INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **18**: 1493–1517 (1998)

HOMOGENEITY ADJUSTMENTS OF *IN SITU* ATMOSPHERIC CLIMATE DATA: A REVIEW

THOMAS C. PETERSON[a,*], DAVID R. EASTERLING[a], THOMAS R. KARL[a], PAVEL GROISMAN[a], NEVILLE NICHOLLS[b], NEIL PLUMMER[b], SIMON TOROK[c], INGEBORG AUER[d], REINHARD BOEHM[d], DONALD GULLETT[e], LUCIE VINCENT[e], RAINO HEINO[f], HEIKKI TUOMENVIRTA[f], OLIVIER MESTRE[g], TAMÁS SZENTIMREY[h], JAMES SALINGER[i], EIRIK J. FØRLAND[j], INGER HANSSEN-BAUER[j], HANS ALEXANDERSSON[k], PHILIP JONES[l] and DAVID PARKER[m]

[a] *National Climatic Data Center, NOAA, 151 Patton Avenue, Asheville, NC 28801, USA*
[b] *Bureau of Meteorology, Melbourne, Victoria, Australia*
[c] *School of Earth Sciences, University of Melbourne, Melbourne, Victoria, Australia*
[d] *Central Institute of Meteorology and Geodynamics, Vienna, Austria*
[e] *Climate Research Branch, Environment Canada, Downsview, Ontario, Canada*
[f] *Finnish Meteorological Institute, Helsinki, Finland*
[g] *Meteo France, SCEM/CBD, Toulouse, France*
[h] *Hungarian Meteorological Service, Budapest, Hungary*
[i] *National Institute of Water and Atmospheric Research Ltd., Auckland, New Zealand*
[j] *Norwegian Meteorological Institute, Oslo, Norway*
[k] *Swedish Meteorological and Hydrological Institute, Norrköping, Sweden*
[l] *Climate Research Unit, University of East Anglia, Norwich, UK*
[m] *Hadley Centre, Meteorological Office, Berkshire, UK*

Received 17 July 1997
Revised 20 April 1998
Accepted 30 April 1998

ABSTRACT

Long-term *in situ* observations are widely used in a variety of climate analyses. Unfortunately, most decade- to century-scale time series of atmospheric data have been adversely impacted by inhomogeneities caused by, for example, changes in instrumentation, station moves, changes in the local environment such as urbanization, or the introduction of different observing practices like a new formula for calculating mean daily temperature or different observation times. If these inhomogeneities are not accounted for properly, the results of climate analyses using these data can be erroneous. Over the last decade, many climatologists have put a great deal of effort into developing techniques to identify inhomogeneities and adjust climatic time series to compensate for the biases produced by the inhomogeneities. It is important for users of homogeneity-adjusted data to understand how the data were adjusted and what impacts these adjustments are likely to make on their analyses. And it is important for developers of homogeneity-adjusted data sets to compare readily the different techniques most commonly used today. Therefore, this paper reviews the methods and techniques developed for homogeneity adjustments and describes many different approaches and philosophies involved in adjusting *in situ* climate data. © 1998 Royal Meteorological Society.

KEY WORDS: homogeneity; climate data; data adjustment techniques; metadata

1. INTRODUCTION

Climate data can provide a great deal of information about the atmospheric environment that impacts almost all aspects of human endeavour. For example. these data have been used to determine where to build homes by calculating the return periods of large floods, whether the length of the frost-free growing season in a region is increasing or decreasing, and the potential variability in demand for heating fuels. However, for these and other long-term climate analyses—particularly climate change analyses—to be accurate, the climate data used must be homogeneous. A homogeneous climate time series is defined as one where variations are caused only by variations in weather and climate (Conrad and Pollak, 1950).

* Correspondence to: National Climatic Data Center, NOAA, 151 Patton Avenue, Asheville, NC 28801, USA.

CCC 0899–8418/98/131493–25$17.50

Unfortunately, most long-term climatological time series have been affected by a number of non-climatic factors that make these data unrepresentative of the actual climate variation occurring over time. These factors include changes in: instruments, observing practices, station locations, formulae used to calculate means, and station environment (Jones *et al*., 1985; Karl and Williams, 1987; Gullett *et al*., 1990; Heino, 1994). Some changes cause sharp discontinuities while other changes, particularly change in the environment around the station, can cause gradual biases in the data. All of these inhomogeneities can bias a time series and lead to misinterpretations of the studied climate. It is important, therefore, to remove the inhomogeneities or at least determine the possible error they may cause.

The authors have put a great deal of effort into developing ways to identify non-climatic inhomogeneities and then to adjust the data to compensate for the biases these inhomogeneities produce. Several techniques have been developed to address a variety of factors that impact climate data homogenization such as the type of element (temperature vs. precipitation), spatial and temporal variability depending on the part of the world where the stations are located, length and completeness of the data, availability of metadata (data about data), and station density. Each team has developed a different philosophy regarding data adjustments since their requirements and missions have been quite different.

For example, the best technique for a dense regional rain gauge network in the humid extra-tropics might not work well for a sparse subtropical semi-arid network. Metadata in the form of station history documentation that details instrumentation, locations, observing practices, etc. may be digital or in paper archives or not available at all. What level of confidence is to be required before making an adjustment in time series? Numerically, one can use a 95 or 99% confidence level generated by a test statistic, but how would metadata weigh into such analysis? Some of these decisions are made based on the specific goal, for example, if one is trying to produce a homogeneous version of a far flung network a different approach might be required than if the goal was selecting and adjusting only the best stations from a dense network to produce a homogeneous subset. Some decisions are made based on years of experience with the specific data and metadata involved. But other decisions are, of necessity, based on the resources available: e.g. careful analysis of station history documentation can be very labour intensive.

Homogeneity-adjusted data sets of precipitation as well as land and marine air temperature are used by a wide variety of researchers worldwide. Users of these data should be aware of the differences and similarities between both data characteristics and adjustment techniques. Also, since the densest networks and sources of most meteorological metadata are usually country specific, the next big push in homogeneity adjustments will likely be the development of a large number of homogeneity-adjusted country data sets. Therefore, this article will not only review different homogeneity detection and adjustment techniques, but also explain the philosophy behind each of the approaches. The review starts with a description of direct methods, which rely on metadata and studies of the effects of specific changes in instrumentation, and indirect methods, which use a variety of statistical and graphical techniques to determine inhomogeneities. Next is a description of various ways these tools have been successfully combined. The article ends with a discussion which includes an assessment of a limited comparison of the results along with their implications.

2. DIRECT METHODOLOGIES FOR HOMOGENEITY TESTING

2.1. Use of metadata

Of all the homogeneity tools available, the most commonly used information comes from station history metadata files. Station moves, changes in instrumentation, problems with instrumentation, new formulae used to calculate mean temperature, changes in the nearby environment such as buildings and vegetation, new observers, changes in the time of observations, and documentation from comparison measurement studies that might have taken place when instruments changed are all relevant information in assessing homogeneity. These metadata can be found in station records, meteorological yearbooks, original observation forms, station inspection reports and correspondence, and various technical

WB Form 500-1
(4-61)

UNITED STATES DEPARTMENT OF COMMERCE
WEATHER BUREAU
STATION HISTORY

RENDITION: ( ) Original;
(X) Supplement No. 12

OFFICE PREPARING FORM WBAS, Spokane, WAS H.

STATION Spokane COUNTY Spokane STATE Washington INTERNATIONAL INDEX NUMBER 72785 DATE PREPARED 4-7-64

NUMBER OF LOCATION	LOCATION	TYPE OF STATION	AT THIS LOCATION		AIRLINE DISTANCE AND DIRECTION FROM PREVIOUS LOCATION	LATITUDE	LONGITUDE	ELEVATION ABOVE MEAN SEA LEVEL		
			FROM	TO				GROUND (H)	ASSIGNED STATION (H_p)	ACTUAL BAROMETER (H_z)
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
11a	Felts Field, Spokane	WBAS	5/14/32	1/1/41	4½ miles NE of city center	47° 40'	117° 20'	1955	1968	1967.65
11b	" "	"	1/1/41	1/1/42	"	"	"	"	"	"
11c	" "	"	1/1/42	2/3/42	"	"	"	"	"	"
11d	" "	"	2/3/42	2/20/42	"	"	"	"	"	"
11e	" "	"	2/20/42	12/8/47	"	"	"	"	"	"
12	Geiger Field, Spokane	"	12/8/47		6 miles SW of city center	47° 37'	117° 31'	2357	2365	2366.47

NUMBER OF LOCATION	ELEVATION ABOVE GROUND									REMARKS (Reason for move, changes in observation programs, effects of buildings or terrain, etc.)
	WIND INSTRUMENTS	EXTREME THERMOMETERS	PSYCHROMETER	TELEPSYCHROMETER*	RAIN GAGES: TIPPING BUCKET	RAIN GAGES: WEIGHING	RAIN GAGES: 8 INCH	(s)	(t)	
(a)	(l)	(m)	(n)	(o)	(p)	(q)	(r)			(u)
a	42	28	27				25			City Office records were official for climatological purposes thru this period.
11b	42	28	27		25	26	25			City Office and WBAS consolidated at Felts Field 1/1/41.
11c	42	7	6		25	26	25			Thermometers moved from roof to ground 1/1/42.
11d	42	7	6		3	4	3			Raingages moved from roof to ground 2/3/42.
11e	53	7	6		3	4	3			New wind mast raised anemometer height 2/20/42.
12	29	7	6		3	4	3			WBAS moved to Geiger Field 12/8/47. (See remarks)

REMARKS CONTINUED:
This rendition of Form 500-1 is to correct Felts Field entries on original rendition. See rendition supplement No. 11 12/1/63 for changes at Geiger Field location.

