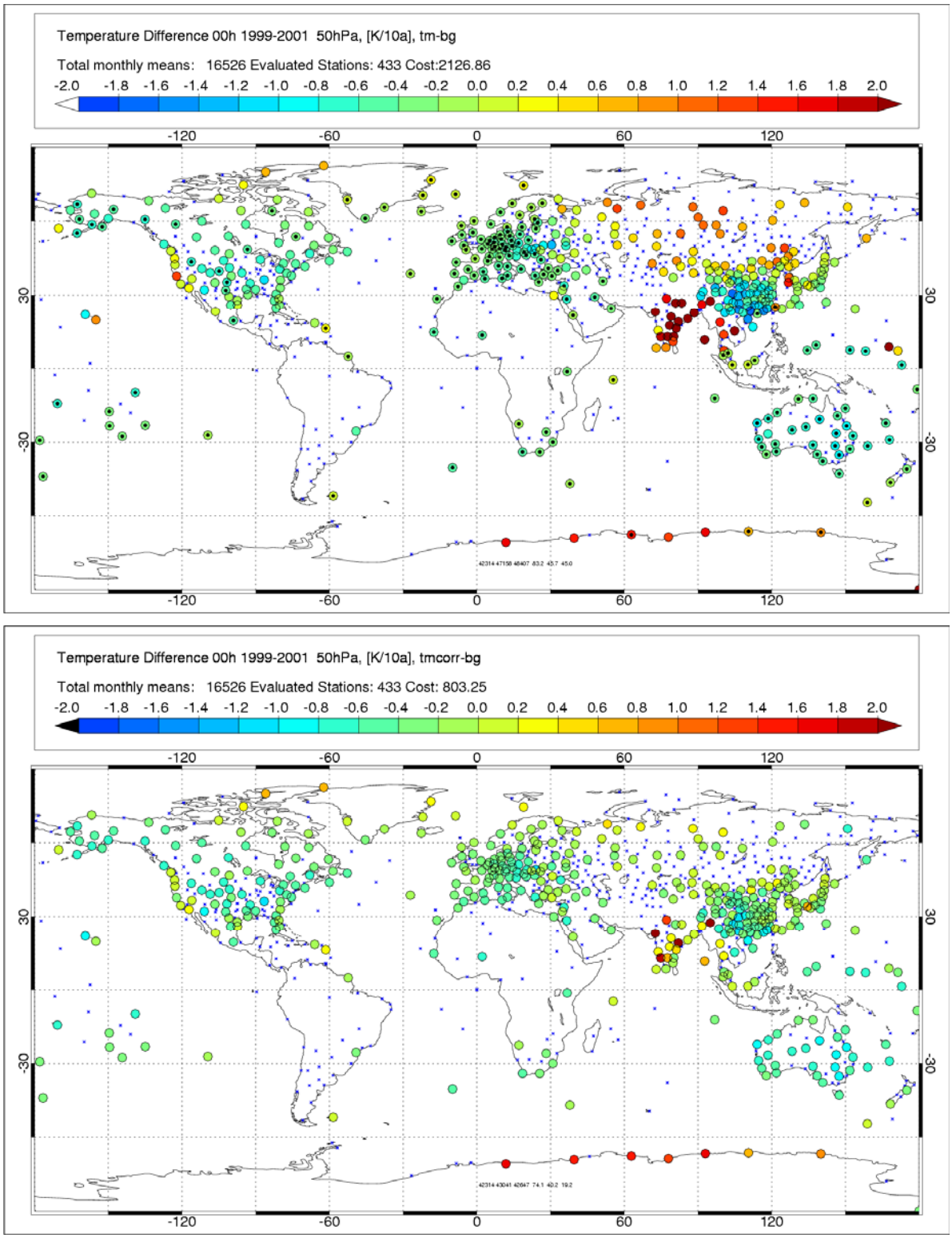

Figure 34: obs-bg temperature difference averaged over 1999-2001 in 50 hPa. Upper panel: obs unadjusted. Lower panel: obs from Table 1 adjusted to bg. Radiosonde sites using Vaisala radiosondes according to Aguilar et al. (2001) are marked with a central black dot in the upper panel.

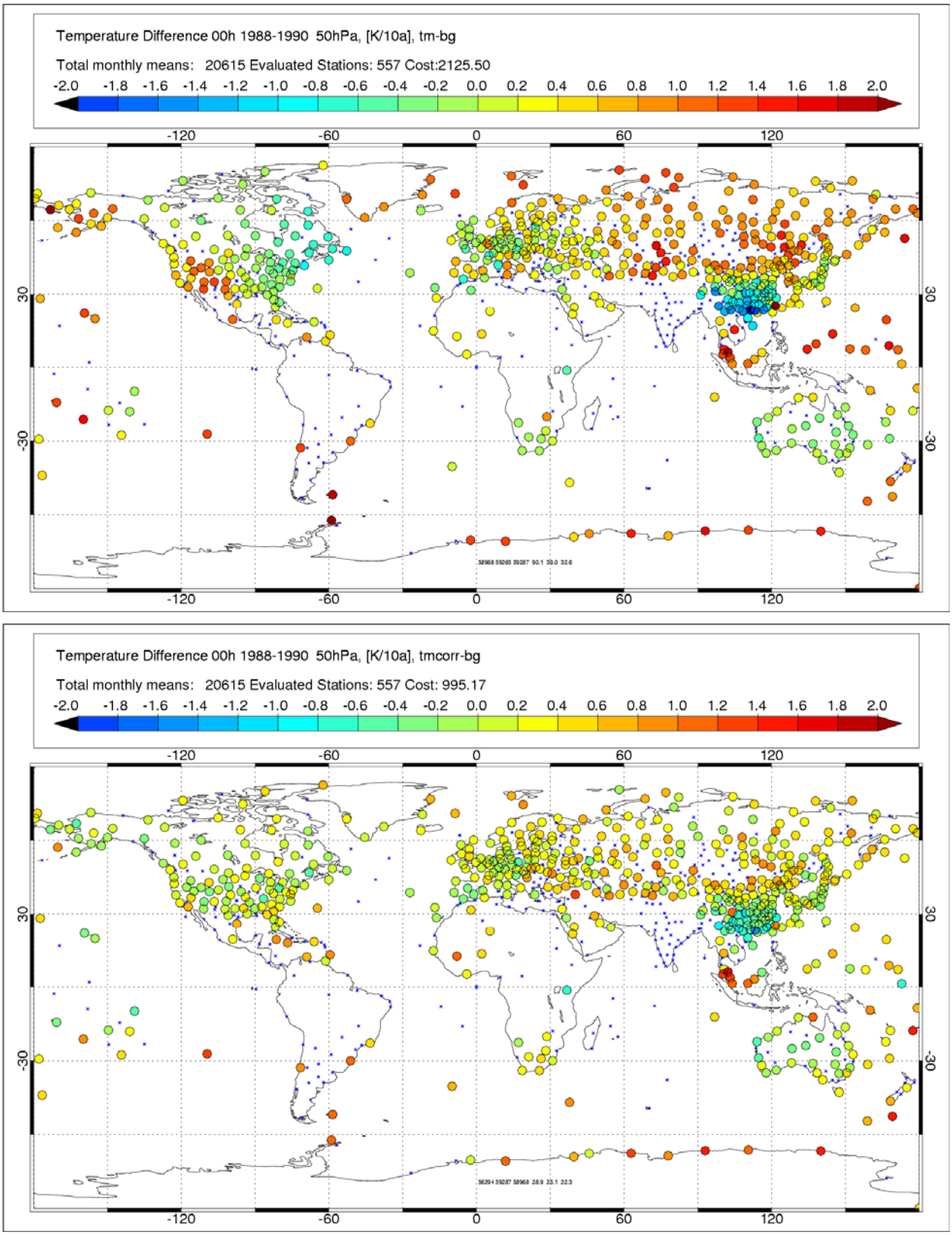

Figure 35: obs-bg differences for the period 1988-1990 from uncorrected obs (upper panel) and from obs adjusted with RAOBCORE (lower panel).

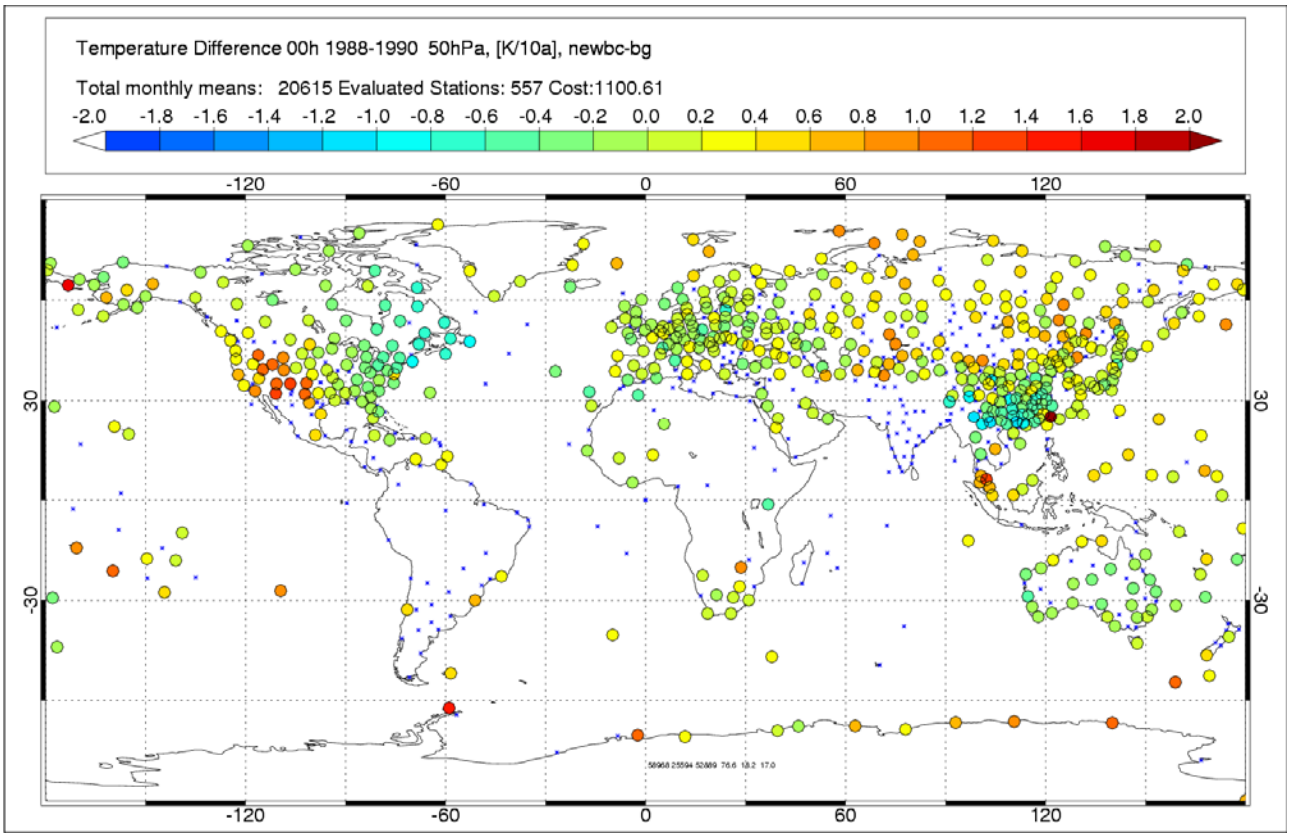

Figure 36: obs-bg differences for the period 1988-1990 from obs corrected with the improved version of Andrae et al. (2004). Spatial heterogeneity of differences is larger than in Figure 35.

When the radiosondes are homogenized and the most recent part is adjusted, the obs-bg differences using the corrected obs should be spatially much more homogeneous than the uncorrected differences.

Figure 35 shows that this is indeed the case for the period 1988-1990 (the starting period for the planned interim reanalysis). There are a few outliers, which are related to time series with large gaps which could not be bridged by the homogenization algorithm. There is a clear overall improvement, also compared with the bias adjustment method used in ERA-40 (Figure 36), which leaves spatially more heterogeneous obs-bg differences than the RAOBCORE-adjusted temperatures.

Figure 37 shows the uncorrected obs-bg differences for the periods 1964-1966 and 1980-1982, 1988-1990, 1999-2001, Figure 38 shows the corrected obs-bg differences. A clear improvement is visible also for these periods although the effect of data gaps leads to more outliers in the earlier periods. While the RAOBCORE-adjusted temperatures seem ready for use in reanalyses of the satellite period, some improvements are still necessary in the earlier periods.