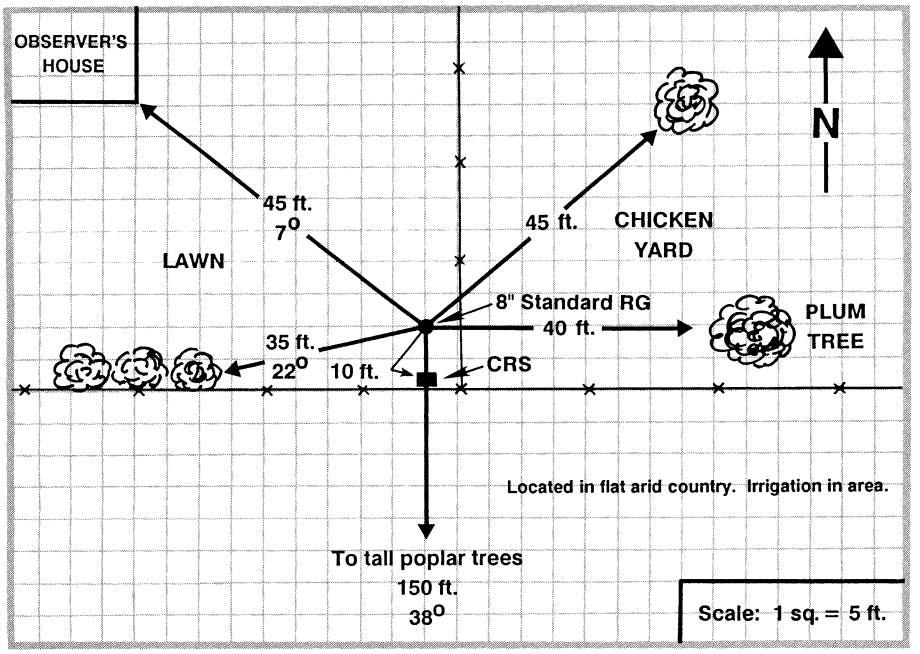
*Or hygrothermometer

Figure 1. One sample page of the station history file for Spokane, Washington describing some changes that can cause inhomogeneities in the climate data

manuscripts. Metadata can also be acquired from interviews with persons responsible for station operations. Figure 1 is an example of a station history form from the United States indicating station moves that might impact observed temperatures. An example of a different type of metadata is given in Figure 2.

The quality of station history metadata varies with time and the best station history data is not necessarily the most recent. In fact, the old descriptive reports with evaluations of station locations and the observers' skills by meteorologists are often more useful than modern, objective reports. However, one type of station metadata that has become more common in recent years is a photograph of the station. Examining a series of pictures may reveal changes not mentioned in the inspection reports. Another important type of metadata is information that applies to more than one station. It is crucial to know the changes that have affected a large fraction of the network. For example, the whole Finnish station network changed to a new type of precipitation gauge in 1909 and 1981 (Heino, 1994) and, according to a document published in 1904, all the thermometers in British East Africa were being moved from a variety of exposures to grass hut shelters (Peterson and Griffiths, 1996). These systematic types of changes may need to be addressed differently than changes involving only individual stations.

The advantages of metadata are that specific information contained is very relevant and it can provide the researcher with precise knowledge of when the discontinuity occurred and what caused it. Unfortunately, metadata are often not complete, missing or sometimes actually erroneous (e.g. when the author of the metadata completes the form from memory several years after the change occurred). Even with rather good Norwegian metadata information back to at least the turn of the century, Førland (1994) discovered important changes that are not reported in the proper way in the inspection reports. Sometimes there are also problems with interpreting the metadata appropriately (e.g. a change in latitude or longitude in 1880 may be more likely to be due to improved surveying than a station move).



EXPOSURE DESCRIPTION : CRS 180/10/- TREES 240-250/35/22 180/150/38 TREE 90/40/- 045/45/- HOUSE 310-330/45/7

Figure 2. Sketch of instrument exposure at Oakley, Idaho for the period June 28, 1949–January 9, 1973, as presented in the observer records and its description in the metadata. These metadata indicate the obstacles that could effect the wind over the orifice of the precipitation gauge (Groisman *et al.*, 1996)

The biggest problem with station history metadata is obtaining the relevant information for determining discontinuities in a useable form. Some sources of metadata contain large amounts of irrelevant information, making extracting the important pieces time consuming and tedious. However, a few countries or institutions have digitized their metadata for selected stations or selected time periods. While time consuming, this ultimately offers researchers access to station history information without the expensive burden of performing a station by station searches through paper archives. Once digital, these metadata have been accessed through spread sheets (e.g. Australia), relational data bases (e.g. United States), and for the North Atlantic Climatological Dataset (NACD) several European countries have agreed on a common structure for metadata files so the metadata can be shared between offices (Frich *et al.*, 1996).

2.2. Side by side comparisons of instruments

Instrument type changes are usually accompanied by comparison measurements. Ideally these are conducted at each station so there are overlapping time series between the old and new instrumentation, but often such comparisons are only performed at a limited number of locations. For example, the differences between shielded and unshielded precipitation gauges were studied at more than 20 sites all over Norway (Førland *et al.*, 1996). Though side by side comparisons should continue for at least a full year in order to be able to assess the seasonal variability in the differences between the instruments, a few of these comparisons have continued for many decades. For instance, temperature measurements in Stevenson Screens and Glaisher Stands were taken at Adelaide, Australia, for over 60 years (Nicholls *et al.*, 1996).

Results of instrument comparisons are often published as institute reports or other 'gray literature' not easily available to the international research community (Nordli *et al.*, 1997). There are also several locations, such as Valdai (Russia), Jokioinen (Finland), and Sterling (Virginia), where many instruments are kept in the same field to facilitate comparisons between sensors. Such sites and special intercomparison studies (e.g. Huovila *et al.*, 1988) provide valuable information that can be used to understand how the data are likely to change when the instrument changes from one type to another. However, in areas of large local climatic gradients, such as Norway and the western United States, it is difficult to extrapolate adjustment factors based on these results, because the biases in observations are dependent on the climate of wind (for rain gauges) or of solar radiation (for thermometer shelters).

2.3. Statistical studies of instrument changes

A clear example of a statistical study of the effect of an instrument change was done in response to a change from liquid-in-glass thermometers in wooden Cotton Region Shelters (CRS) to a new thermistor maximum–minimum temperature system (MMTS) at many U.S. stations. Quayle *et al.* (1991) used station metadata to determine which U.S. stations remained unchanged with CRS and which stations switched to MMTS as their only change (e.g. no major relocations at the same time). The five most highly correlated nearby CRS stations for each MMTS station were used to create a local CRS temperature time series for each MMTS station. After creating an MMTS − CRS time series for each MMTS station, the hundreds of MMTS − CRS time series in the study were averaged, centering each station's available data based on the month the MMTS was installed. The results smoothed out the individual station noise but preserved the average effect of the change in instrumentation (see Figure 3). Interestingly, examination of Figure 3 indicates that the step function move from pre- to post-MMTS took several months. This is attributed to inaccuracies in the metadata indicating the date of the change. Averaging many stations together like this can produce a clear signal to indicate the average magnitude of the discontinuity. However, this is just a regional average; the exact effect at individual stations may vary somewhat depending on local environmental or climatic factors such as the amount of direct sunlight on the shelter.

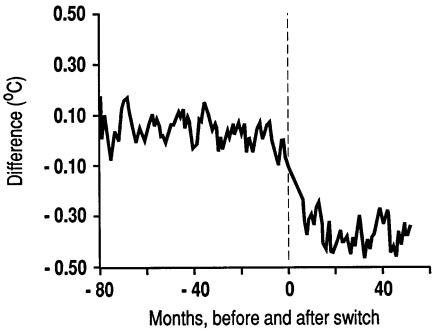


Figure 3. Aggregated differences in mean monthly maximum temperature between stations that switched to thermistor sensors at time 0 and neighbouring stations that continued to use liquid in-glass thermometers (from Quayle *et al.*, 1991)

3. INDIRECT METHODOLOGIES FOR HOMOGENEITY TESTING

3.1. Use of single station data

Station data are used in most homogeneity testing techniques but primarily in conjunction with metadata or comparisons with neighbouring stations. Using only data from an individual station is problematical because the change (or lack of change) one detects may be caused (or masked) by real changes in climate. However, there are some isolated stations without adequate neighbours where more reliance must be given to the individual station data alone. Additionally, station data can be used to date a change when metadata are imprecise. This is particularly true when multiple elements are available (e.g. a change in pressure is often better able to date a move than precipitation data).

Zurbenko *et al.* (1996) describes a filter that has been applied to single station data to date a discontinuity. Specifically, as the moving average filter approaches a region of the time series where there may be a discontinuity as indicated by increased variability or magnitude of the slope, the half-length of the moving average on the potential discontinuity side is smaller than the half-length on the other side of the data point. This process, which is iterative, can smooth out the noise of the time series while retaining discontinuities as distinct breaks.

Rhoades and Salinger (1993) have derived a number of statistical procedures for homogenizing isolated station data. Although adjustments for discontinuities are necessarily more subjective, a variety of graphical and analytical techniques were found useful in deciding homogeneity adjustments. These involve graphical analysis, simple statistical tests using annual and subannual differences over symmetric intervals, and a mathematical procedure to identify the most prominent change points in the time series independently of discontinuities identified by metadata. These procedures provide rules on when adjustments should be made.

3.2. Development of reference time series

A change in a station's time series may indicate inhomogeneities or may simply indicate an abrupt change in the regional climate. To isolate the effects of station discontinuities from regional climate change, many techniques use data from nearby stations as an indicator of the regional climate. Any significant variation from that regional climate signal is assumed to be due to inhomogeneities. While not a homogeneity detection method *per se*, use of nearby stations' data either directly or in the development of a reference series is integral to many methods.

The method used to form the reference time series can be important and may need to be tailored specifically to the network and adjustment methodology particularly because the homogeneity of the stations contributing to the reference series can usually not be assessed ahead of time. In some cases (e.g. U.S. historical climatology network), metadata have been used to determine which nearby stations would not be expected to have discontinuities during specific time periods. Another approach is that used by

Potter (1981). He created a reference series for a 19-station network that did not vary with time by using the mean of all the other 18 stations in his network for each candidate. After the homogeneity test was run on all the stations, he created a new reference series as before but excluding those stations with inhomogeneities.

Like Potter, Alexandersson runs homogeneity tests, then uses homogeneous data to create reference series which is used to rerun the tests. In Alexandersson (1986) three different techniques were used to create precipitation reference series. The first was an arithmetic mean of the homogeneous and complete stations. The second method was an arithmetic mean of normalized data so stations that were not serially complete could be used. The third method used a weighted mean of normalized data where the weighting was based on a distance function that was determined by spatial correlation. Alexandersson (1994) used the optimal interpolation technique for creating a reference series by using correlation coefficients between all sites and minimizing the coefficients of variation of the sequence of ratios (many inhomogeneity testing techniques are based on difference series when dealing with temperature and ratio series when dealing with precipitation). However, this technique was not successful due to an over sensitivity to the exact values in the correlation matrix. In more recent work (Alexandersson and Moberg, 1997) squared correlation coefficients (positive) of first difference series ($FD_i = (T_{i+1} - T_i)$, for year i, Peterson and Easterling, 1994) have been used as weighting factors when reference series are created.