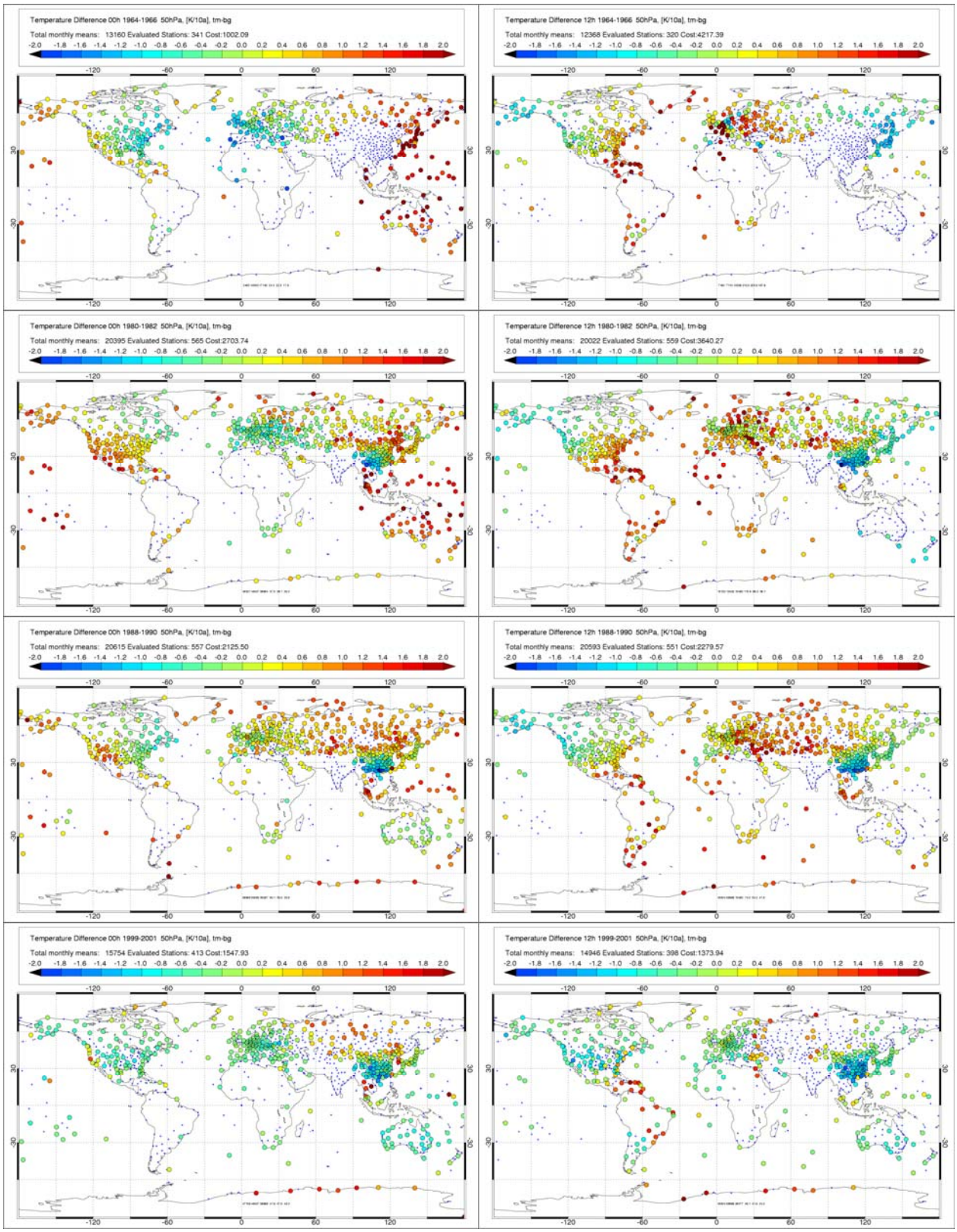

Figure 37: obs-bg differences for the periods 1964-1966, 1980-1982, 1988-1990 and 1999-2001 calculated from the unadjusted radiosonde time series.

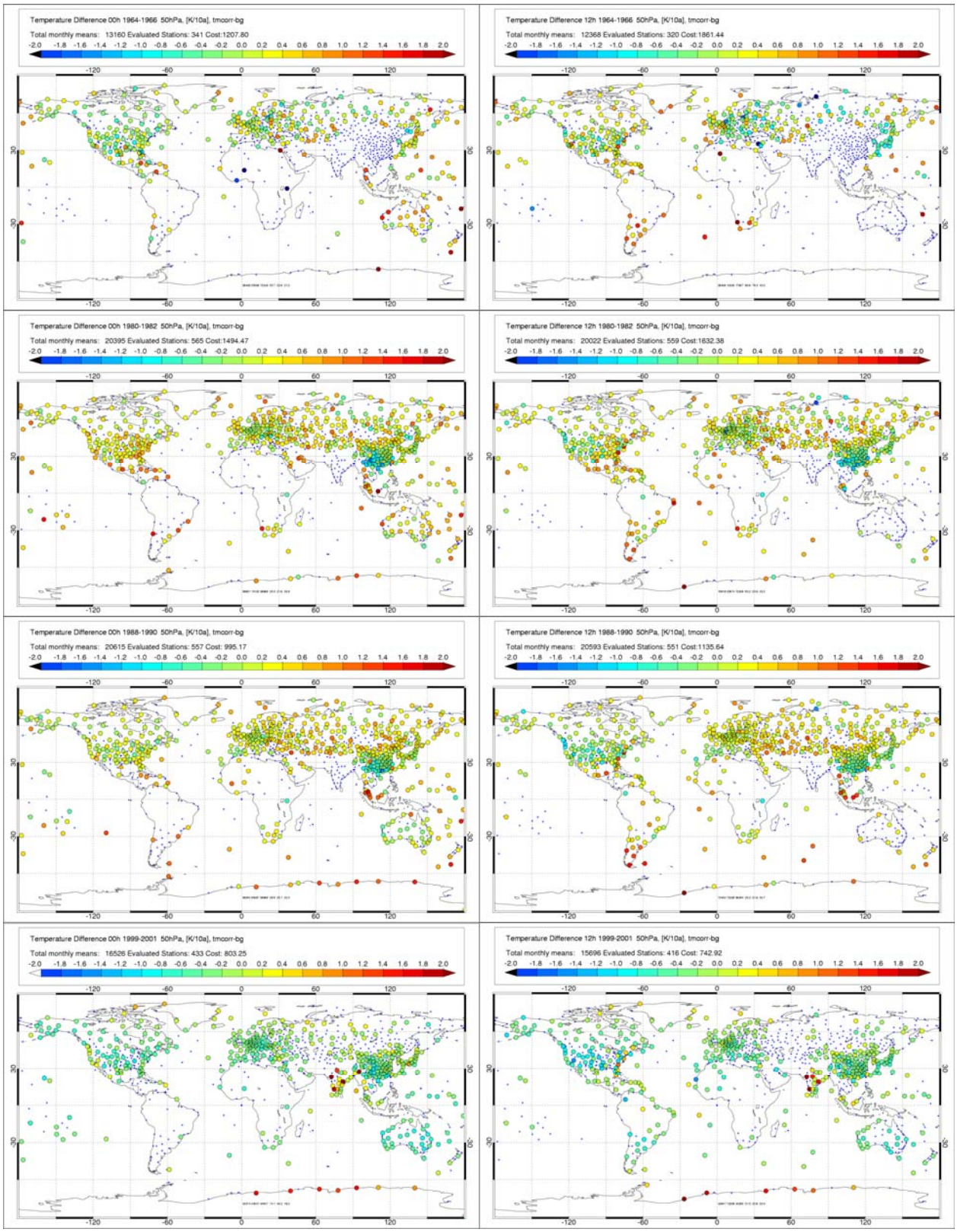

Figure 38: obs-bg differences for the periods 1964-1966, 1980-1982, 1988-1990 and 1999-2001 using the adjusted radiosonde time series.

5.5 Day-night time differences

Day-night time differences are a classical measure for the magnitude of the radiation error. While the diurnal cycle in the stratosphere is not zero, it is typically less than 0.5K in 50 hPa. The analysis of annually averaged obs(12GMT)-obs(00GMT) differences is a valuable tool for estimating radiosonde temperature biases and of the correction performance independent of the ERA-40 bg. The analysis of the bg(12GMT)-bg(12GMT) differences allows checking of how the radiation errors evident in radiosonde observations also affect the bg temperatures.

Figure 39 shows the 12GMT-00GMT obs difference in the pre-satellite era (1964-1966). Large differences are visible, which depend heavily on longitude and on the operating country. Over the US, for example, the 12GMT-00GMT difference is negative in the western US and Alaska, where it is daytime at 00GMT. In the central US the difference is small since there is dawn/dusk both at 00GMT and 12GMT. In the eastern parts of the US the difference is positive. Over Europe the difference is generally positive, but not everywhere. The West-German radiosondes had a “reverse” radiation error at this time; the French radiosonde had an extremely large radiation error. Over Scandinavia the radiation error is relatively small already at this time. Over Asia, the Russian as well as the Japanese radiosondes had a relatively large radiation error. The Chinese radiosondes did not reach the 50 hPa level at this time.

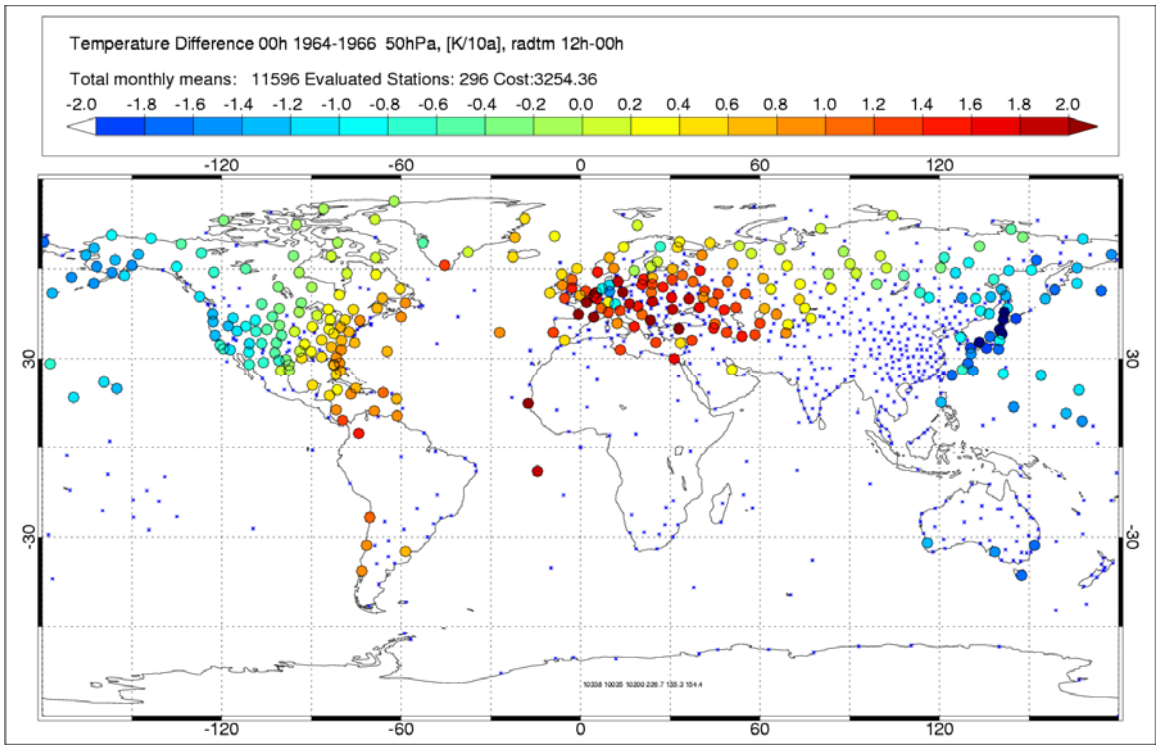

Figure 39: 12GMT-00GMT differences as measured from radiosondes in 50 hPa. Dependence of bias on solar elevation is evident over the US. There is strong spatial heterogeneity over Europe.

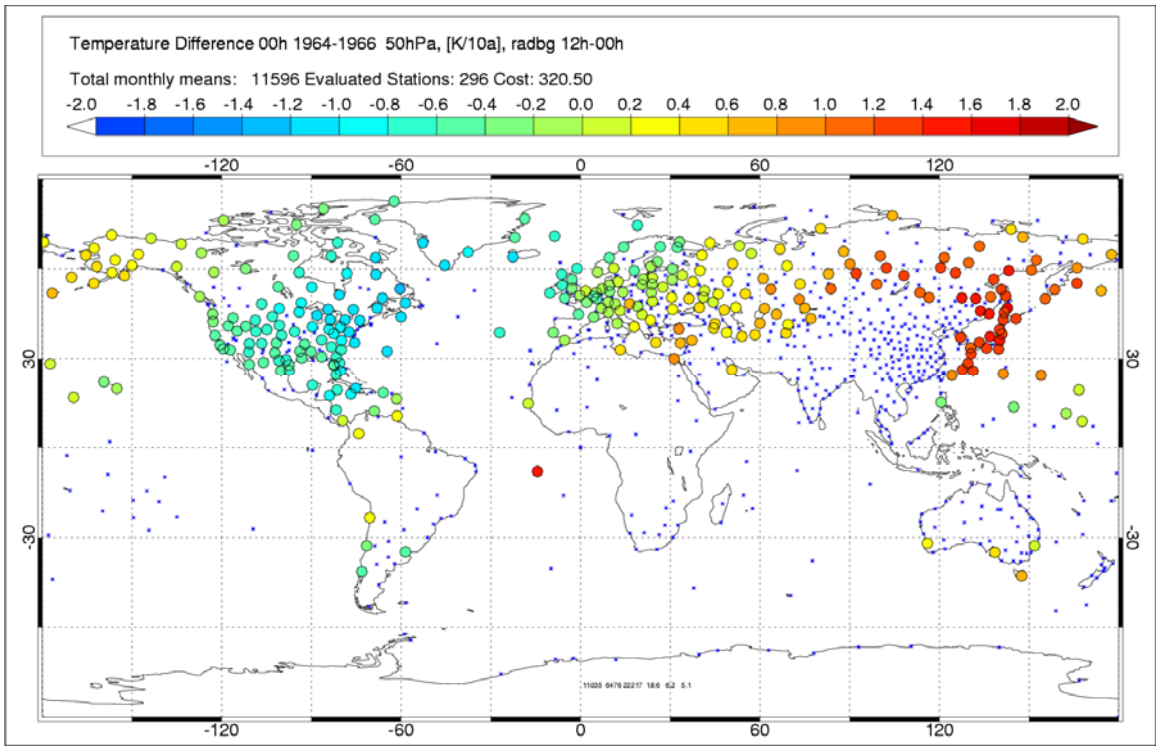

Figure 40: 12GMT-00GMT bg temperature difference in 50 hPa in the pre-satellite era. Too warm radiosonde temperatures at daytime cause too warm bg 12 hours later. Therefore the (spuriously strong) day-night differences measured by the radiosondes induce day-night differences of the bg with the opposite sign.

Figure 40 shows the 12GMT-00GMT bg differences. Since no satellite data have been available to correct the radiation error, the analyses are also biased and the bg forecasts keep the daytime positive bias of the radiosondes for the next 12 hours. At nighttime just the opposite is true. This effect leads to a 12 hour phase shift of the 12GMT-00GMT bg differences compared to the obs differences. Besides the phase shift it has to be noted that the 12GMT-00GMT differences of the bg are much smaller compared to the obs differences, especially over Europe and the US.

In later periods as shown in Figure 1 and Figure 41, the 12GMT-00GMT observed temperature differences are still spuriously large over some regions. The situation over Europe, China and Japan has markedly improved, however. The 12GMT-00GMT differences of the bg in Figure 1 are much smaller since the bg is heavily influenced by two TOVS satellites that do not have a solar-elevation dependent radiation error. Even over the US, which is densely covered with radiosondes, the satellite observing system had enough influence to correct the spurious 12GMT-00GMT differences in the observations.