Young (1993) also used several techniques to create reference time series to adjust sea-level pressure. For any given time element (i.e. month or year) his technique uses the median estimate of three techniques: multiple discriminant analysis, multiple linear regression, and the normalized anomaly; each of which provides a prediction for the candidate series. The sequence of predictions then becomes the reference series.

A major problem for the global historical climatology network (GHCN; Peterson and Vose, 1997) is that station coverage in some parts of the world, such as Africa, varies considerably with time. For example, GHCN has nine stations with temperature data in Zimbabwe but the data from four of these stations end in 1970. Therefore, to make maximum use of the available data, Peterson and Easterling (1994) created reference series from a network of stations that can change with time by choosing the best stations available for each year.

Building a completely homogeneous reference series using data with unknown inhomogeneities may be impossible, but several techniques were used to minimize potential inhomogeneities in the reference series. The first of these seeks the most highly correlated neighbouring station, but performs the correlation analysis on the first difference series. A change in thermometers would alter only one year of data in a first difference series whereas with the original data such a change alters all following years.

The second minimizing technique built a first difference reference series one year at a time by calculating the correlations without including the target year's data. Thus, if the candidate station's first difference series value for a year was excessively warm due to a discontinuity, the determination of that year's first difference reference series data point would not be impacted at all by that discontinuity. In creating each year's first difference reference series data point, the approach used the five most highly correlated neighbouring stations that had enough data accurately modeling the candidate station data such that the probability of this similarity being due to chance was less than 0.01 as determined by a multivariate randomized block permutation test (MRBP; Mielke, 1991).

The final Peterson and Easterling (1994) technique to minimize inhomogeneities in the reference series used the mean of only the central three values of the five highest correlated neighbouring stations to create each first difference reference series' data point. This is based on the assumption that if there was a significant discontinuity in one of the five stations that year, that station would most likely have the highest or lowest value. The final step in creating the reference series turned the first difference reference series into a station time series ($T_1 = 0$; $T_2 = T_1 + FD_1$) and adjusted the values so the final year's value of the reference series equaled the final year's temperature from the candidate series.

Other techniques, such as principal component analysis, may also produce very good reference series. While neighbouring station data are central to many homogeneity adjustment approaches, there are times when such data are not good enough. Peterson and Easterling (1994) determined that the correlation (r,

of the first difference series) between the reference series and the candidate station had to be 0.80 or higher to be reliable enough to use. Therefore, adequate reference series could not be made for many remote stations. Also, for some countries that change all their instruments at the same time (e.g. Finland), neighbouring station data cannot provide insights into those inhomogeneities.

3.3. Subjective methods

Subjective judgement by an experienced climatologist has been an important tool in many adjustment methodologies (Jones *et al.*, 1985; Jones *et al.*, 1986c; Plummer *et al.*, 1995) because it can modify the weight given to various inputs based on a myriad of factors too laborious to program. For example, when viewing a graphical display revealing a station time series, a neighbouring station time series and a difference series (candidate − neighbour), a subjective homogeneity assessment can factor in the correlation between the stations, the magnitude of an apparent discontinuity compared to the variance of the station time series, and the quality of the neighbouring station's data along with other information such as the relevance and reliability of the available station metadata. Subjective judgement can be particularly helpful in an initial inspection of the stations' data and when the reliability of certain inputs (e.g. metadata) varies.

Double-mass analysis (Kohler, 1949) can provide additional insights for a subjective assessment. A double-mass curve analysis plots the cumulative sum of the candidate station against the cumulative sum of a nearby station. Most double-mass plots are roughly linear so a sudden change to a new slope indicates a discontinuity. However, it is impossible to determine whether the indicated discontinuity occurred at the candidate or nearby station. To account for this problem, Rhoades and Salinger (1993) use plots of parallel cumulative sums (CUSUM) at several nearby stations at the same time.

3.4. Objective methods

3.4.1. Potter's method. Plummer *et al.* (1995) uses the technique commonly known as Potter's method (Potter, 1981) to generate a test statistic for each data value and an estimate of the maximum probable offset, or adjustment, for the data value. Potter's method is a likelihood ratio test between the null hypothesis that the entire series has the same bivariate normal distribution and the alternate hypothesis that the population before the year being tested has a different distribution than the population after the year in question. This bivariate test closely resembles a double mass curve analysis. One part of the test statistic depends on all points on a time series while another part depends only on the points preceding the year in question. The highest value of the test statistic will be in the year preceding a change in the mean of the candidate station time series. Potter (1981) applied this technique to ratio series of the candidate station's precipitation and a composite reference series.

3.4.2. Standard normal homogeneity test. Alexandersson (1986) developed the standard normal homogeneity test (SNHT) which is widely used. There are now variations in this test to account for more than one discontinuity, testing for inhomogeneous trends rather than just breaks, and inclusion of change in variance (Alexandersson and Moberg, 1997). Like Potter's method, the SNHT is a likelihood ratio test. The test is performed on a ratio or difference series between the candidate station and a reference series. First this series is normalized by subtracting the mean and dividing by the S.D. In its simplest form, the SNHT's test statistic is the maximum of T_v.

$$T_v = v(\bar{z}_1)^2 + (n - v)(\bar{z}_2)^2$$

where $\bar{z}_1$ is the mean for the series from data point 1 to v and $\bar{z}_2$ is the mean of the time series from v to the end, n.

At the Norwegian Meteorological Institute (DNMI), which uses the standard normal homogeneity test to test for homogeneity of series of precipitation, temperature and pressure, the output from the test includes the time series of a test parameter T, the 5 and 10% significance level for its maximum value ($T_{\max}$), and the adjustment value which should be used if there is an inhomogeneity at the time step when $T_{\max}$ occurs (Hanssen-Bauer *et al.*, 1991).

3.4.3. Multiple linear regression. In Canada, various approaches were first investigated (Gullett *et al.*, 1990, 1991), and a new technique based on multiple linear regression was developed by Vincent (1990, 1998) to identify steps and trends in temperature series. The technique is based on the application of four regression models to determine whether the tested series is homogeneous, has a trend, a single step, or trends before and/or after a step. The dependent variable is the series of the tested station and the independent variables are the series of a number of surrounding stations. Additional independent variables are used to describe and measure steps and trends existing in the tested series. To identify the position of a step, the third model is applied successively for different locations in time, and the one providing the minimum residuals sum of squares represents the most probable position in time of a step in the tested series.

The procedure consists of the successive application of the four models (Vincent, 1998). Each time, the residuals are analyzed to assess the fit. The autocorrelation for residuals several distances or lags apart are obtained. Consecutive statistically significant autocorrelations identified at low lags indicate the poor fit of the model; in this case, the fitted model is rejected and the next model is applied instead. When the autocorrelations become not significantly different from zero, the fitted model adequately describes the tested series. The estimated parameters corresponding to steps and trends provide the magnitude of each inhomogeneity. If there is significant autocorrelation after the application of the fourth model, the series is divided at the identified step and each segment is tested separately. In this way, the technique systematically divides the tested series into homogeneous segments. Adjustments are applied to bring each segment into agreement with the most recent homogeneous part of the series.

3.4.4. Two-phase regression. Solow (1987) described a technique for detecting a change in the trend of a time series by identifying the change point in a two-phase regression where the regression lines before and after the year being tested were constrained to meet at that point. Since changes in instruments can cause step changes, Easterling and Peterson (1995a,b) developed a variation on the two-phase regression in which the regression lines were not constrained to meet and where a linear regression is fitted to the part of the (candidate − reference) difference series before the year being tested and another after the year being tested. This test is repeated for all years of the time series (with a minimum of 5 years in each section), and the year with the lowest residual sum of the squares is considered the year of a potential discontinuity. A residual sum of the squares from a single regression through the entire time series is also calculated. The significance of the two phase fit is tested with (i) a likelihood ratio statistic using the two residual sums of the squares and (ii) the difference in the means of the difference series before and after the discontinuity as evaluated by the Student's t-test.

If the discontinuity is determined to be significant, the time series is subdivided into two at that year. Each of these smaller sections are similarly tested. This subdividing process continues until no significant discontinuities are found or the time series are too short to test (< 10 years). Each of the discontinuities that have been identified are further tested using a multiresponse permutation procedure (MRPP; Mielke, 1991). The MRPP test is non-parametric and compares the Euclidean distances between members within each group with the distances between all members from both groups, to return a probability that two groups more different could occur by random chance alone. The two groups are the 12-year windows on either side of the discontinuity, though the window is truncated at a second potential discontinuity. If the discontinuity is significant at the 95% level (probability $(P) = 0.05$), it is considered a true discontinuity. The adjustment that is applied to all data points prior to the discontinuity is the difference in the means of the (station − reference) difference series' two windows.

3.4.5. Rank order change point test. Using a test based on the ranks of values from a time series has the benefit that it is not particularly adversely affected by outliers. Lanzante (1996) describes such a non-parametric test related to the Wilcoxon-Mann-Whitney test. The test statistic used is computed at each point based on the sum of the ranks of the values from the beginning to the point in question (Siegel and Castellan, 1988). The first step is determining the rank of each point in the time series and then making a series of the sum of the ranks (SR_i). Next an adjusted sum (SA_i) series is calculated for the series of length n where: $SA_i = |(2SR_i) - i(n+1)|$. The maximum value of SA_i, except for the last point,

is considered the point of possible discontinuity. If x is the point where the maximum value of the series SA_i occurs, then the test statistic z is calculated:

$$z = \frac{(SR_x - x(n+1)/2 + d)}{[x(n-x)(n+1)/12]^{0.5}}$$

where $d = 0$ if $SR_x = x(n+1)/2$, $d = +0.5$ if $SR_x < x(n+1)/2$, and $d = -0.5$ if $SR_x > x(n+1)/2$. If $x > 10$ and $(n-x) > 10$, then a two-tailed test using a normal probability table can be used to assess the test statistic's significance (Lanzante, 1996).

3.4.6. Craddock test. Developed by Craddock (1979), this test requires a homogeneous reference series though sometimes long enough homogeneous sub-periods are sufficient (Boehm, 1992). The Craddock test accumulates the normalized differences between the test series and the homogeneous reference series according to the formula:

$$s_i = s_{i-1} + a_i \bullet (b_m/a_m) - b_i$$

where a is the homogenous reference series, b is the time series to be tested and a_m and b_m are the time series means over the whole period. If the tested climate element becomes zero (or near zero) it has to be transformed by an additive constant to avoid dividing by zero. For temperature this can be done by using K instead of °C.

In Austria, the Craddock test was found to produce a clear signal in the case of pressure and temperature (Boehm, 1992). More difficulties occur with precipitation due to the temporal and spatial variability of this element, but all in all the results are sufficient. However, significant problems occurred when the Craddock test was applied to snow records (Auer, 1992).

3.4.7. T-test. The usual Student's t-test (Panofsky and Brier, 1968) has also been used to assess homogeneity. For example, this test has been used at the Norwegian meteorological institute (DNMI) for temperature time series when metadata has already indicated a specific date for a major change.

3.4.8. Caussinus-Mestre technique. The Caussinus-Mestre method simultaneously accounts for the detection of an unknown number of multiple breaks and generating reference series. It is based on the premise that between two breaks, a time series is homogeneous and these homogeneous sections can be used as reference series. Each single series is compared to others within the same climatic area by making series of differences (temperature, pressure) or ratio (precipitation). These difference or ratios series are tested for discontinuities. When a detected break remains constant throughout the set of comparisons of a candidate station with its neighbours, the break is attributed to the candidate station time series.

For detection purposes, the formulation described by Caussinus and Lyazrhi (1997) is used which allows the determination of a normal linear model with an unknown number of change-points and outliers. They formulated it as a problem of testing multiple hypotheses, and provided a Bayes invariant optimal multi-decision rule for detecting a set of such perturbations based on a penalized log-likelihood statistic. The penalty term curbs the increase in the likelihood and picks the solution with the right number of breaks most of the time. For practical computation, in the case of multiple breaks in a Gaussian sample, an adapted step-by-step procedure (called double-step) is used (Caussinus and Mestre, 1996). A two factor linear model is used for adjustments. Additional information for climatic studies is provided by an estimator of the amplitude of the minimum detectable inhomogeneity.

3.4.9. Multiple analysis of series for homogenization (MASH). The MASH method, developed by Szentimrey (1994, 1995, 1996) in the Hungarian meteorological service, also does not assume a reference series is homogeneous. Possible break points and shifts can be detected and adjusted through mutual comparisons of series within the same climatic area. The candidate series is chosen from the available time series and the remaining series are considered reference series. The role of the various series changes step by step in the course of the procedure. Depending on the climatic elements, additive or multiplicative models are applied. The multiplicative models can be transformed into additive models by conversion to logarithms.

Several difference series are constructed from the candidate and weighted reference series. The optimal weighting is determined by minimizing the variance of the difference series, in order to increase the efficiency of the statistical tests. Providing that the candidate series is the only common series of all the difference series, break points detected in all the difference series can be attributed to the candidate series.

A new multiple break points detection procedure has been developed which takes the problem of significance and efficiency into account. The significance and the efficiency are formulated according to the conventional statistics related to type one and type two errors, respectively. This test obtains not only estimated break points and shift values, but the corresponding confidence intervals as well. The series can be adjusted by using the point and interval estimates.

4. APPROACHES FOR CREATING HOMOGENIZED CLIMATE DATA SETS

All of the above techniques and sources of information have been used in homogeneity adjustment approaches. But they have been used in many different ways. How one combines metadata and a statistical technique depends on a wide variety of factors such as the completeness of metadata (which can vary from country-to-country and decade-to-decade), the expected uses of the data, and the human resources available for the work. An equally important aspect is the individual researcher's philosophy towards homogeneity adjustments that has developed over years of working with data and metadata. This Section describes several different adjustment approaches used in both country specific and global data sets. Understanding the approaches and the rationale behind the approaches can be important for any user of homogeneity-adjusted data.

4.1. Country or region specific approaches

4.1.1. Australia. Two different approaches have been used to adjust Australian climate data. Torok and Nicholls (1996) make a subjective decision on the position and magnitude of adjustments, based on an objective statistical test (Easterling and Peterson, 1995a,b), comparisons between the candidate and median reference series, individual comparisons with independent stations, visual analysis of the diurnal temperature range (DTR), and station history documentation. Mean monthly maximum and minimum temperatures were considered separately, as they behave differently under different climatic conditions, are influenced by different non-climatic factors and sometimes have different periods of record.

The comparison of temperature with other parameters at the same station, such as rainfall, or to proxy data, such as tree-rings, was not feasible for calculating adjustments to temperature. These parameters were either less reliable or had a shorter period of record than temperature, or correlations between the parameter and temperature were not sufficiently high. Comparison of annual values with the long-term averages at a candidate station was found to be an inadequate method of homogeneity testing. A comparison with a simple average of surrounding stations was also found to be inadequate.

An 'ideal' neighbouring temperature record was created for each candidate station using the median interannual temperature differences from surrounding stations. In the first step, neighbour stations were identified within 6° of latitude and longitude of the candidate station. The use of only nearby stations ensured that stations with climates dissimilar to that of the candidate were not used, and minimized computer processing time. Stations were excluded if they were within 6° of latitude and longitude but in a location likely to be climatologically dissimilar to the candidate station (according to a previously determined climate classification), such as on different sides of a mountain range. Stations with temperatures that were possibly affected by urbanization were not included in the reference series.

All neighbouring first difference (*FD*, see Section 3.2) series were correlated with the candidate *FD* series. Only stations which were strongly, significantly and positively correlated with the candidate station were used in the development of the neighbouring record (it was necessary for the correlation to be at least 0.7, significance level to be below 0.05 and the number of years of record to be greater than ten for a nearby station to be selected as part of the neighbour data set). All available years of temperature data

were graphed and analyzed. The values of *FD* at the qualifying neighbouring stations were sorted and the median value of *FD* for each year was used to compile the reference series. (The median value was used to ensure that discontinuities at a neighbouring station had minimal influence on the reference series.) The value therefore represents how the temperature should have changed from one year to the next in the area surrounding the candidate station. The total number of neighbour stations used to compile a median reference series was typically about 65, although the number was much less than this in sparse data areas. Only about half of these would be available in each year, because of missing data and because of the short periods of record at some neighbour stations.

The median *FD* series was then converted to a median reference temperature series by the addition of *FD* values to the temperature in the first year of record at the candidate station. It was assumed that the reference series is free from discontinuities, and therefore any change in the relationship between the candidate and median reference series is likely to be due to a change at the candidate station. This will appear as an abrupt change or spurious trend in the difference series (candidate − median reference series). Change points in the difference series were objectively identified using automated 2-phase regression techniques developed by Easterling and Peterson (1995a,b).

To complement the objective test, both the median and candidate series for each station and the series of differences between the two were printed for visual analysis. Any years with an apparent step change in the mean difference were recorded. Individual graphical comparisons were conducted with other stations that had been excluded from the median reference series, such as major cities in the vicinity of the candidate station. In these cases it was ensured that positive trends identified in the difference series due to the growth of the nearby city were not interpreted as a cooling at the candidate station.

Visual analysis of the diurnal temperature range (DTR) series was also used as a method to identify problems, as discontinuities often affect the maximum and minimum temperatures differently. Next, the station history documentation files were studied in order to investigate moves and other changes likely to affect the temperature record over the period of record covered by the long-term data set. Finally, based on all this information, a subjective decision was made on which adjustments were necessary and the magnitude of each adjustment.

Another approach to Australian data (Plummer *et al.*, 1995) concluded that the relatively sparse Australian station network made it difficult to construct an appropriate reference series for all areas. So it was decided to perform Potter's test between the candidate's data and data from each neighbour individually. Through the redundancy of a multiple station comparisons, the most suspect (in terms of data quality) stations in the comparison were identified. Iterating through each station in the comparison, the test produced a time series of potential inhomogeneities detected at a station based on the comparison with each of its neighbours. The test was applied on sub-periods of data where it was determined that more than one inhomogeneity was involved.

Using information from statistical plots of this information along with the significance of the inhomogeneities at the 99, 95, and 90% confidence levels, graphical comparisons of anomaly differences and available metadata, the time series for stations which were considered highly likely to contain an inhomogeneity were adjusted (annual followed by appropriate seasonal adjustments). The multiple station test was repeated using the adjusted data for the neighbour (or candidate) station and the next most obvious adjustment made, if necessary. The process was repeated until the data from the candidate station had been suitably adjusted or required no adjustment. The annual and seasonal adjustments were calculated by the testing procedure as an average of the individual comparison tests. However, these were recalculated if it was obvious that the adjustment suggested by the technique was an outlier.

The amount of available metadata varied markedly between stations and over the period of operation of individual stations and so supporting metadata was not made a requirement for making an adjustment. The final decision to adjust a neighbour's or candidate's data was still quite subjective but based on: the number of stations in which the comparison was indicating an inhomogeneity, the statistical significance of those discontinuities, coincident metadata suggesting a possible inhomogeneity, coincidence in both maximum and minimum temperatures, and evidence from graphical plots.

4.1.2. Austria. At Austria's Central Institute for Meteorology and Geodynamics (CIMG), a combined approach using metadata and the Craddock test (Section 3.4.6) has been used successfully to test the homogeneity of a rather dense (the mean distance between stations is approximately 40 km) temperature and precipitation network. The method is described in Auer (1992) and Boehm (1992). With this combined approach it is not necessary to trust the completeness of station history descriptions. For example, in earlier times the relocation of instruments within 10 m was often not recorded. In addition, changing environment (growing trees or cutting of trees) is seldom contained in station descriptions. And the frequency of station checks is too low in many cases to fully trust the metadata. On the other hand, statistical techniques are not without drawbacks as well, raising questions as to whether one should consider the significance levels the only criterion in the homogeneity adjustment decision. To illustrate a problem with relying solely on significance levels, Austrian glaciers have retreated to less than 50% of their former size since 1850 accompanied by a non-significant temperature rise of about 1°C (signal to noise ratio of about 1).

Adjustments are done at CIMG in two steps. In a first step, the *a priori* known inhomogeneities are eliminated. A typical example of a known inhomogeneity was one caused by the change of the evening observation time from 2100 to 1900 h from 1970 to 1971. In that case, hourly temperature data sets could be used to adjust all mean temperatures to 'true means' based on 24 hourly readings per day. For the second step (dealing with the *a priori* unknown inhomogeneities) the solution at CIMG is to adjust time series at break points supported by metadata and also at break points that are not supported by metadata if the signal is of the same (or greater) magnitude as metadata-verified signals. According to the different spatial correlation coefficients of the various elements, a correlation coefficient of at least 0.7 between test series and reference series is required. The adjustments are on monthly data and adjustment values normally show a marked annual course.