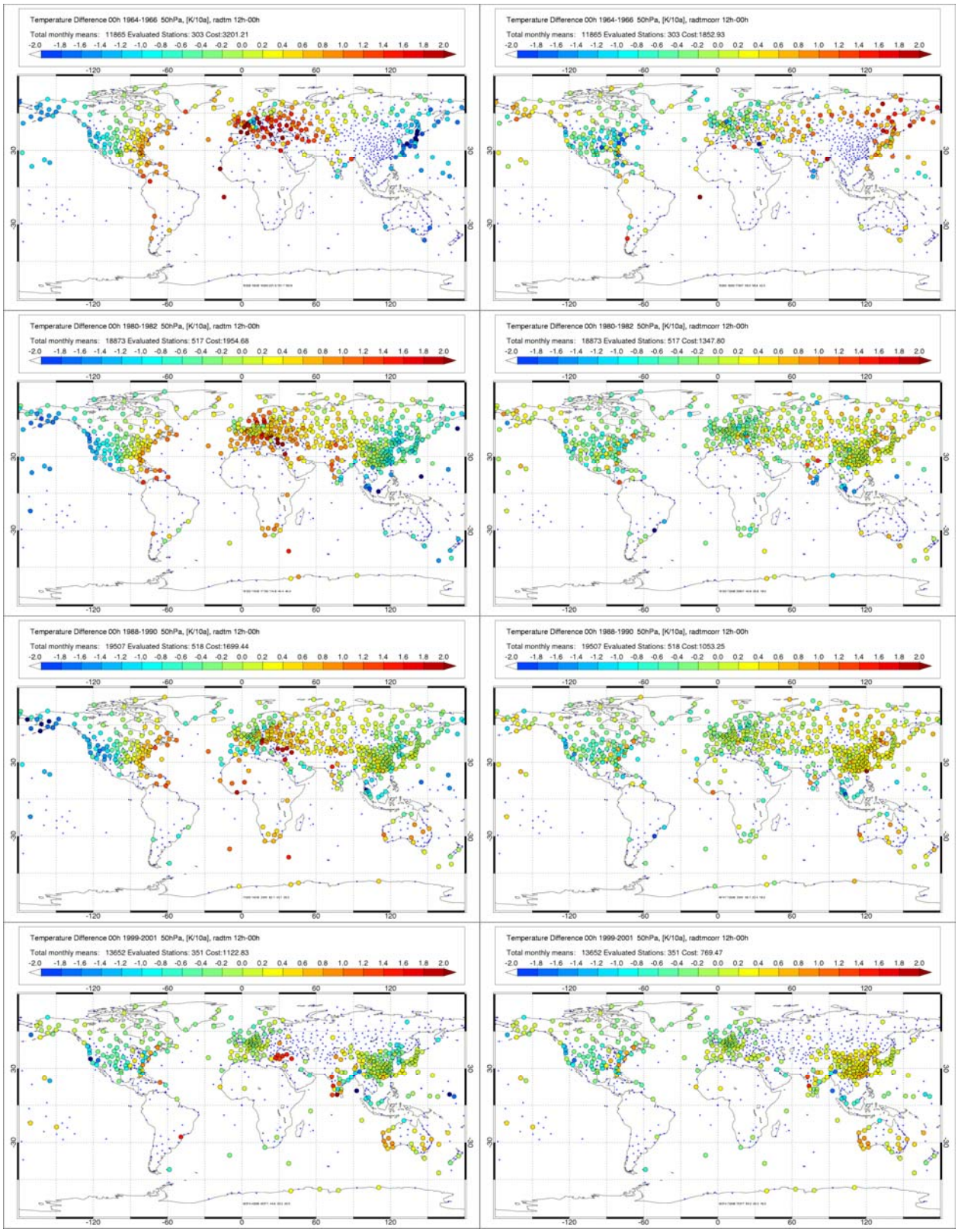

Figure 41: obs(12GMT)-obs(00GMT) differences for the periods 1964-1966, 1980-1982, 1988-1990 and 1999-2001 for the uncorrected (radtm, left) and corrected (radtmcorr, right) radiosonde observations.

Figure 41 shows how the RAOBCORE adjustments affect the obs(12GMT)-obs(00GMT) differences. The lowest two panels show this difference in 1999-2001. The radiosondes listed in Table 1 have been adjusted to the bg in the right panel. One sees that this measure removes the spurious obs(12GMT)-obs(00GMT) difference at most places, e.g. over Turkey. One sees, however, that a few stations with strong radiation error remain over the US. Some of the US stations have not been upgraded in order to allow intercomparison between the old and more modern radiosonde types. The two middle panels show the effect of the RAOBCORE correction in the 1980s. The radiation error is significantly reduced. In the period 1964-1966 the radiation error is also reduced. There is even some evidence of overcorrection over Japan and Alaska, however. The overcorrection is probably a result of the 12GMT phase shift of the bg(12GMT)-bg(00GMT) difference at this time (see Figure 39) and must be addressed in future versions of RAOBCORE.

5.6 Some further examples of adjustments at specific radiosonde stations

In the beginning of this report, the station Jan Mayen has been examined as an example to explain the analysis tools developed for break detection and adjustment. This section is devoted to describing adjustment results from other radiosonde stations with interesting station history.

5.6.1. Stations in NW-Russia

From 1997-2001 a joint project between the Russian meteorological agency and the Finnish Meteorological Institute has been carried out to foster the introduction of Vaisala-RS80 temperature and humidity sensors on Russian radiosondes (ROSHYDROMET, 2001). Three test sites have been equipped with the new sensors: Murmansk (22113) from September 1999 onwards, St. Petersburg (26063) from December 1998 onwards and Kandalaksha (22217) from January 2000 onwards.

Figure 42 shows difference time series for these three stations for the stratospheric layer (50-150 hPa) used for break detection. For Murmansk and St. Petersburg, 12GMT-00GMT differences are shown while for Kandalaksha the 12GMT bg-obs difference is shown since only very few 00GMT observations have been available. In all three time series, one can see the radiation error before 1999. It has lead to spurious day-night differences of ca. 0.6K at Murmansk and St. Petersburg and to a positive bias of the 12GMT observations compared to the ERA-40 bg at Kandalaksha. All three time series show a maximum of the SNHT around 1999. Figure 43 shows the break profiles detected by RAOBCORE for the three stations. Negative temperature adjustments are suggested for all three stations for the 12GMT observations. For the 00GMT observations the adjustments are about neutral to slightly negative at high altitudes. The dates in the title of the three panels specify the detected time of the break. They are quite close to the dates noted above, although the CARDS metadata are not accurate (CARDS suggests 199910 for all three stations). Figure 44 shows the time series after adjustment. The annual mean bias due to the radiation error is removed from the time series prior to 1999 at all three stations.

Two notable problems remain: Firstly, the dependence of the radiation error on the solar angle which is the main cause for the variability of the corrected time series is not removed. It should be possible to correct the radiation error with an approach similar to Andrae et al. (2004). Secondly, the time series of Kandalaksha before 1980 is not correctly adjusted since RAOBCORE cannot yet cope with data gaps larger than 4 years.

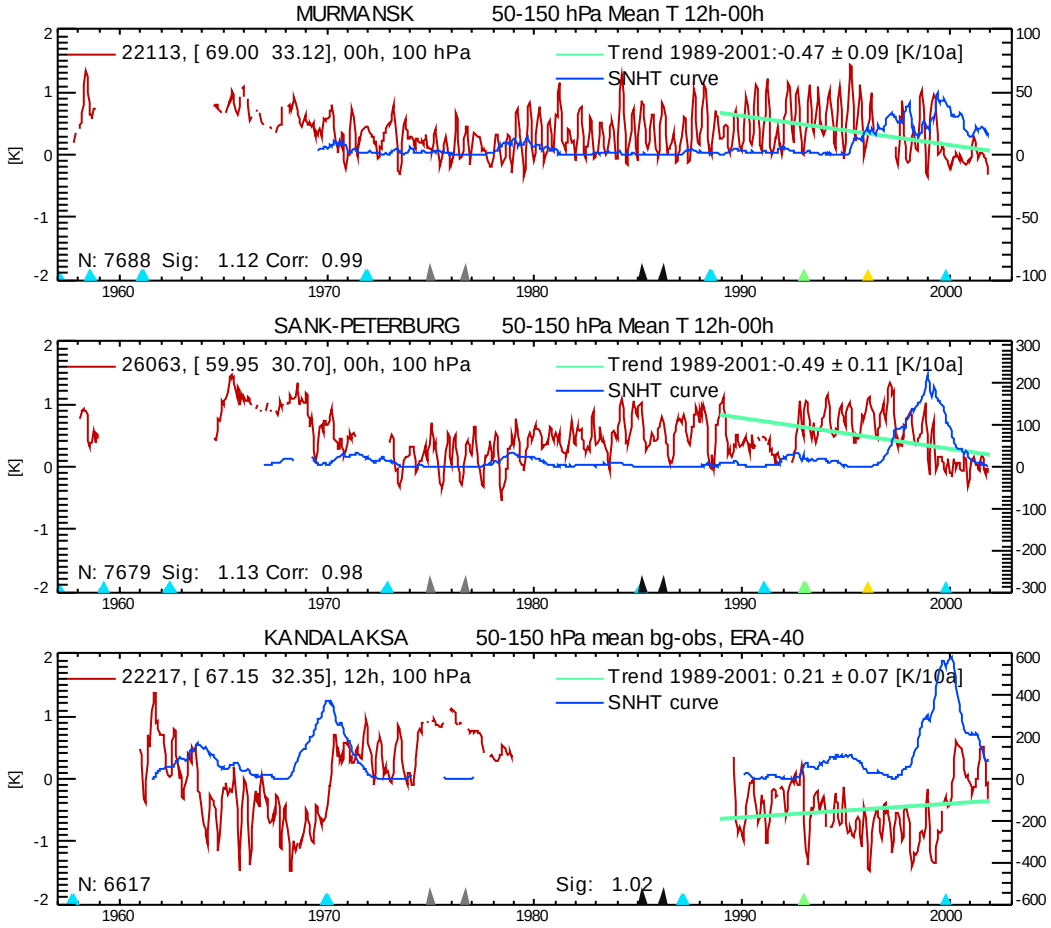

Figure 42: Stratospheric layer mean difference time series for three radiosonde stations of the RF95-NW project before adjustment with RAOBCORE: a) obs(12GMT)-obs(00GMT) for Murmansk, b) obs(12GMT)-obs(00GMT) for St. Petersburg, c) bg-obs difference at 12GMT at Kandalaksa. At Kandalaksa too few obs(12GMT)-obs(00GMT) have been available. Therefore the bg-obs difference at 12GMT is shown instead.

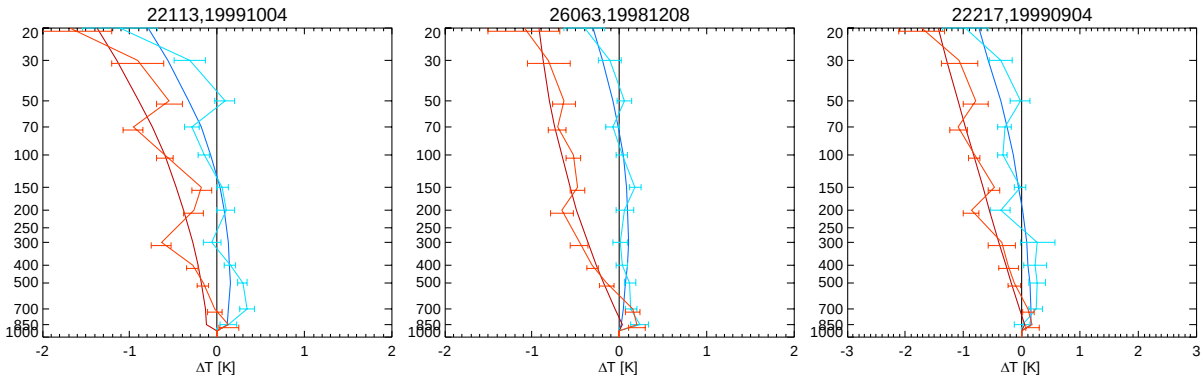

Figure 43: Profiles of adjustments suggested by RAOBCORE algorithm. The dates in the headers refer to the breakpoint detected by RAOBCORE.

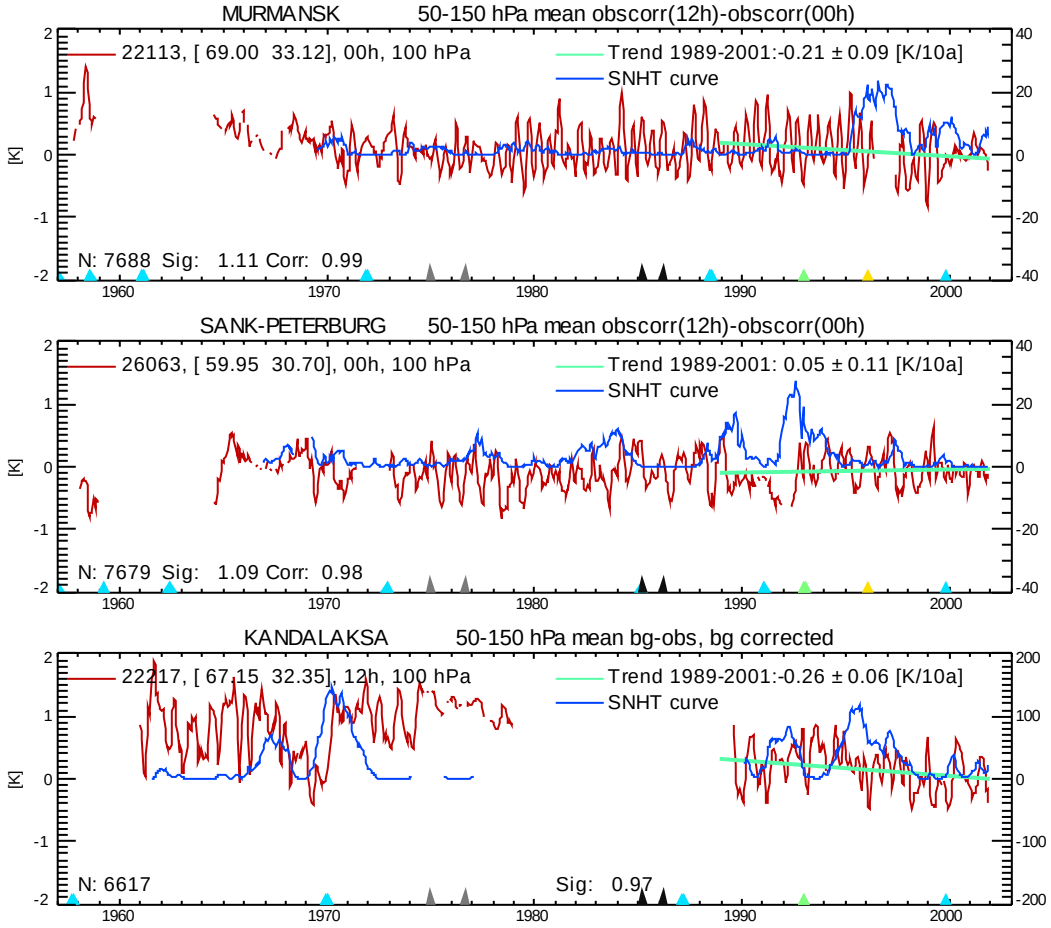

Figure 44: Stratospheric layer mean difference time series for three radiosonde stations of the RF95-NW project after adjustment with RAOBCORE: a) obs(12GMT)-obs(00GMT) for Murmansk, b) obs(12GMT)-obs(00GMT) for St. Petersburg, c) bg-obs difference at 12GMT at Kandalaksa.

5.6.2. Breaks and their adjustment at Bethel (70219, Alaska), Trappes (7145, France), Saigon (48900, Vietnam) and Darwin (94120, Australia)

In Bethel (Alaska) VIZ radiosondes have been used until 1989. From July 1989-July 1995 Space Data radiosondes have been in use. From November 1995 onwards Vaisala RS80 radiosondes have been used. The three radiosonde types have rather different measurement characteristics as is visible from Figure 45. As can be seen from the time series of the adjustment, RAOBCORE is capable of adjusting this time series. Figure 46 shows that the breaks have an almost linear profile in a log(p) vertical coordinate system. Again the dependence of the obs-bias on interannual variations of the solar angle is evident and should be corrected in a later version of RAOBCORE. No metadata events are reported between 1961 and 1989 but several small breaks have been detected and corrected. A detailed investigation in the individual station history would be necessary to decide whether all these small breaks can be traced back to changes in observation practice or are the result of a too sensitive detection algorithm.

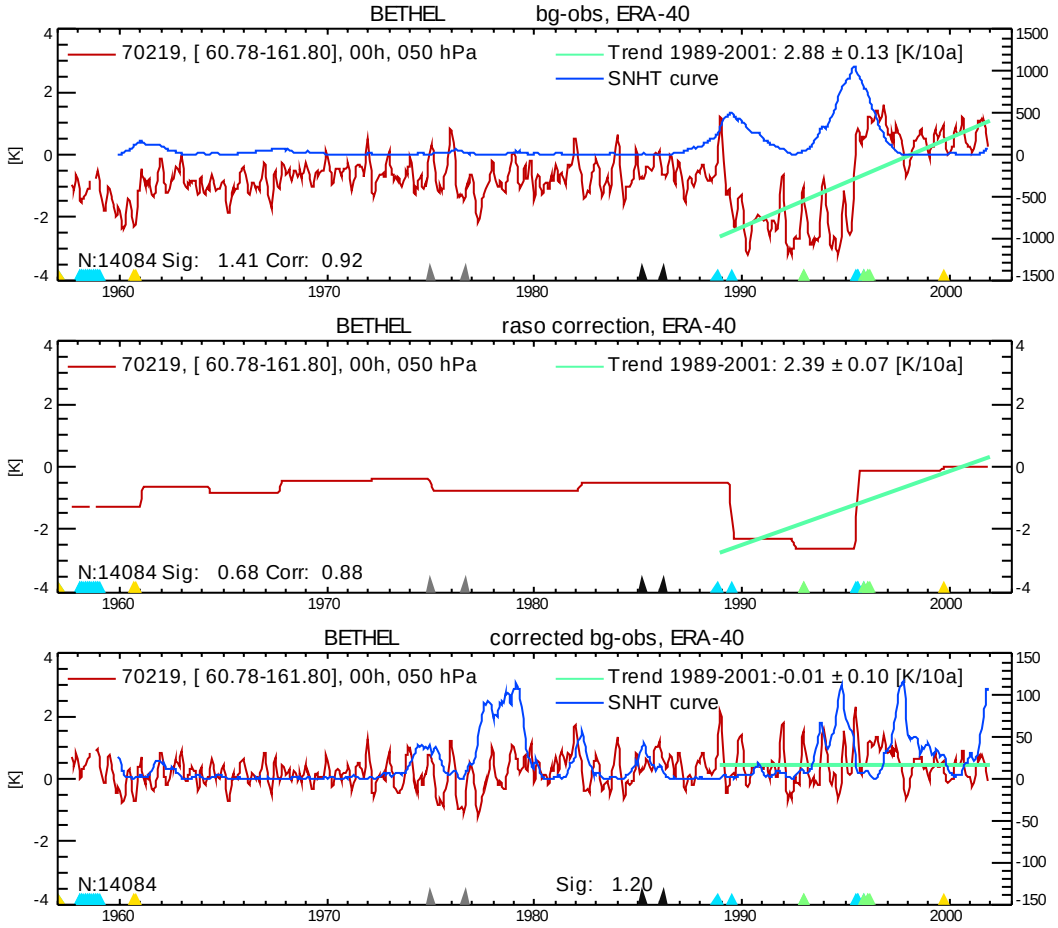

Figure 45: Adjustment of 50 hPa time series of radiosonde station Bethel (Alaska) at 00GMT. a) Uncorrected bg-obs series at 00GMT, b) adjustment applied to radiosonde time series, c) bg-obs series after adjustment

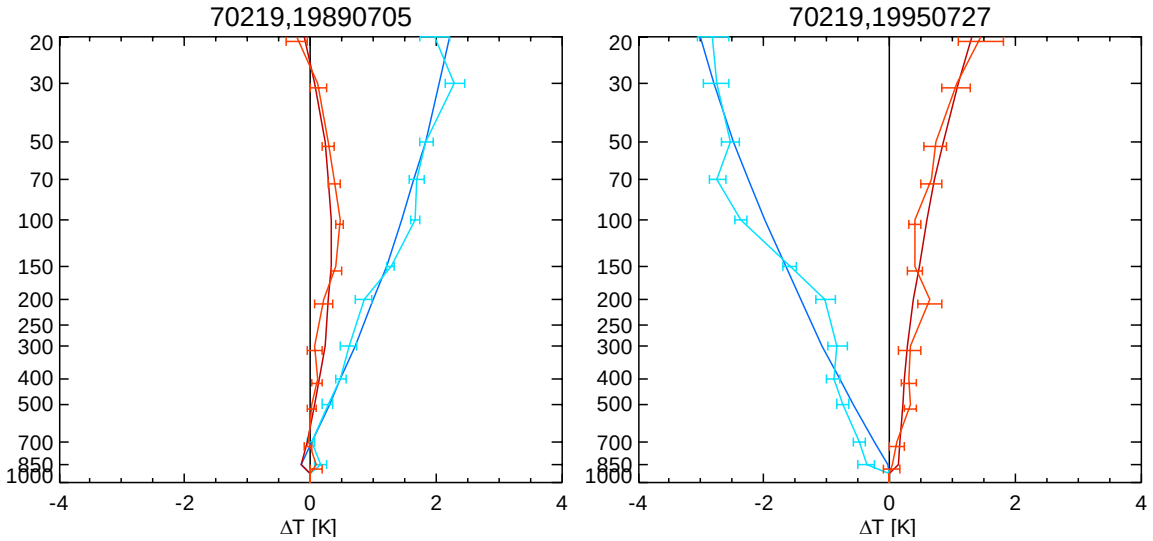

Figure 46: Break profiles for two major breakpoints in the time series of Bethel. The first break was caused by a change from VIZ to Space Data radiosondes, the second break by a change from Space Data radiosondes to Vaisala radiosondes.

The radiosonde station Trappes is affected by well known problems with MESURAL radiosondes in the 1960s. The temperature measurements had very strong radiation errors prior to 1972, as is visible from the bg-obs difference series in Figure 47. Most of the errors could be removed although some residual break is still evident in 1971/1972 in the lower panel of Figure 47. The reason is the large uncertainty involved in the break estimate (right panel of Figure 48). Only few launches have been accepted by the ERA-40 data assimilation system between 1969 and 1972 (on the issue of quality control see also Figure 63).

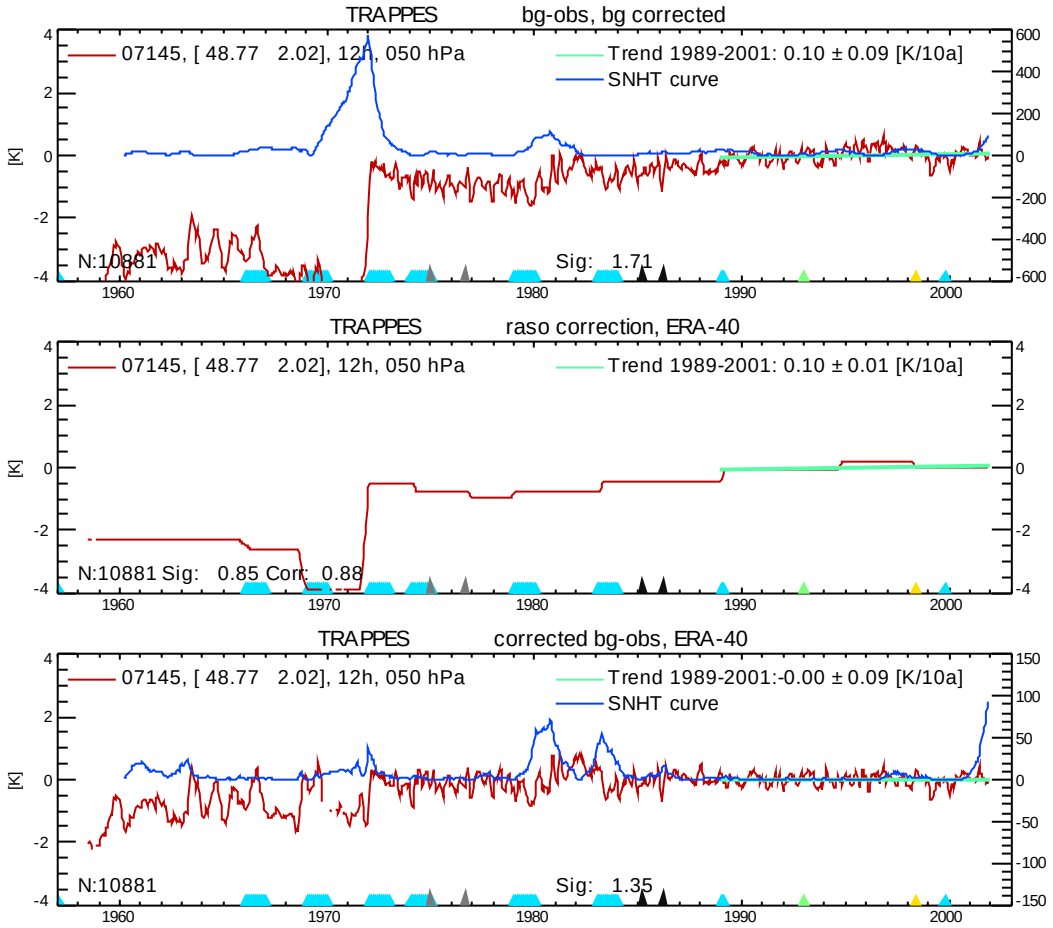

Figure 47: Unadjusted time series, adjustment and adjusted time series at 12GMT at 50 hPa for radiosonde station Trappes (France).

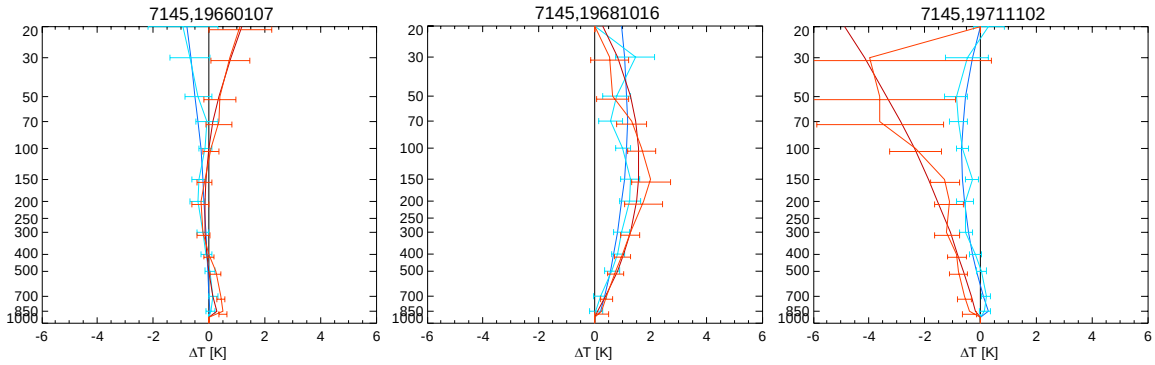

Figure 48: Break profiles for most critical part of time series at Trappes. Note large uncertainties in estimate of largest break in 1971. The radiosondes between 1969 and 1971 only rarely reached levels above 100 hPa.

The radiosonde station Saigon (Figure 49 and Figure 50) has a quite interesting station history as well. Before 1977 French radiosondes were used at this site. Between 1977 and July 1994 Russian A22 radiosondes were used. From July 1994 onwards Vaisala radiosondes have been used. Time series at the 200 hPa level have been chosen for this site since only few launches reached higher levels between 1977 and 1994. The break profiles show that the A22 radiosondes had a large positive bias already in the troposphere. Since the stratospheric levels above 100 hPa are mostly missing between 1978 and 1994, the adjustment method had to resort to analysis of the tropospheric (300-700 hPa) levels. This may be the reason why the breakpoint in 1978 is detected too early which led to the spike in the corrected profile. It is also the reason why no correction is performed in 1982 where there is an apparent break in the 200 hPa bg-obs series but no break is visible at all in the tropospheric bg-obs series (not shown). Although RAOBCORE leads to an improved time series, it is obvious that there is still potential for improvement.