The time series are adjusted step-by-step beginning with the most recent sub-period. The length of a sub-period is chosen to have only one break per sub-period (with new testing after each adjustment). For temperature series, mean differences in the sub-periods before and after the break are used, for precipitation, mean ratios are used. Reference series are chosen from those being homogeneous in the respective sub-period only. They can be single series or a number of highly correlated series and they can change from sub-period to sub-period, thus avoiding the danger of adjusting all series to one initial reference series and possibly damping real regional differences. One result of testing and adjusting long-term Austrian temperature and precipitation series was—and this should be of more than local relevance—the realization that long series are never homogeneous. The typical length of a homogeneous sub-period is only 30–40 years. Further results are published in Auer and Boehm (1994a,b).

In 1997, CIMG began an extension and a complete re-analysis of its long-term series. It will include about 20 other climate elements in order to obtain true climate time series and to overcome the single- or bi-elemental approach which is typical for climate change research up to now. In addition to the Craddock test and the SNHT, the new methods by Szentimrey (1996) and Caussinus and Mestre (1996) are used. The multi-elemental approach allows 'internal testing' (using different elements of the same station) such as sunshine duration versus cloudiness or extremes versus means. Adjustments are done by linear regressions instead of constant differences or ratios. The results (10–20 stations with *ca*. 20 climatic elements, original and adjusted series, and metadata information) are expected to be ready by the end of 1998.

4.1.3. Canada. Homogeneity testing and data adjustments have been applied to the maximum and minimum temperature series of 210 'base' climate stations in order to create a database of long-term, complete and homogeneous series. Data sets have been extended back in time as near as possible to 1895, however in northern Canada, systematic observations are available generally only from the early 1940s. Considerable spatial disparity also exits between areas north and south of 60°N, with a greater network density in the southern part of the country. For each base station, four to six neighbouring stations are chosen for estimation of missing monthly values and for relative homogeneity testing. The selection of neighbours is based on distance from the test site, elevation, landscape and vegetation features, and

similarity of temperature regime as determined through correlation statistics. Neighbours within 50–100 km of the base are preferred but sometimes they can be as far as 500 km in the north.

Using a multiple linear regression (Section 3.4.3; Vincent, 1998), annual mean maximum and minimum temperatures are tested separately. After identification of any non-climatic steps in the base series, the station history files (metadata) are consulted to determine, if possible, the causes of the inhomogeneities. While the history files are extremely useful, they are, however, usually incomplete. A conservative approach is adopted in which adjustments are carried out only if the magnitude of the identified step in the annual temperature series is large (greater than 0.5°C). Steps of less than 0.5°C are adjusted only if their cause can be retrieved from the station history files. The philosophy used is that it is better not to adjust than to erroneously adjust since some adjustments can actually make the data more biased than if no adjustment had been applied. An adjustment factor is derived for each month by applying the third model (see Section 3.4.3) to each of the 12 series of monthly values at the identified position of the step in the annual series. After adjusting, it is necessary to verify that the new adjusted series more closely reflect the general temperature pattern of the area and this is done by comparing the new linear trend with those of the neighbouring stations.

Identification and adjustment of inhomogeneities in Canadian daily precipitation data sets have also been carried out using a very different approach (Mekis and Hogg, 1997). The method employed for precipitation do not rely on the availability of neighbour stations, but rather use the station history files for identifying the precise dates for known instrument changes. This method removes discontinuities resulting from network-wide or systematic changes but not individual station discontinuities caused by local site alterations. Daily rainfall and snowfall are therefore adjusted specifically for inhomogeneities caused by precipitation gauge changes and by changes in observing procedures. These absolute inhomogeneities typically occur in the data series of most stations in a region at roughly the same time, and can usually be verified through the station history files. Precipitation data sets have been assessed and adjusted for 69 Canadian stations which are relatively evenly distributed across the country, and which cover the period 1895–1996 as much as possible. As for temperature, relatively few stations are available across northern Canada. Work continues on the expansion of the Canadian adjusted daily precipitation database.

4.1.4. Finland. At the Finnish Meteorological Institute (FMI), the philosophy has been to use two complementary methods: analyzing metadata and statistical tests of station data. Relative homogeneity tests, like SNHT, are not capable of detecting simultaneous, similar-size discontinuities. Therefore, first discontinuities biasing the whole network were adjusted according to metadata and comparison measurements. These adjustments can be verified at least near country borders with data from a neighbouring country. Next adjustments were made to account for changes in temperature averaging methods and changes precipitation gauge types (Heino, 1994). Finally, the homogeneity of individual series was tested with SNHT (Tuomenvirta and Drebs, 1994) in a two step process.

The testing was done on annual mean temperatures and precipitation totals. The 95% confidence limit was used to classify the test parameter maximum as a homogeneity break. The exact date and reason of discontinuity was searched from the station history (year suggested by SNHT $\pm$ 2 years). If the station history did not indicate any reason for the discontinuity, the time series was classified as suspect. The limiting requirement of finding a physical reason for a break in the metadata before adjusting cannot be used unless metadata information is fairly comprehensive, but the metadata are rarely complete. However, in this way, the danger of adjusting time series to be homogeneous relative to each other, but not necessarily correctly describing the climate can be avoided (Tuomenvirta and Heino, 1996).

The role of metadata also becomes important when data are sparse (remote places and/or 19th century network). A simple example is a case where there are only two series available and one discontinuity is found. Statistical tests cannot tell which of the series should be adjusted. In these kinds of situations, metadata is used to guide testing and adjusting procedures. The SNHT is ideally meant for detecting one break at a time. Therefore, metadata were used to divide time series into overlapping periods containing at most one potential discontinuity.

In order to limit the amount of test results, annual values were tested. However, the adjustments should be made on the monthly level to be able to study seasonal series. On an annual level, the minimum size of discontinuities detected was roughly 0.2°C and 1.04/0.95 as ratio for precipitation. Testing annual values created a risk of not detecting seasonal discontinuities that cancel each other out on an annual level. In the future, one possible improvement for the testing of Finnish series that is being considered is to repeat testing on a seasonal level. Also, an adjustment method where the size of the adjustment is a linear function of the temperature anomaly has been developed (Tuomenvirta and Alexandersson, 1995).

4.1.5. New Zealand and Pacific Islands. Two techniques for homogeneity analysis have been developed for New Zealand and South Pacific islands (Rhoades and Salinger, 1993; Salinger *et al*., 1995). When station history information suggested a significant change, together with the neighbour station comparisons and statistical tests suggesting significance, homogeneity adjustments to the data were made. For stations with several neighbours, the decision to adjust can be taken with some confidence. Unfortunately, for isolated stations this is not the case. However, large shifts can be recognized and adjusted for, though with some uncertainty. The case is particularly strengthened when the three methods and other outside evidence converge.

Where neighbour stations exist, the following features are identified from station histories: (i) site changes; (ii) changes in the environment (e.g. airport extensions, building alterations, growth of vegetation, and urban expansion); and (iii) instrument changes. Times of possible discontinuities are identified from the station histories. Next, plots of cumulative sums (CUSUM, Section 3.3) of observations for temperature with CUSUM plots from neighbour stations, and rainfall ratios of a target station with its

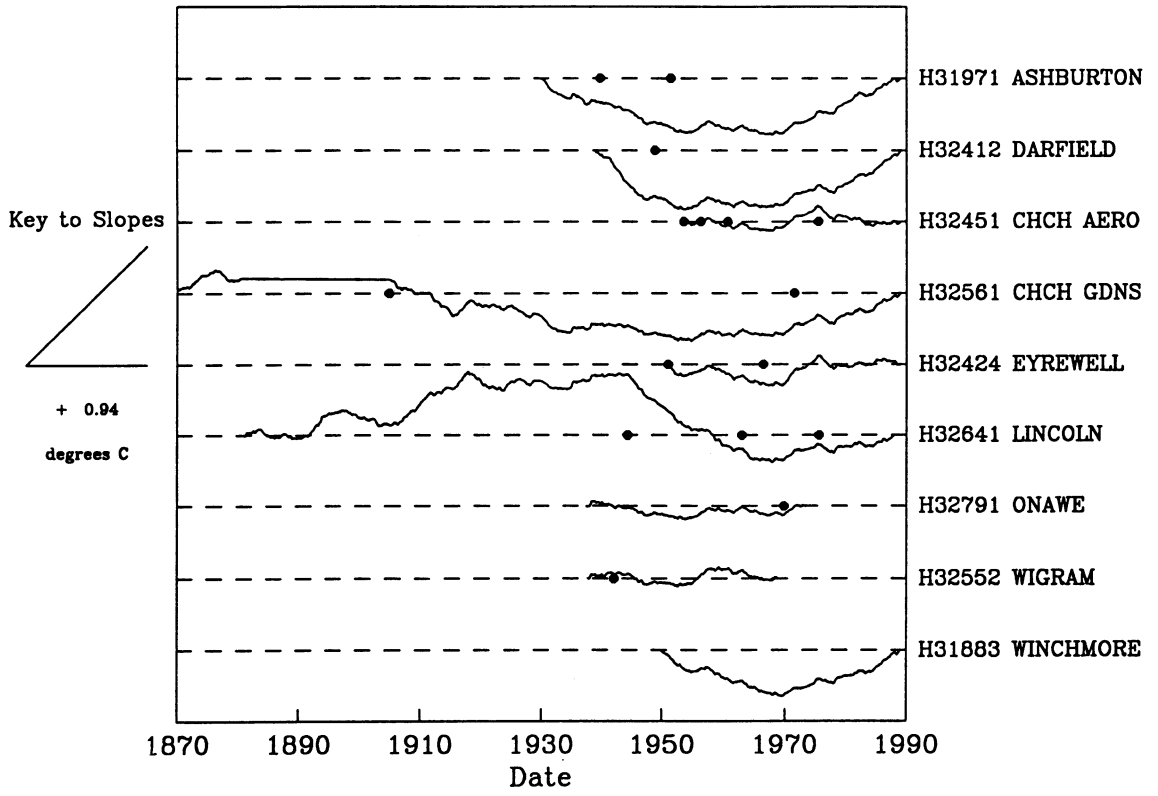


Figure 4. Parallel CUSUMS of mean daily minimum temperature from stations in and around Christchurch, New Zealand. Dots show the times of site changes determined from station histories. The seasonal cycle has been removed. Change points that occur at the same time in many graphs most likely reflect variations in the climate, while changes that affect the graph of only one station are indicative of inhomogeneous data (from Rhoades and Salinger (1993), reproduced courtesy of the Royal Meteorological Society)

neighbour's, were made to visually detect changes (see Figure 4). These were compared to identify important discontinuities from the metadata.