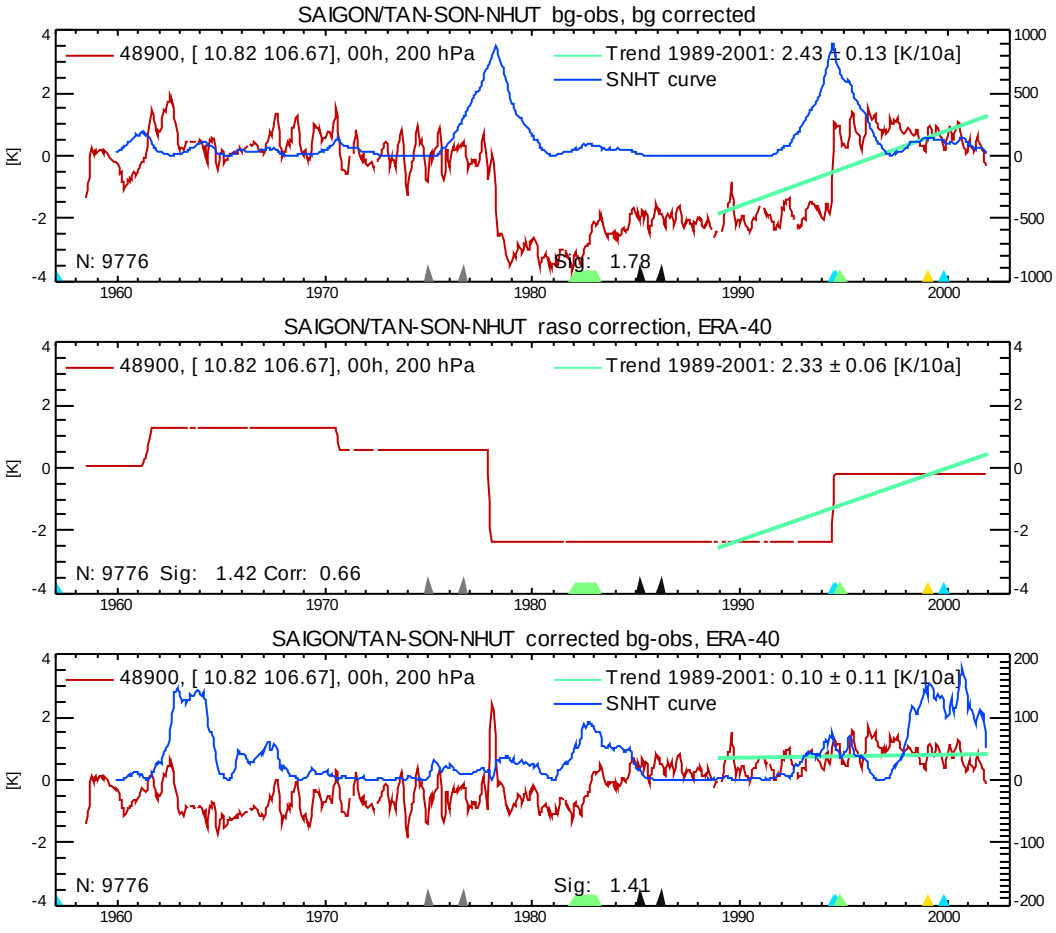

Figure 49: Unadjusted bg-obs time series, adjustment and adjusted bg-obs time series at the 200 hPa level at 00GMT for radiosonde station Saigon.

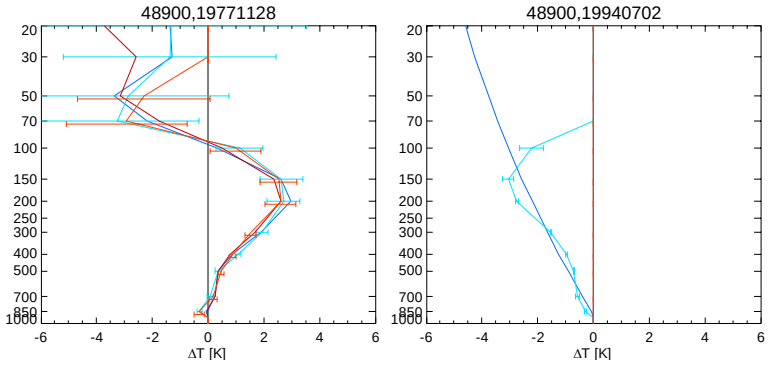

Figure 50: Adjustment profiles at the two largest breakpoints at Saigon. Before 1977 French radiosondes were in use at this site, then Russian A22 radiosondes, from 1994 onwards Vaisala radiosondes have been used. Only few launches reached 70 hPa between 1977 and 1994, leading to uncertain break estimates above this level.

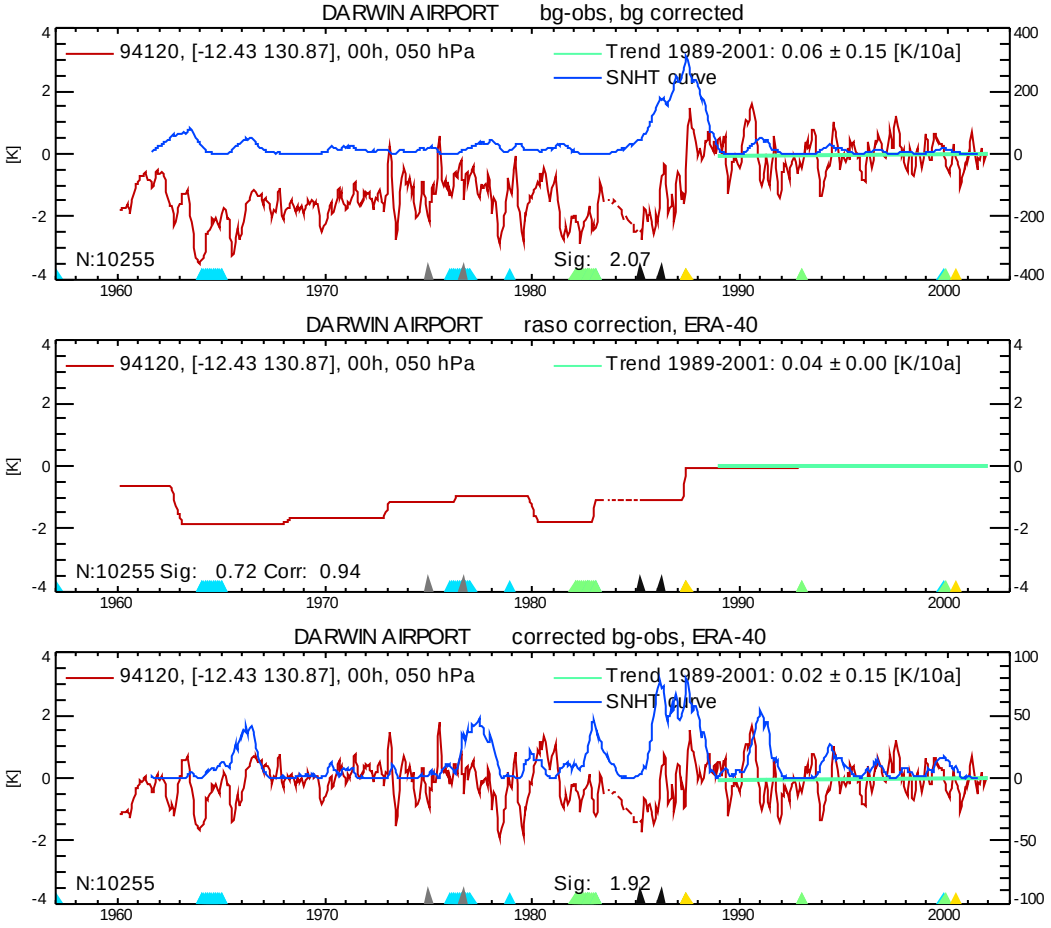

Figure 51: Unadjusted bg-obs time series, adjustment and adjusted bg-obs time series for radiosonde station Darwin.

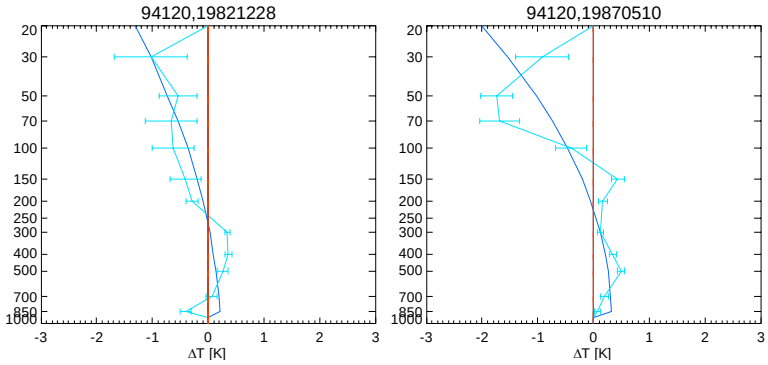

Figure 52: Adjustment profiles at two largest breakpoints at Darwin. In 1982 the radiation correction has changed according to CARDS metadata, from 1987 onwards Vaisala radiosondes have been in use instead of Philips MKIII radiosondes.

The last example shows time series and break profiles for radiosonde station Darwin (Australia). Major event at this site was the switch from Philips MKIII to Vaisala RS80 radiosondes in May 1987. The change is clearly indicated in the time series of Figure 51. A breakpoint is found at the right location but the adjustment applied at the 50 hPa level seems too small. Figure 52 shows, that the unsmoothed break profile suggests a larger correction at 50 hPa. In this case the vertical smoothing applied by RAOBCORE is probably too strong.

Despite some shortcomings that become obvious in all these examples, one can see that the automatic correction algorithm can cope with a variety of different situations and that the corrections applied make physical sense.

5.7 Robustness of the adjustment results – some sensitivity experiments

The RAOBCORE dataset presented here is the result of many decisions depending on thresholds, sampling strategies and interpolations. The optimal choice of these parameters is not well known and they may be varied within sensible bounds. In this section the results of two sensitivity experiments are discussed to give some hints how robust the adjustment results are with respect to the correction of the global mean bg and with respect to the use of metadata.

5.7.1. Adjustment of the bg time series

As stated above, the bg temperature time series contains breaks in the global mean that need to be removed before detecting breaks in radiosonde time series. This section shows results of an experiment where the bg has *not* been corrected. Without a corrected bg, the adjustment results deteriorate in almost every respect. Firstly, the number of breaks detected increases from ~4600 to 4840. Despite the higher number of detected breaks the trend heterogeneity *J* of the corrected trends increases slightly (see Figure 53 and Figure 54), suggesting that false break detections have occurred.

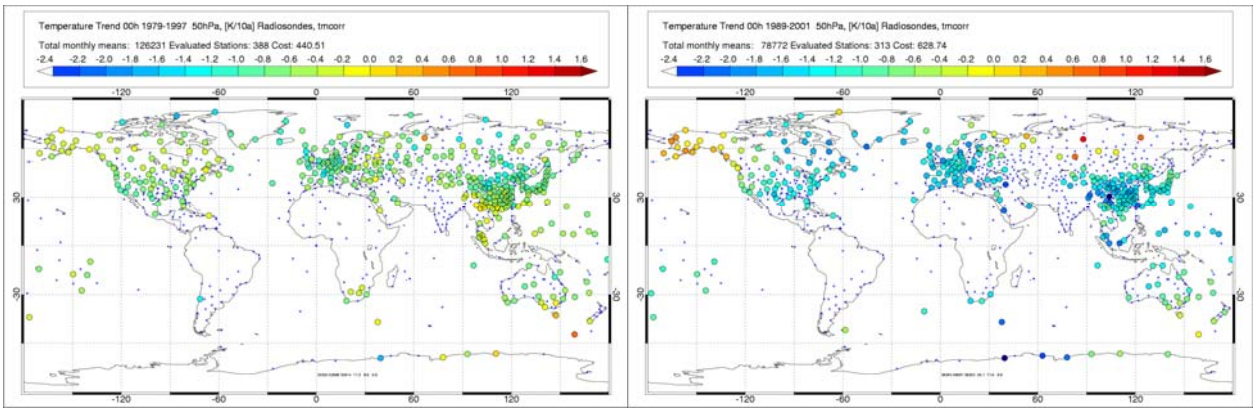

Figure 53: Temperature trends 1979-1997 and 1989-2001 in 50 hPa, corrected with RAOBCORE, but without adjusting the bg with its global mean bg-obs difference. Compare with Figure 27

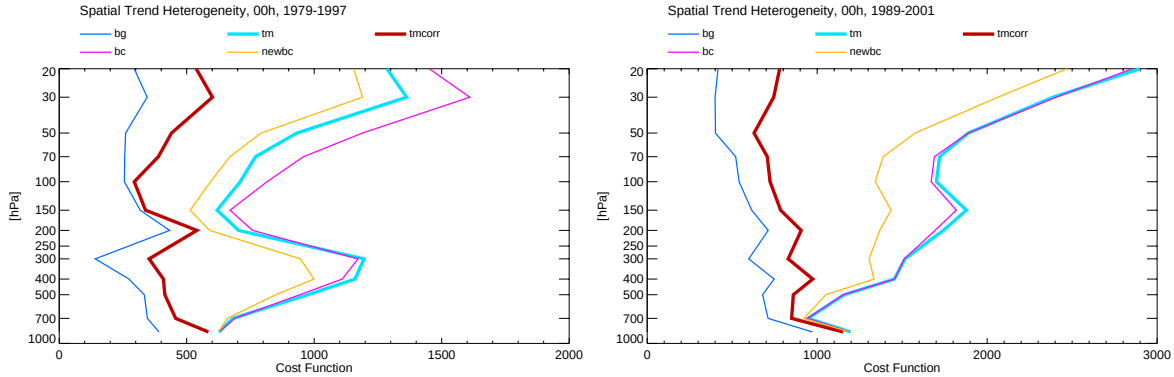

Figure 54: Temperature trend cost profile 1979-1997 and 1989-2001 in 50 hPa, corrected with RAOBCORE, but without adjusting the bg with its global mean bg-obs difference. Compare with Figure 29.

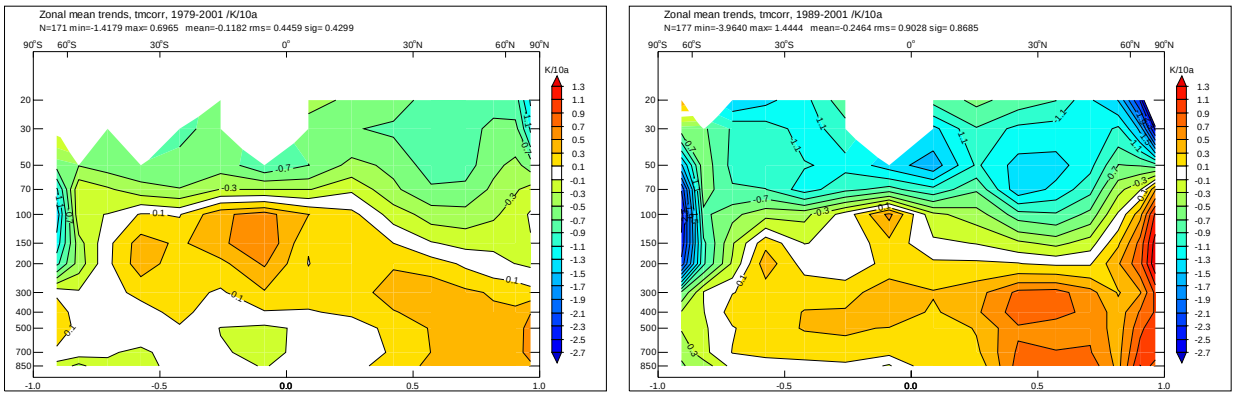

Figure 55: Zonal mean temperature trends 1979-2001 and 1989-2001 corrected with RAOBCORE, but without adjusting the bg with its global mean bg-obs difference. Compare with Figure 30 and Figure 31.

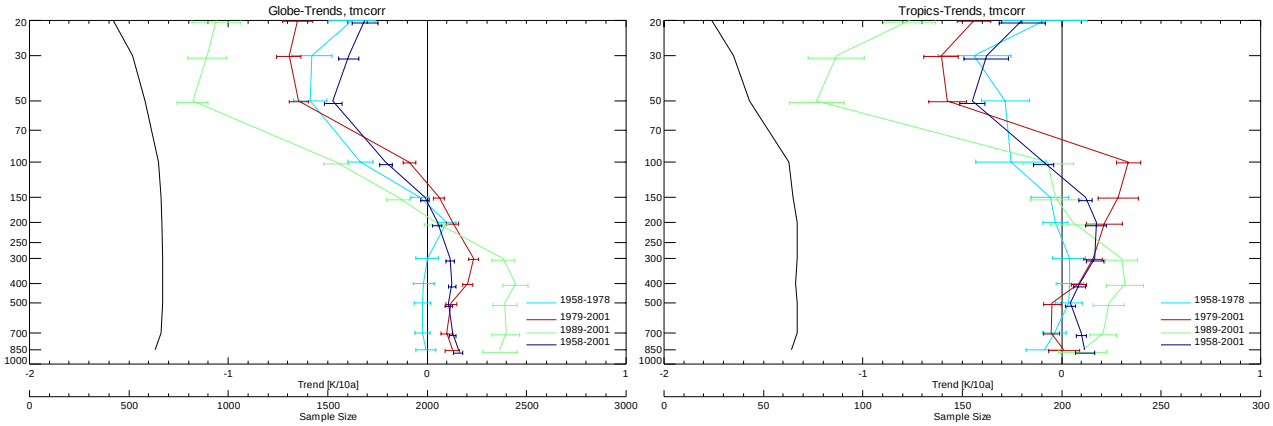

Figure 56: Mean radiosonde temperature trend profile for the globe (left) and the tropics (right), corrected with RAOBCORE, but without adjustment of the bg with its global mean bg-obs difference. Compare with Figure 32 and Figure 33.

The second problem, which is more serious, is that the regional and global mean trends of the corrected radiosondes are drawn to the bg trends. These trends are not trustworthy most notably in the tropics from 1979 onwards when spuriously large heating by condensation occurred, especially in the 1990s. The strong tropospheric warming in Figure 55 and Figure 56 is not visible or to a much lesser extent visible in the plots shown in Figure 30-Figure 33 where the bg has been adjusted before applying RAOBCORE. This artificial trend is introduced if a break of a radiosonde time series is adjusted with a reference series that has a spurious warming trend. In that case the temperature estimate of the bg after the break is too high and the temperature estimate before the break is too low. Shifts in the radiosonde time series towards cooler temperatures are overestimated and shifts towards warmer temperatures are underestimated in this case. Further it is more likely to detect shifts toward cooler temperatures than shifts toward warmer temperatures. Since most radiosondes have at least one break in the period 1979-2001 this effect occurs almost everywhere.

5.7.2. Use of metadata for break detection

RAOBCORE is capable of using metadata information for break detection. It is an interesting question how the metadata influence the adjustment results. The results of two experiments, one without use of metadata and one where adjustments are only applied in case of metadata events, are discussed here.

As can be seen from Figure 57, RAOBCORE has detected fewer breaks than in the standard configuration (shown in Figure 22) in both experiments. Figure 58 and Figure 59 show that the spatial trend heterogeneity for the late period has deteriorated in both experiments as well. This suggests that both sources of information are useful for break detection. The results deteriorate more if only metadata are used. Many undocumented breaks are not corrected in this case. The negative impact is less pronounced if metadata are missing. This suggests that the statistical break detection method used in RAOBCORE is capable of detecting most breaks even if metadata are missing. The metadata issue has little impact on the zonal mean trends shown in Fig. 59. The most likely explanation for this result is that in most cases the largest breaks are not only well documented but can also be reliably detected by the statistical method.

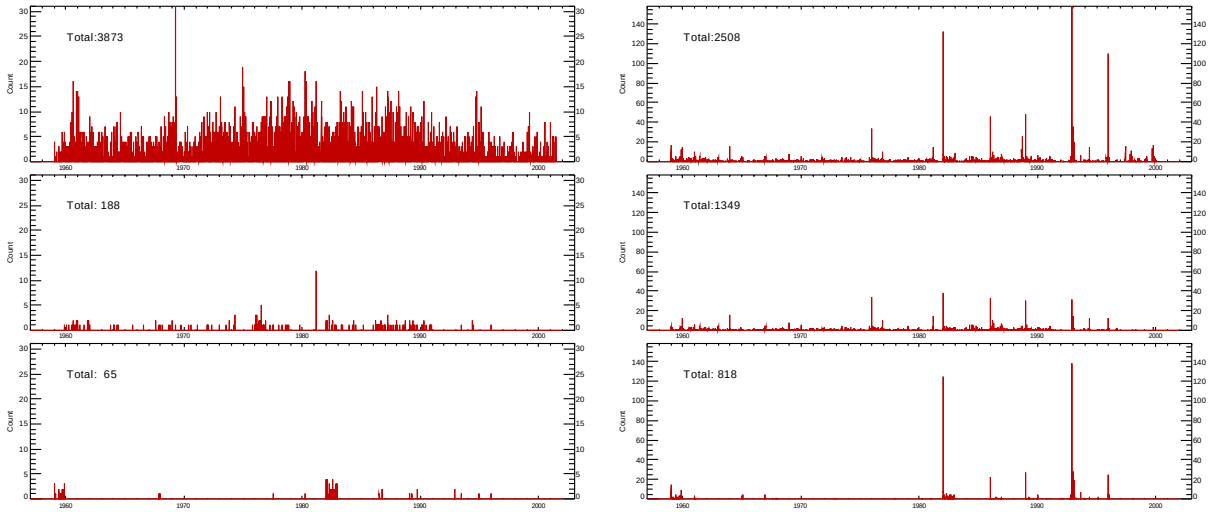

Figure 57: Break detection statistics from RAOBCORE using no metadata but information from the SNHT (left panel) and from RAOBCORE using metadata information only.

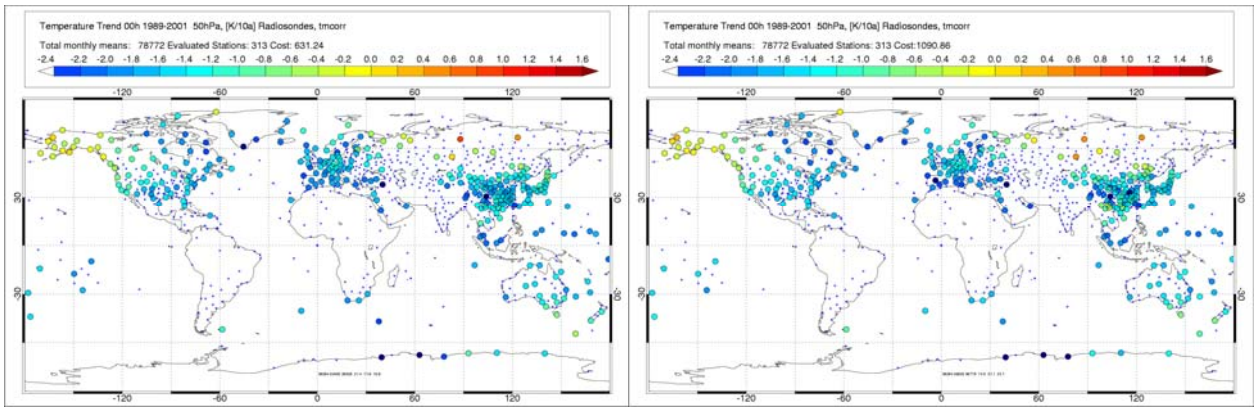

Figure 58: Radiosonde temperature trends in 50 hPa for period 1989-2001, corrected with RAOBCORE using no metadata but information from the SNHT(left panel) and with RAOBCORE using metadata information only (right panel).

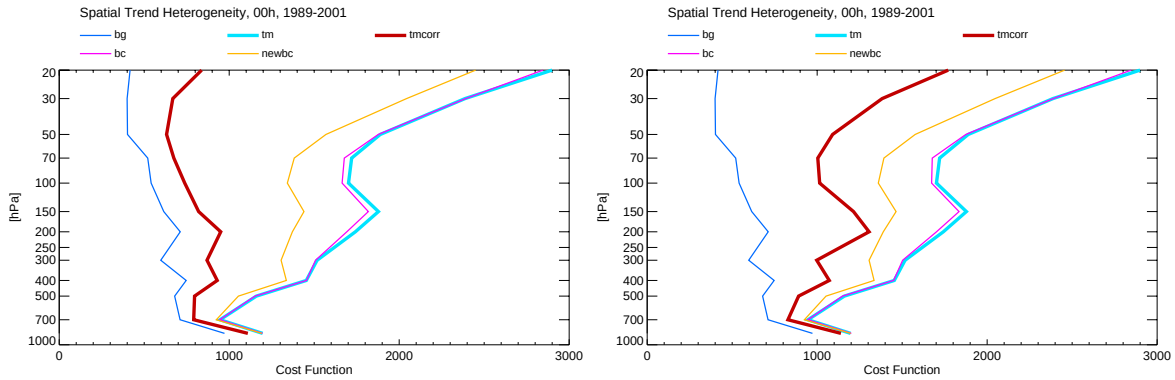

Figure 59: Trend cost profiles for radiosonde temperatures corrected with RAOBCORE using no metadata but information from the SNHT(left panel) and with RAOBCORE using metadata information only (right panel).

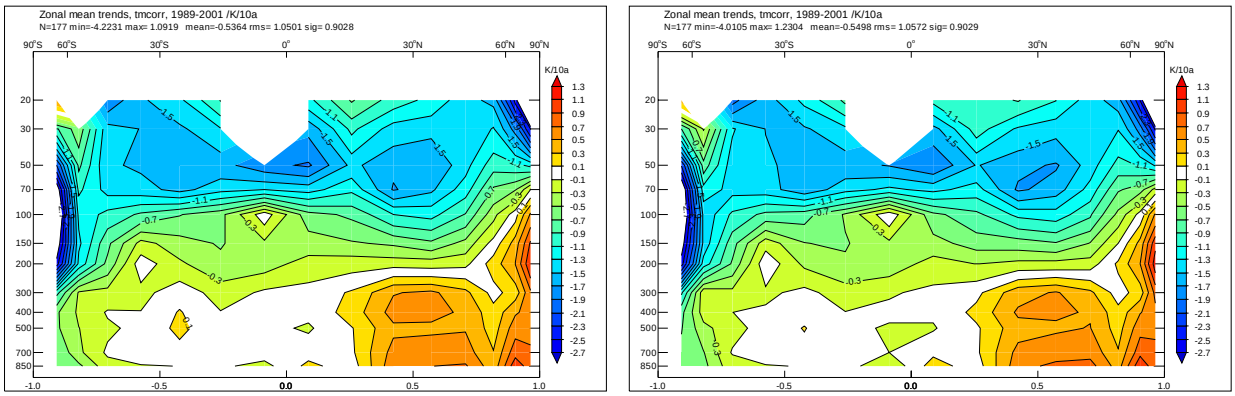

Figure 60: Zonal mean trends 1989-2001 from radiosonde temperatures corrected with RAOBCORE using no metadata but information from the SNHT(left panel) and with RAOBCORE using metadata information only (right panel).

One can think of many more sensitivity experiments. One could use different levels for calculating tropospheric and stratospheric means, one could use nonparametric tests for break detection and robust estimators for calculating the adjustments, one could change the vertical smoothing method of the break profiles etc. One could use subsampling methods such as bootstrapping for calculating uncertainty estimates. These experiments are left for future study.

5.8 Adjustment of the merged IGRA+ERA-40 radiosonde dataset

As noted in the introduction, a combined dataset that contains the union of the ERA-40 and IGRA radiosonde datasets has been created. This section gives a brief glimpse on the results achieved with the merged IGRA+ERA-40 radiosonde dataset. The merged dataset could have been used throughout this report. This may have led to some confusion, however, because of the different types of quality control used in ERA-40 and IGRA.

Figure 61 shows uncorrected and corrected trends at 50 hPa at 00GMT from the merged dataset. The merged dataset contains about 10% more stations for which trends could be calculated (compared with Figure 23). Some radiosonde data of the mid-1990s especially from the US are missing in the ERA-40 dataset but are available in the merged dataset. The performance of the RAOBCORE algorithm in reducing the spatial trend homogeneity is not much different if applied to the merged dataset, as can be seen from Figure 62. The outliers remaining in the merged dataset due to less stringent quality control in the IGRA dataset do not seem to have an adverse effect on the performance.