A *t*-statistic was used for estimating the size of site-change effects. Temperature series were differenced with neighbour's before and after the identified discontinuity for monthly series for 1, 2 and 4 years. For rainfall, logarithms of monthly rainfall ratios between the target station and its neighbours are taken. From these, an estimated discontinuity and S.E. can be calculated. Adjustments were made for the discontinuity when it was significant at the 5% level.

The isolated Pacific Island stations do not usually have nearby neighbours. Therefore, adjustments for discontinuities requires a much greater uncertainty and a greater degree of subjectivity (Salinger *et al.*, 1996). However, station histories were available, so recorded potential discontinuities could be identified. A visual analysis of the data was first used. Next the size and error of the potential discontinuity was identified by using a comparison of annual (means and totals) before and after the discontinuity. There is a danger of a longer-term trend being confused with a discontinuity, so other steps were taken. Subannual differences were also tested using symmetric intervals about the time of the discontinuity point. For South Pacific sites, monthly differences spanning 1, 2 and 4 years around the discontinuity point were used. Finally, the most prominent change points in the temperature or rainfall time series, where the level seems to change from one value to another were mathematically identified. For a given number of change points, the optimal partitioning set is efficiently determined by dynamic programming, as discussed by Seward and Rhoades (1986). When all three methods converged to give significant discontinuities, adjustments were made.

4.1.6. Norway. A study of 165 Norwegian precipitation series showed that 70% of the series were inhomogeneous (Hanssen-Bauer and Førland, 1994). The most frequent reasons for inhomogeneities in Norwegian precipitation series were relocation of the gauges (47%), changes in buildings and vegetation in the environments of the gauges (18%), and installation of windshields (9%, it should be noted that to prevent inhomogeneities, a large number of old precipitation stations in Norway are still run without a windshield). Changes in the gauge's height above the ground are usually so small that they have little effect on the measurements, except in areas with much drifting snow.

The most frequent reasons for inhomogeneities in temperature series are relocation, changes of temperature screen, changes in environments and changes in observational hours (Nordli *et al.*, 1997). Changes in the height above ground of the thermometer often lead to inhomogeneities in series of temperature from one specific observational hour and in the series of maximum and minimum temperatures. For daily temperature means, the influence of such changes is usually minor. The reasons for the inhomogeneities found in series of air pressure are improper use of instrument corrections and change in instrument height above sea level.

At the Norwegian Meteorological Institute (DNMI), it is thought that it will be impossible to quantify inhomogeneities by using metadata alone. Statistical analyses have indicated that, while some major changes at a station have not led to any inhomogeneities at all, some small relocations may have substantial influence on the wind exposure of a precipitation gauge and, thus, lead to severe inhomogeneities. On the other hand, while a statistical test could stand on its own feet, it also implies larger uncertainties and larger possibilities of adjusting series which are, in fact, homogeneous. Therefore, a higher significance threshold is used for statistically-identified inhomogeneities that are not supported by metadata. Inhomogeneities are adjusted for: (i) if the value of standard normal homogeneity test statistic lies between the 5 and 10% significance levels and there is metadata documentation to support that change, or (ii) if the value of test statistic exceeds the 5% level (even without support in metadata).

The best approach is a hybrid where metadata are used as a supplement to the statistical testing, but alas, sufficient metadata is not always available. Up to 10 years pass between station inspections, and not all inspectors have known what to look for in the way of potential inhomogeneity producing changes. Collection of additional information has occasionally enabled DNMI to explain inhomogeneities which were not supported by information in DNMI's metadata files.

In practice, several tests are run for each station, using different reference series and periods. The choice of reference series in the final test, thus, leaves some room for subjective decisions. Two aspects should be considered in this context. As far as possible, reference series should represent areas in different directions from the test station. This may be impossible in areas where the station coverage is poor. In such cases, one should be especially reluctant to introduce adjustments unless they are confirmed by metadata. A study of the homogeneity of climate series from the Norwegian Arctic (Nordli *et al.*, 1996) revealed systematic differences between the precipitation trends in different parts of the area, which led to apparent inhomogeneities between series from these different parts. It is also important to be aware of changes (e.g. in instrumentation) which happened more or less simultaneously at several stations in an area. If possible, at least one test should be run against reference series which are not affected by this change.

The effect of inhomogeneities on temperature is often opposite during winter and summer. Therefore, testing of annual mean temperatures is not recommended if the series are intended for studies of seasonal trends. Single months may be tested, but the noise level is then increased. As a result, in Norway, available metadata and seasonal test results are used to determine if and when there is an inhomogeneity. Adjustment values are then calculated on a monthly basis. But the monthly adjustment values need to be smoothed in order to give a reasonable variation throughout the year.

For precipitation, both seasonal and annual test results are used in addition to the metadata, to state if and when there is an inhomogeneity. The noise level is larger for seasonal testing, therefore, many inhomogeneities are most easily detected using annual values. Accordingly, some series were tested and adjusted only on an annual basis. These series should not be used for studies of seasonal trends. For this purpose, seasonal adjustment factors were calculated for some selected series, and monthly adjustment factors were estimated from these. As with temperature, the adjustment values for precipitation may vary throughout the year (Hanssen-Bauer *et al.*, 1995; Hanssen-Bauer *et al.*, 1997).

In Norway there are three major reasons for such variability: (i) the effect of inhomogeneities on rain and snow is quite different (Førland *et al.*, 1996). For example, a change in wind speed of 1.0 m s^{-1} implies a change in gauge catch of 10% for snow, but has a small effect on catch efficiency for rainfall (Førland, 1994). Thus, changes in the environment around the gauge will lead to different adjustment factors for rain and snow; (ii) in mountainous areas, the orographic enhancement varies substantially throughout the year. Accordingly, the adjustment factor for a relocation will be different for different seasons. Finally, in areas where drifting snow is a serious problem, the effect of a relocation on the catch efficiency of snow and rain may even be opposite (Nordli *et al.*, 1996).

4.1.7. Sweden. The main homogeneity testing tool at the Swedish Meteorological and Hydrological Institute (SMHI) is the standard normal homogeneity test (Section 3.4.2; Alexandersson, 1986), but there are some variations of this technique which can also be used. For example, metadata is often used as a guide for exact dating when possible (though this is often not possible before about 1925 because of poor documentation).

Generally, SMHI's procedure often utilizes the following logic. Someone requests a homogenized series for station Y. There are already some fairly well tested and reasonably homogenized series available to serve as reference series. Suppose that X_1–X_k are chosen as the reference series out of these series for testing Y. Plots are then made of the sequence of ratios or differences for Y compared with a weighted mean value of the X (see Figure 5). Then, after some iterative work, which as a rule includes checking the metadata archive and testing subsets of the series, the SNHT program is supplied with some dates of breaks so that one output is monthly adjustments that can be used to get homogenized (or, perhaps more honestly, less inhomogeneous) series. In practice it is a bit cumbersome to use trends for adjustments (Alexandersson and Moberg, 1997) so up till now SMHI have mainly used distinct breaks.

4.1.8. U.S. wind-driven winter precipitation adjustments. There are several reasons for precipitation under-catch by various standard gauges used worldwide (Sevruk, 1982), but the main factor is wind-induced turbulence over the gauge orifice. The resulting error is most pronounced for snow measurements. The existing U.S. cooperative rain gauge network measures rainfall with a bias of 3–10% and snowfall

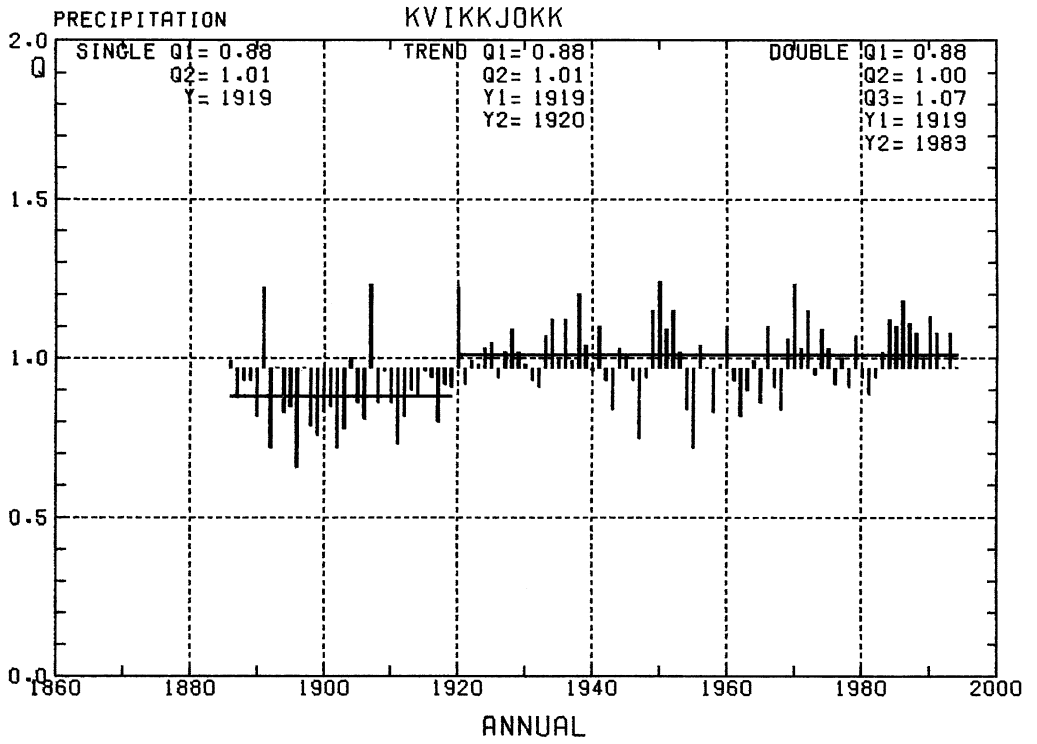


Figure 5. Ratio series of annual precipitation values where data from Kvikkjokk, Sweden are used for the Y-series and a homogeneous reference series for the X-series of the ratio. The final adjustment was 1.16 (+16%) on annual values 1886–1919

with a bias of up to 50% or more (Larson and Peck, 1974; Groisman and Legates, 1994). Over the last century, the U.S. has experienced a systematic increase of the mean angle from the gauge to nearby obstacles which can be explained by official instructions to place the gauge in more protected sites (Alter, 1937; Riesbol, 1938). Over time, this can cause an increase in measured winter precipitation that can result in significant spurious positive trends in frozen and liquid precipitation derived from the observed precipitation data even when there is no actual change in precipitation. To adjust for this bias, Groisman *et al.* (1996) devised an approach that uses detailed metadata.

The metadata used are information about the topography of the stations, descriptions of the gauges used, gauge exposure (information about wind obstacles, their distance from the gauge, and their height), gauge elevation, and changes in these variables with time. These metadata files were digitized for as much of the period of record as possible and special software was developed to apply the metadata in a bias adjustment procedure. The mean direction, distance, and zenith angle from the gauge to each obstacle was calculated for every significant change throughout each station's history (Figure 2). Similar information has been recorded for anemometers at nearby first order stations. This has been done to ensure correct wind time series which are used in the adjustments.

The regional adjustment coefficients for light rain and drizzle measurements increase the measured monthly rainfall values in the range of 2–10%. The adjustments for frozen precipitation measured by standard U.S. rain gauges are much larger and are based on the relationships between snow under-catch by the gauge and wind speed over its orifice during the precipitation event (Larson and Peck, 1974; Golubev *et al.*, 1995). These relationships are described by steep functions of wind speed for each type of gauge; as wind speed increases, gauge efficiency dramatically decreases. Therefore, a knowledge of the wind speed over the gauge orifice during precipitation events is critical to the proper evaluation of adjustments for frozen precipitation. To make these calculations, the algorithm uses a roughness parameter derived from metadata and information about wind speed during precipitation events deduced from nearby first order station data in addition to gauge type information (Groisman *et al.*, 1996).

4.1.9. U.S. historical climatology network. The monthly maximum, mean, and minimum temperature and precipitation data in the United States Historical Climatology Network data set (U.S. HCN; Easterling *et al*., 1996a) are adjusted for a number of known biases, as well as unknown biases. The first adjustment to the data is for the time-of-observation bias in monthly temperatures calculated from stations that observe maximum and minimum temperature only one time per day (Karl *et al*., 1986). When the observing time differs from midnight, a bias in monthly averaged quantities is introduced by the shifting of observations at the beginning or end of the month to the adjacent month. The next known bias adjustment is for the effects of the introduction of the maximum–minimum temperature system (MMTS) into approximately half the stations in the network (Quayle *et al*., 1991; see Figure 3).

Metadata play an important role in the preceding adjustments, but the use of metadata is probably best illustrated by the station move adjustment routine (Karl and Williams, 1987). In this procedure the station histories are examined to determine when a change (e.g. move, instrument change, etc.) at a station (candidate station) has occurred that could affect the time series homogeneity. If such a change has occurred, the station histories for the stations nearest to the candidate station are used to find 3–5 stations with no documented changes in the 5 years on either side of the change to use as reference stations. Seasonal candidate time series are then compared with the reference time series in a statistical procedure to determine if the change has resulted in a statistically significant discontinuity and, if so, an adjustment value is applied to the candidate time series. The last procedure is to adjust the temperature time series at each station for urban warming. This adjustment is a regression-based approach with population as the predictor (Karl *et al*., 1988).

4.2. Global approaches

4.2.1. Jones global data set. The Jones *et al*. (1986a,b,c, see also Jones, 1994) data set is a global monthly mean temperature data set where the vast majority of the stations have undergone homogeneity assessment. The assessment of station homogeneity is performed by comparison of the annual temperature time series with that from neighbouring stations. Station difference time series plots are subjectively studied for jumps and/or trends. Using n stations in a small group there will be $(n-1)n/2$ comparisons. At least three stations enables the errant station (the one with the problem) at a particular time to be isolated (if $n=2$ there is only one comparison, so either could be wrong).

All comparisons are done visually. Stations revealing multiple jumps (≥ 3) and/or trends in the difference series are deemed non-homogeneous and not correctable. These stations are removed from the data base. The homogeneity assessment is only done on the annual data so compensating errors during the year will not be detected. The ability to detect jumps visually depends on the station variability. It is much easier to detect jumps in the tropics than in the higher latitudes because the year to year variability is smaller in the tropics compared to higher latitudes.

Having located the jumps in the station time series, a reason for the change may be apparent in the station history file (Bradley *et al*., 1985). Even if there is no reason, adjustment is undertaken. Adjustment uses from one to three neighbouring stations (homogeneous or previously adjusted stations) to calculate monthly adjustment factors for the errant station using at least a 10-year period before and after the jump. Adjustment factors, which are calculated on a monthly basis because they generally have an annual cycle, are added to all years before the jump.

Sites cannot reliably be adjusted for jumps near the ends of records. Each station is assigned a first reliable year (FRY) and this is generally the first year of records. If the difference plots reveal a few odd years at the beginning of a station time series, these are ignored in later analyses by the FRY value. Only data for years after and including this year are used. In some regions and even in earlier times for data-dense regions, comparisons with neighbouring stations may not be possible. For these sites, the annual mean temperature time series is visually inspected and the data are either accepted, a FRY value assigned, or the period of bad data is deleted.

4.2.2. Global Historical Climatology Network. The Global Historical Climatology Network (GHCN; Peterson and Vose, 1997) includes data sets of both original and homogeneity-adjusted time series of mean monthly maximum, minimum, and mean temperature. Because of the paucity of available station history information for many of the 7280 GHCN temperature stations around the world, no metadata beyond latitude and longitude are used in the GHCN adjustment methodology. First a reference series for each station is made as described in Section 3.2 and Peterson and Easterling (1994) then the candidate − reference series is tested using a 2-phase regression technique described in Section 3.4.4 and Easterling and Peterson (1995a,b). This determines the date and the magnitude of the adjustments.

All GHCN homogeneity testing is done on annual time series that have at least 20 years of data. Annual reference series are more robust than monthly series, but the effects of most discontinuities vary with the season. Therefore, monthly reference series were created and the differences from them in intervals before and after the annual discontinuity values were compared for each month. The potential monthly adjustments were then smoothed with a 9-point binomial filter and all the monthly adjustments were adjusted slightly so that their mean equaled the adjustment determined by the annual analysis. There is one significant exception to the GHCN adjustment methodology. GHCN includes data from the U.S. HCN. Since both U.S. HCN and GHCN are produced at the same institution, it was deemed inappropriate to produce two different homogeneity-adjusted versions for the same station. Therefore, GHCN simply incorporates the U.S. HCN adjusted version. However, in the future some GHCN methodology is expected to be incorporated into the U.S. HCN approach.

4.2.3. UK Meteorological Office marine air temperature data. Folland *et al.* (1984) introduced the use of night marine air temperatures (NMATs) to largely avoid the daytime heat-island effect caused by solar heating of ships' structures. Diurnal ranges of marine air temperatures (MAT) measured in recent decades on ships' decks are typically 2°C in the tropics (Parker *et al.*, 1995), in contrast with typical diurnal ranges of 0.4°C for SST (Bottomley *et al.*, 1990), emphasizing the magnitude of the daytime bias in ships' MATs. Following the recommendation of Parker *et al.* (1995), the selection of 'night' observations has been refined to minimize the effects of residual solar heat on ships' decks after sunset. However, systematic biases remain in NMATs; these result from the steady increase of height of ships' decks above sea-level (Folland *et al.*, 1984; Bottomley *et al.*, 1990); probable nonstandard observing practices during the Second World War yielding excessively high temperatures (Folland *et al.*, 1984); poor exposure owing to cargo stacked on deck to avoid tariffs following the opening of the Suez Canal in 1869 (Bottomley *et al.*, 1990) yielding unrealistically high NMAT relative to adjusted SST in the Mediterranean and northern Indian Ocean between 1876 and 1893; and possible non-standard observing practices also giving too-high NMAT relative to adjusted SST in the Atlantic before the mid-1880s (Bottomley *et al.*, 1990).

Using boundary-layer similarity theory (Large and Pond, 1982), along with information on historical changes in the heights of ships' decks, Folland *et al.* (1984) and Bottomley *et al.* (1990) derived adjustments to compensate for the height changes. Bottomley *et al.* (1990) also adjusted the NMATs for the Second World War, using day marine air temperatures (DMATs) as an approximate guide because these appeared to be more homogeneous. However, the coarse geographical resolution of these adjustments appears to be inadequate in some areas, leaving a residual warm NMAT peak which appears to be unrealistic in the light of trends of collocated SST (Folland and Salinger, 1995; Parker *et al.*, 1995). Adjustments with a higher geographical resolution have subsequently been calculated.

Finally, the above mentioned nineteenth century biases in NMATs were not found to be corrigible without reference to adjusted (Folland and Parker, 1995) SST anomalies, which were therefore used either as straight replacements (Mediterranean and northern Indian Ocean, 1876–1893) or to calculate adjustments which forced the local 30-year average NMAT anomaly in a given calendar month to equal the corresponding average adjusted SST anomaly (Atlantic, 1856–1885; Bottomley *et al.*, 1990; Parker *et al.*, 1995). These procedures are regarded as valid because of good similarity between anomalies of NMAT and SST in subsequent years, and the unlikelihood of substantial systematic changes in air-sea heat fluxes over decades (Barnett, 1984). However, these adjustments mean that for these regions and periods, the corroborative value of the NMATs is lost.

The adjustments applied to NMAT by Bottomley *et al.* (1990) and Parker *et al.* (1995) are, on the other hand, corroborated by the agreement between trends of adjusted NMAT and coastal and island air temperature (Folland and Salinger, 1995; Parker *et al.*, 1995). In an alternate approach, Jones *et al.* (1986a) used anomalies of regional, mainly coastal, land surface air temperature to adjust anomalies of nearby MAT. This was possible because anomalies of NMAT and nearby 'coastal' land surface air temperature are found to be similar in recent data over periods as long as a decade (Folland and Salinger, 1995; Parker *et al.*, 1995), even though the absolute values differ considerably. Accordingly, their adjustments differed from those of Folland *et al.* (1984) and Bottomley *et al.* (1990).