Figure 63 illustrates how the different quality control in ERA-40 and IGRA affects the data availability. The radiosonde station in Tamanrasset located in the African Sahel had a grossly positive bias in the 1960s, particularly between 1969 and 1971. As in Trappes shown above, MESURAL radiosondes have been in use at that time. Most of the 50 hPa data have been rejected by the ERA-40 quality control algorithm while they are retained in IGRA and therefore also in the merged dataset. In the 1990s three smaller spikes are evident in the merged bg-obs series which cannot be found in the ERA-40 bg-obs series. The good satellite data coverage during this period allowed the ERA-40 quality control system to identify these relatively small spikes as erroneous.

Figure 63 shows that many of the data rejected by the ERA-40 quality control system are not outliers, but are simply strongly biased data which do contain useful information if they are adjusted. It will be interesting to see in future data assimilation experiments if the number of rejected radiosonde data is significantly reduced if they are bias-corrected with RAOBCORE before the assimilation process.

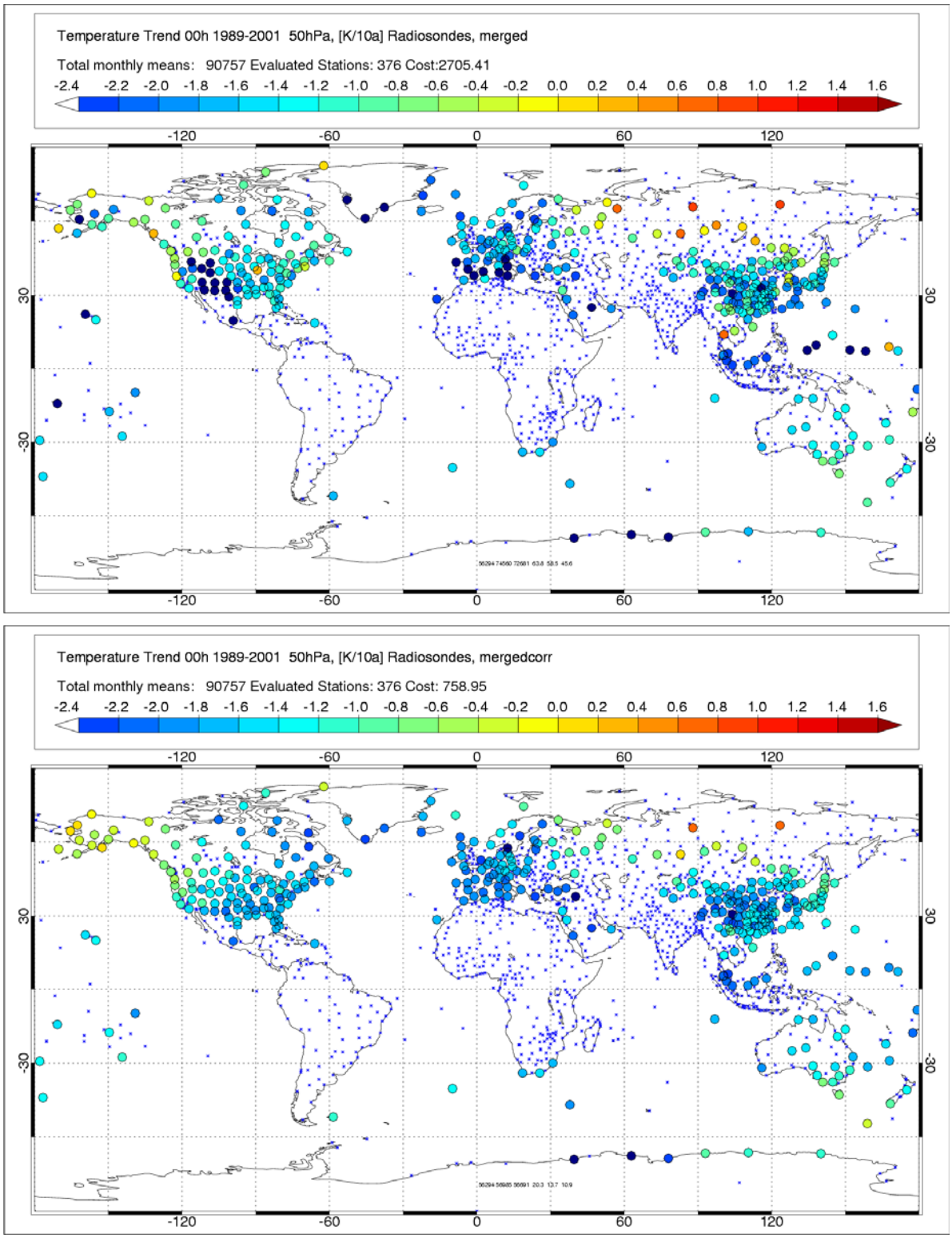

Figure 61: Uncorrected and corrected trends 1989-2001 at 50 hPa at 00GMT from the merged IGRA+ERA-40 dataset. About 10% more stations could be evaluated compared to Figure 23. Note increased number of stations over the US.

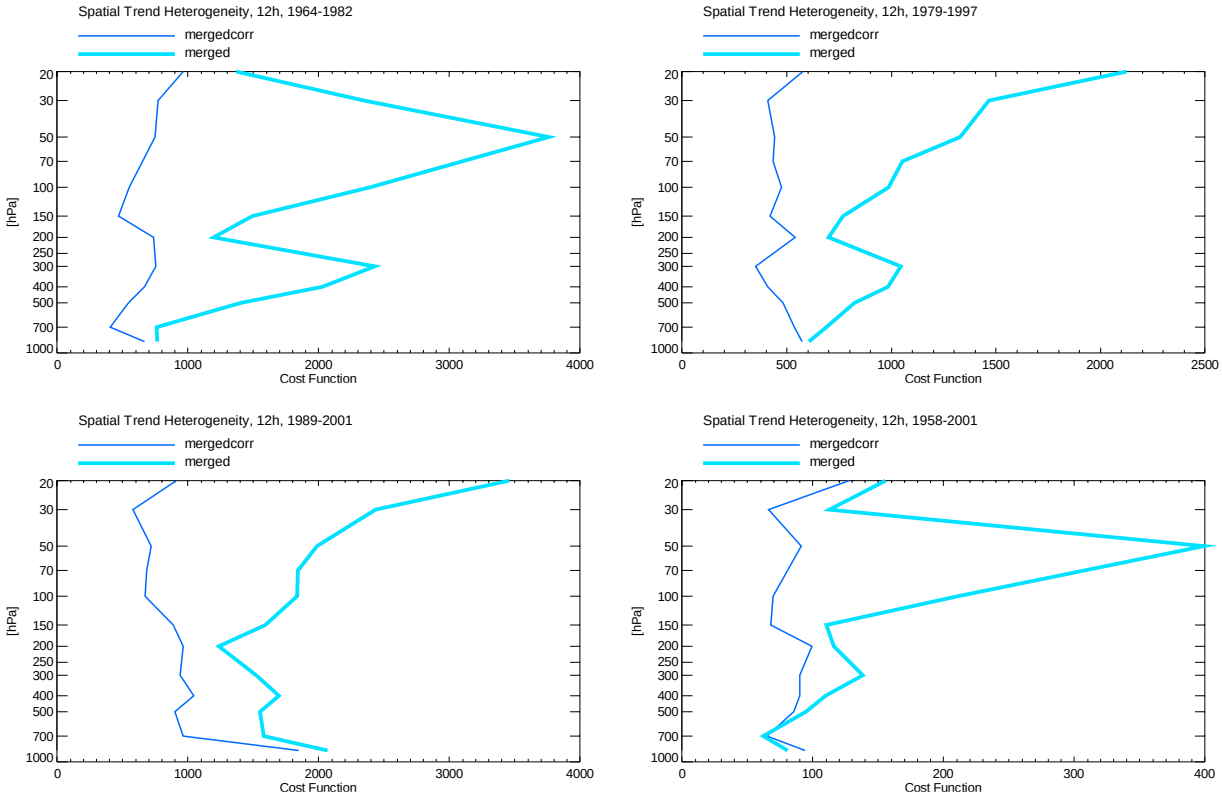

Figure 62: Trend costs at 12GMT for the merged radiosonde dataset for four different time periods 1964-1982, 1979-1997, 1989-2001, 1958-2001. "merged" corresponds to uncorrected merged dataset, "mergedcorr" to merged dataset corrected with RAOBCORE. Profiles are similar to those in Figure 29 ("tm","tmcorr") although they are not strictly comparable since the merged dataset contains more stations.

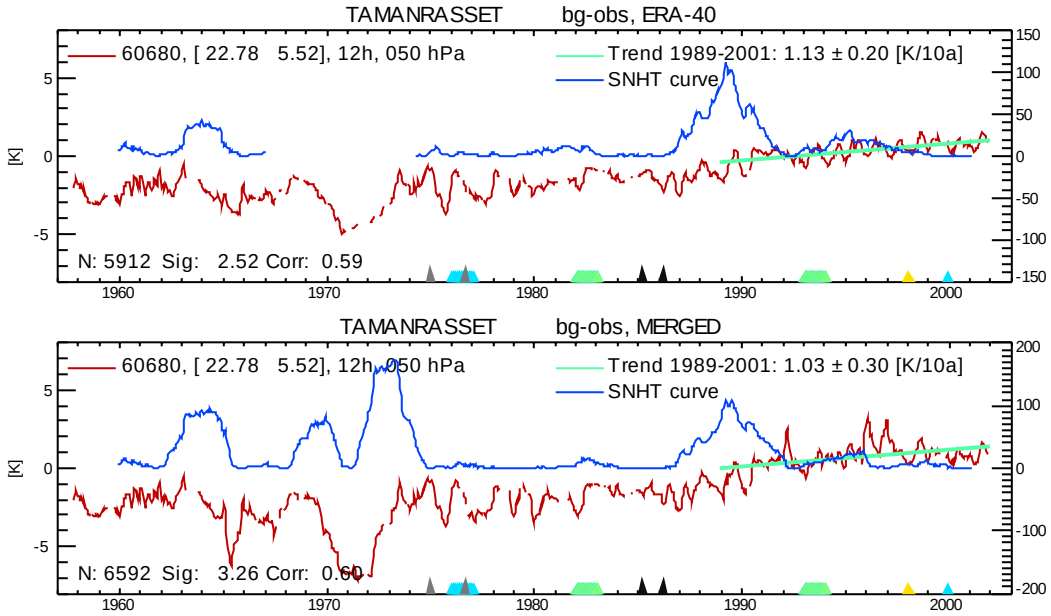

Figure 63: bg-obs difference at GUAN station Tamanrasset (60680) a) from ERA-40 data only and b) from the merged dataset. Note different temperature scale [-8K,8K] instead of [-4K,4K].

The examples given here show that the radiosonde data collection efforts must continue since both IGRA and ERA-40 have large data gaps and a merge of the data significantly improves data availability. Much can be learned from offline calculation of bg-obs difference time series. They are more suitable for data quality control than simple climatological checks and they allow finding out whether rejected data are real outliers or are strongly biased but correctable data. Further it will be important for a future reanalysis to analyze post-ERA-40 (2002-) bg-obs time series from operational assimilations before the radiosonde data are used in these reanalyses.

5.9 Effect of the radiosonde bias adjustment on analyses

Only little time was left towards the end of the year of the Marie Curie Fellowship for performing assimilation experiments with the IFS. Two experiments have been performed with IFS cycle 28R4. The satellite data used in ERA-40 with static bias adjustment have been used in both experiments. In the reference experiment (438) the radiosondes were left uncorrected, in the second experiment (436) radiosondes were corrected with RAOBCORE. Figure 64 shows the temperature difference 436-438 at the 50 hPa level at 00 GMT for the month September 1986 between the two analyses. The difference is negative especially over eastern Australia since the warm bias evident in the Australian radiosondes of this time has been reduced. The maximum difference of about 0.8K is shifted to the east since the information is transported downstream. There are also spots with negative differences over Antarctica. These are collocated with the three Australian radiosonde stations 89564 (Mawson), 89571 (Davis) 89611 (Casey) and with the Japanese station 89532 (Syowa) which changed the radiosonde type in 1987 as well. A tendency towards negative differences is also notable in the tropical western Pacific where island stations operated by Australia also used Philips MKIII radiosondes before 1988.

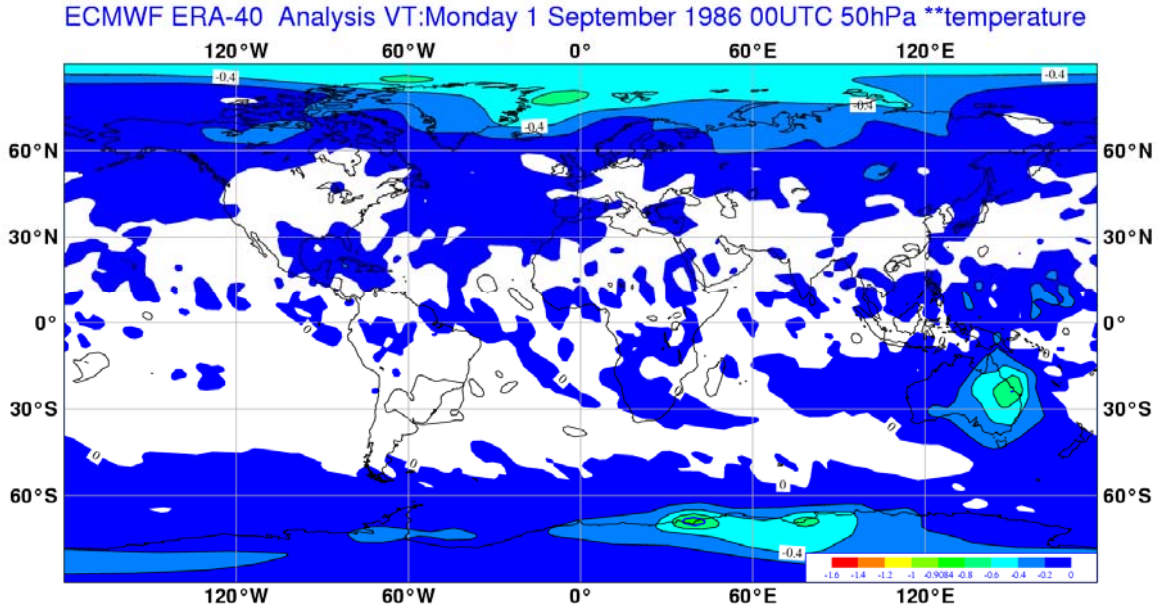

Figure 64: Effect of radiosonde bias adjustment on monthly mean 50 hPa analyses at 00GMT in September 1986. Figure shows difference between analyses without radiosonde bias adjustment and analyses using RAOBCORE-adjusted radiosonde data. Note effect over Australia and Antarctica due to adjustment of Philips MKIII radiosondes.

The negative differences over the north polar regions come from the adjustment of some of the most northerly radiosonde stations (see Figure 35) which exhibit very strong stratospheric cooling trends especially in the period 1979-1997 (see Figure 26 and Figure 27). Further analysis of this region is required where both the operating conditions for the radiosondes are extreme and the model biases of the ERA-40 assimilating model are large.

6. Conclusions and outlook

In this report temperature time series from the global radiosonde network as available in the ERA-40 (Uppala et al. 2005) analysis feedback dataset and in the IGRA (Durre et al. 2005) database have been investigated in order to detect and correct artificial breaks due to changes in observation practice. These breaks affect the suitability of radiosonde temperature time series for climate change studies as well as for future reanalyses and therefore they must be removed as far as possible.

A fully automatic detection and correction algorithm (RAOBCORE) based on comparison of the observed time series with time series of 6h background forecasts as calculated by the ERA-40 data assimilation system has been developed.

Layer mean bg-obs and obs(12GMT)-obs(00GMT) differences have been analyzed using a modified form of the Standard Normal Homogeneity Test (SNHT, Alexandersson and Moberg, 1997) with an analysis window of 6 years or less. It has been modified such that equal sampling of the annual cycle of bg-obs differences before and after a break is ensured. These measures have proven necessary in order to reduce false detections of breaks.

The SNHT yields time series of break probabilities that can be combined in principle with a priori probabilities coming from knowledge of individual station histories using Bayes' theorem. In practice the a priori probabilities are not known so that the probabilities are not well determined. Nevertheless the Bayes formula can be used to define a score, whose maxima are regarded as breakpoints. The time series are adjusted at the breakpoints by adding the vertically smoothed difference of the averages before and after the breakpoints to the older parts of the time series. The averages are again calculated such that the annual cycle is sampled equally before and after the break.

The trends from series corrected with RAOBCORE are spatially much more homogeneous than the trends from the uncorrected series for four different time periods (1964-1982, 1979-1997, 1989-2001, 1958-2001) investigated. The same is true for bg-obs differences at different time periods and for obs(12GMT)-obs(00GMT) differences as well. In general the results are best for the most recent periods, particularly from 1989 onwards.

The adjustments lead to spatially more smooth trend patterns than the adjustment algorithm used in ERA-40, which had to determine the adjustment without using long time series of bg-obs differences, since these differences were not available during the production of the ERA-40 dataset (Andrae et al. 2004). The corrected radiosonde time series can be recommended especially for 1989 onwards. The adjustments for the 1980s also seem reliable enough so that the corrected dataset can be used for contributing to controversial issues in climate change detection, e.g. about the vertical structure of temperature change in the atmosphere (Seidel et al., 2004; Fu et al. 2004). The adjustments for earlier parts of the records look promising but need more thorough investigation before they can be recommended as input to future reanalyses.

The satellite bias correction algorithm used in ERA-40 (Harris and Kelly, 2001) worked better than the algorithm used in ERA-15. Nevertheless some shifts in the global mean bg time series remained (especially 1975, 1976 and 1986) and it has been necessary to adjust the bg series by its weighted global mean bg-obs

difference before applying RAOBCORE. This was possible since the spatial pattern of these satellite-induced shifts is quite smooth.

The homogeneity adjustments only remove spurious breaks in a time series. The resulting time series is unbiased only if the most recent part of the time series is unbiased. If this is not the case, the whole time series still has a constant bias. These constant biases are clearly visible in the spatial patterns of the bg-obs differences in the period 1999-2001. Estimating these biases is a separate and difficult problem since one needs to know the absolute biases of the measurements. In this report time series that have been obviously biased even in their most recent part have been adjusted to the ERA-40 bg temperatures. This can be only an intermediate solution since the ERA-40 bg itself has a bias. In a later stage these biases must be assessed by intercomparison with homogenized neighboring reference radiosonde temperature time series.

The corrections found with RAOBCORE are available in ASCII format and can be used for assimilation experiments with IFS. There has been little time yet to assess the effect of the new radiosonde bias adjustment on analyses performed with an up to date data assimilation system using a recent version of the IFS and 4D-VAR. Data assimilation experiments that address this issue are planned for the near future.

Little time was left also for comparison of the adjustments presented here with those from Lanzante et al. (2003a,b) and Thorne et al. (2005). Especially the intercomparison with data from Thorne et al. (2005) will be important since both the dataset presented here and their dataset contain corrections which are motivated by the time series analysis alone but are not verifiable from metadata. If both methods, which contain quite different assumptions, yield similar adjustments, this will greatly enhance the credibility of both methods. Detailed comparisons with independent homogenization efforts based on detailed error assessments of the used sensors, which are available at selected well documented radiosonde stations, must follow as well.

Apart from the results from the correction algorithm this report has documented gaps in both the ERA-40 and the IGRA radiosonde datasets. The union of both datasets is about 5-10% larger than the individual datasets. Efforts to create a comprehensive radiosonde dataset must therefore continue. In any case one should try to calculate bg-obs differences from ERA-40 or from future reanalysis projects for quality control of historical data that have not been assimilated (yet). Experience with radiosonde temperature data in this report has shown that many more erroneous data can be identified with bg-obs differences than with simple climatological tests since the background fields from an advanced data assimilation system are much more accurate. It has also been demonstrated that it is possible to generate accurate background minus observation (bg-obs) time series for IGRA data.

This report has documented the first serious attempt to use analysis feedback data for automatic temporal homogenization of climate obsevation records. While the developed algorithm still has shortcomings, which need to be addressed, the potential of this approach is now evident. It seems natural to apply this adjustment method to other upper air parameters as well, especially to wind data.

7. Acknowledgements

The research for this report has been funded by EU-contract “RASOHOM” (Contract No. MEIF-CT-2003-503976**)** of the European Commission. I thank Adrian Simmons and Sakari Uppala for their continuous advice and encouragement. I would like to thank also Dian Seidel for hosting a mini-workshop in March 2004 that gave me much useful advice and introduced me to many members of the upper air climate research community, and Peter Thorne as well as John Lanzante for providing databases of their adjustments. NCDC provided the IGRA dataset through its web site.

8. References

Aguilar, E, 2001: The Upper Air Stations History. A brief Description. Internal Report, available from NCDC.

Aguilar, E., I. Auer, M. Brunet, T. C. Peterson and J. Wieringa, 2004: Guidance on Metadata and Homogenization. WMO/TD No. 1186, 53pp. WMO, Geneva.

Andrae, U., N. Sokka and K. Onogi, 2004: The radiosonde temperature bias correction in ERA-40. ERA-40 project report series 15, 35pp.

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

Cardinali, C., S. Pezzulli and E. Andersson, 2004: Influence matrix diagnostic of a data assimilation system. *Q. J. R. Meteorol. Soc.* 130, 2767-2786.

Courtier, P., Andersson, E., Heckley, W., Pailleux, J., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F. and Fisher, M., 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). I: Formulation. *Q. J. R. Meteorol. Soc.*, 124, 1783-1807.

Dee, D., 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, pp. 65-74, available from ECMWF.

DeGroot, M., 1986: Probability and Statistics. Second edition, Addison Wesley, 723pp.

Ducre-Robitaille, J. F., L. Vincent and D. Boulet, 2003: Comparison of Techniques for Detection of Discontinuities in Temperature Series. *Int. J. Climatology* 23, 1087-1001

Durre, I. , R. S. Vose and D. B. Wuertz, 2005: Overview over the Integrated Global Radiosonde Archive. submitted to *J. Climate*

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

ECMWF, 2003: IFS documentation, Cycle 23R4, available online via http://www.ecmwf.int/research/ifsdocs/

Eskridge, R. E., O.A. Alduchov, I.V. Chrnynkh, P. Zhai, A.C. Polansky, S.R. Doty, 1995: A comprehensive aerological reference datasets (CARDS): Rough and systematic errors. *Bull. Amer. Meteor. Soc*. 76, 1759 1775.

Free, M., et al., 2002: Creating climate reference datasets. *Bull. Amer. Meteorol. Soc.*, 89, 891- 899.

Fu, Q., C. M. Johanson, S. G. Warren and D. J. Seidel, 2004: Contribution of stratospheric cooling to satellite-inferred tropospheric temperature trends. *Nature, 429, 55-58.*

Haimberger, L., 2004: Checking the temporal homogeneity of radiosonde data in the Alpine region using ERA-40 analysis feedback data. *Meteorol. Z.* 13, 123-129.

Harris, B. A. and G. Kelly, 2001: A satellite-bias correction scheme for data assimilation. *Q. J. Roy. Met. Soc.*, 127, 1453-1468.

Hernandez, A., G. Kelly and S. Uppala, 2004: The TOVS/ATOVS observing system in ERA-40. ERA-40 Project Report Series No. 16, available from 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

Kelly, G. and X. Li, 2001: Assimilation of TOVS/VTPR/SSMI radiances and use of Australian surface PAOBS. In: Proceedings of Workshop on Reanalyses, 123-148, available from ECMWF.

Lanzante, J.R., S.A. Klein, and D.J. Seidel, 2002a: 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, 2002b: Temporal homogenization of monthly radiosonde temperature data. Part II: Trends, sensitivities, and MSU comparison. *J. Climate,* 16, 241-262.

Caussinus, H. and O. Mestre, 2004: Detection and correction of artificial shifts in climate series. *Appl. Statist.* 53, 405-425.

Nash, J. and F. J. Schmidlin, 1987: WMO International Radiosonde Comparison: Final Report, WMO/TD-No. 195, WMO, Geneva. 103 pp.

Onogi, K., 2000: The long term performance of the radiosonde observing system to be used in ERA-40. *ERA-40 Project Report Series* 2, 77pp.

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

Peterson, T.C., D.R. Easterling, T.R. Karl, P. Groisman, I. Auer, R. Böhm, N. Plummer, N. Nicholis, S. Torok, L. Vincent, H. Tuomenvirta, J. Salinger, E.J. Forland, I. Hanssen-Bauer, H. Alexandersson, P. Jones, and D. Parker, 1998: Homogeneity adjustments of in situ climate data: A review. *Int. J. Climatology*, 18, 1493-1517.

Randel, W., P. Udelhofen, E. Fleming, M. Geller, M. Gelman, K. Hamilton, D. Karoly, D. Ortland, S. Pawson, R. Swinbank, F. Wu, M. Baldwin, M-L. Chanin, P. Keckhut, K. Labitzke, E. Remsberg, A. Simmons and D. Wu, 2003: The SPARC Intercomparison of Middle-Atmosphere Climatologies, *J. Climate* 17, 986-1003.

Richner, H. and P. D. Philips, 1982: The Radiosonde Intercomparison SONDEX. Contributions to Current Research in Geophysics, 11. Birkhaeuser, 1198 pp.

ROSHYDROMET and Finnish Meteorological Institute, 2001: Joint report of the RF95-NW project, 20pp, available from Finnish Meteorological Institute.

Santer, B. D., Wigley, T. M. L., A. J. Simmons, P. W. Kallberg, G. A. Kelly, S. M. Uppala, C. Amann, J. S. Boyle, W. Brüggemann, C. Doutrinaux, M. Fiorino, C. Mears, G. A. Meehl, R. Sausen, K. E. Taylor, W. M. Washington, M. F. Wehner and F. J. Wentz, 2004: Identification of anthropogenic climate change using a second generation reanalysis. *J. Geophys. Res*. 101D, 21104.

Seidel, D. J., J. K. Angell, J. Christy, M. Free, S. A. Klein, J. R. Lanzante, C. Mears, D. Parker, M. Schnabel, R. Spencer, A. Sterin, P. Thorne and F. Wentz, 2004: Uncertainty in Signals of Large-Scale Climate Variations in Radiosonde and Satellite Upper Air Temperature Datasets. *J. Climate* 17, 2225-2240.

Simmons, A.J., 2003: Development of the ERA-40 data assimilation system. In: Proceedings of the Workshop on Reanalysis 5-9 Nov. 2001. ECMWF, 11-30.

Thorne, P., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman P. Brohan and J. R. Knight, 2005: Revisiting radiosonde upper-air temperatures from 1958-2002. submitted to *J. Geophys. Res*.

Uppala, S., 2003: ECMWF Re-Analysis, 1957-2001, ERA- 40. In : Proceedings of the Workshop on Reanalysis 5-9 Nov. 2001. ECMWF, 1-10.

Uppala, S. M., P. W. Kallberg, A. J. Simmons, U. Andrae, V. da Costa Bechthold, M. Fiorino, J. K. Gibson, J. Haseler, A. Hernandez, G. A. Kelly, X. Li, K. Onogi, S. Saarinen, N. Sokka, R. P. Allan, E. Andersson, K. Arpe, M. A. Balmaseda, A. C. M. Beljaars, L. van den Berg, J. Bidlot, N. Bormann, S. Caires, F. Chevallier, A. Dethof, M. Dragosavac, M. Fisher, M. Fuentes, S. Hagemann, E. Holm, B. J. Hoskins, L. Isaksen, P.A.E.M. Janssen, R. Jenne, A. P. McNally, J.-F. Mahfouf, J.-J. Morcrette, N. A. Rayner, R. W. Saunders, P. Simon, A. Sterl, K. Trenberth, A. Untch, D. Vasiljevic P. Viterbo and J. Woollen, 2005: The ERA-40 Re-Analysis. *Q. J. Roy. Meteorol. Soc. (To be published)*

Vincent, L., 1998: A Technique for the Identification of Inhomogeneities in Canadian Temperature Series. *J. Climate* 11, 1094-1104.