5. DISCUSSION AND CONCLUSIONS

Making proper homogeneity adjustments can be tedious and exacting work, but accounting for the inhomogeneities that *in situ* climate data often contain is crucial prior to many types of climate analyses. Without proper adjustments, erroneous conclusions may be inevitable in some cases (Figure 6). The difference in trends between homogeneity-adjusted and unadjusted data can be enormous at an individual station and very significant in regional analyses. However, Easterling and Peterson (1995a,b) found that on very large spatial scales (half a continent to global), positive and negative homogeneity adjustments in individual station's maximum and minimum temperature time series largely balance out so when averaged into a single time series, the adjusted and unadjusted trends were similar. This would probably not be true, however, for global-scale averages of frozen precipitation where the increasing use of shielded gauges has increased the mean gauge efficiency. Also, homogeneity adjustments can still be important even when offsetting adjustments do not alter the regional trend. For example, Figure 7 shows a regional curve of precipitation derived from data from 25 Norwegian stations. While the adjusted and unadjusted trends are very similar, the uncertainty of the curve, as measured by the standard deviations (S.D.s) between each of the station curves is considerably decreased when using homogeneous series (Hanssen-Bauer *et al.*, 1997).

Different homogeneity adjustment techniques naturally produce different results. Deciding which technique, variant of a technique, or combination of techniques is best can be difficult and depends on the task and resources available. But often two differently adjusted versions of the same station's data are much more similar to each other than they are to the unadjusted data. For example, Figure 8 reveals the results of U.S. HCN and Jones' adjustment methodologies for Carlsbad, New Mexico and Figure 9 shows

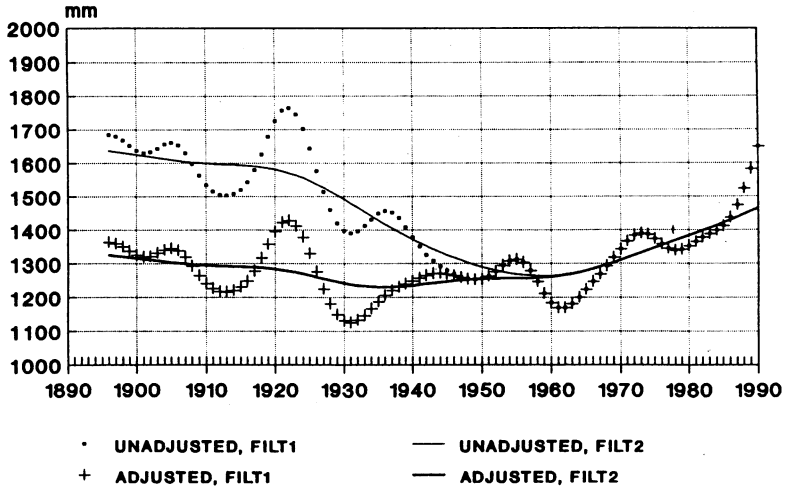


Figure 6. Unadjusted and adjusted time series of annual precipitation at Briksdal in western Norway. The results of climate analyses would be very different using homogeneity-adjusted vs. unadjusted data (from Hanssen-Bauer and Førland, 1994)

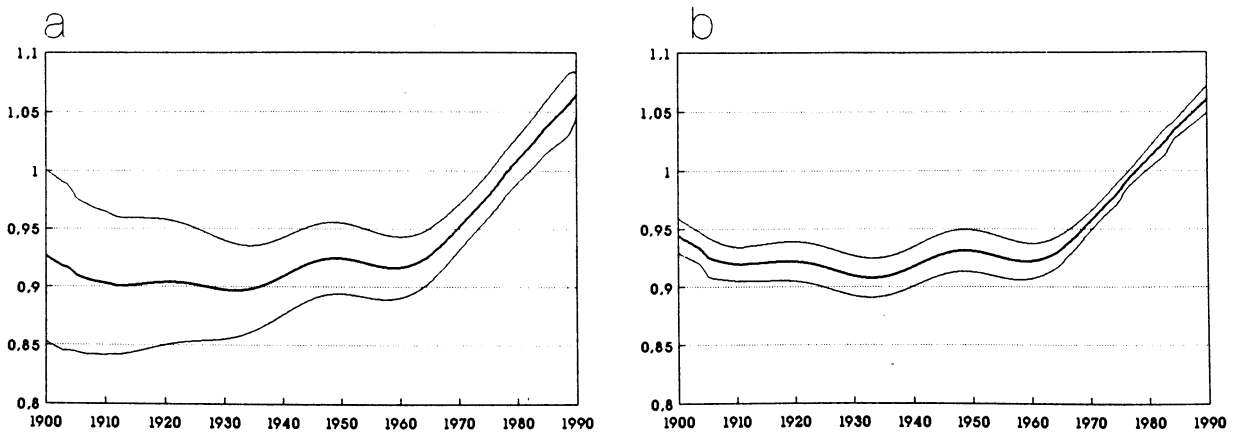


Figure 7. Smoothed trend curves in percent of the 1961–1990 precipitation normal from 25 Norwegian stations (Hanssen-Bauer *et al.*, 1995). While the trends remain the same, the uncertainty, as calculated by the S.D. of the 25 stations, is much larger in the raw data (a) than the homogeneity-adjusted data (b)

a comparison between U.S. HCN and GHCN methodologies for Spokane, Washington. In both cases, the most significant inhomogeneities were detected by the different methodologies at the same time and assessed at similar magnitudes.

While different adjustment methodologies may produce similar results and improve the reliability of the data for many uses, a homogeneity-adjusted time series is not the same as a pristine time series that is homogeneous without needing adjustments. Many approaches can only make average adjustments. For example, when a new instrument is introduced, one may be able to determine the average bias this change introduces but not be able to determine the exact bias at a particular station given its siting and microclimate. Also, techniques that utilize data from neighbouring stations must, by their very nature, introduce a regional climate signal into the individual station time series. Therefore, the most appropriate use of homogeneity-adjusted data—one that takes full advantage of the benefits of homogeneity adjustments and is not adversely impacted by regionality inherent in most adjustment methodologies—is in creating and analyzing area-averaged time series (Easterling *et al.*, 1996b).

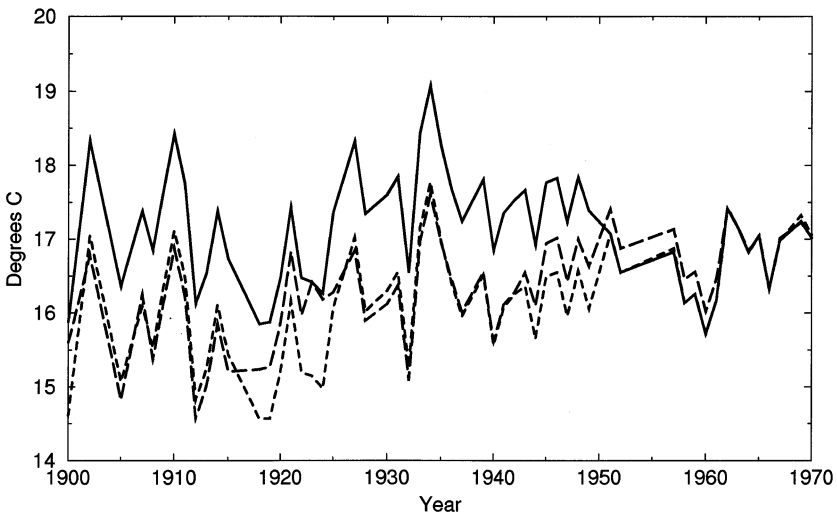


Figure 8. Annual temperature time series for Carlsbad, New Mexico. Solid line is unadjusted data. Long dash is the Jones homogeneity-adjusted data and the short dash is the U.S. HCN adjusted data for the same station. Despite very different methodologies and perhaps some differences in the source data, the long-term trends for the two adjusted time series are in good agreement

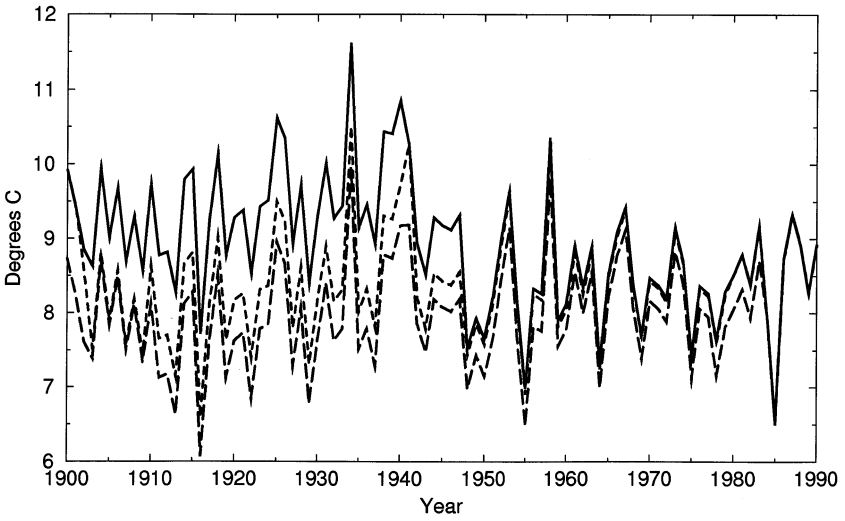


Figure 9. Annual temperature time series for Spokane, Washington. Solid line is unadjusted data. The short dash is the U.S. HCN adjusted version and the long dash is an adjusted time series created with the GHCN adjustment methodology. Again, despite very different adjustment methodologies, trends in the two adjusted time series are in good agreement. Note that the metadata in Figure 1 indicates that the station moved 17 km horizontally and 123 m vertically on 8 December 1947

Because of this regionalization effect, there are some analyses where unadjusted data are preferred. These are often station specific studies that do not involve long-term trend analysis. Therefore, it is important to preserve the original data as well as homogeneity-adjusted versions. Also, original data need to be preserved because new and better approaches to homogeneity adjustments will probably be developed in the near future. During the last decade, considerable work has been done on homogeneity testing and data adjustments and research will continue in this field. Future work includes improving adjustment methodologies, investigating adjustments of daily data, and evaluating the impact of adjustments on extreme values. With continuing efforts to put current climate variability, change, and extreme events into historical perspective, the need for reliable, homogeneous climate data sets will increase.

REFERENCES

Alexandersson, H. 1986. 'A homogeneity test applied to precipitation data', *J. Climate*, **6**, 661–675.

Alexandersson, H. 1994. *Climate Series-a Question of Homogeneity*, The 19th Nordic Meteorological Meeting, Kristiansand, Norway, DNMI-preprint, pp. 25–31.

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

Alter, J.C. 1937. 'Shielded storage precipitation gauges', *Mon. Wea. Rev.*, **65**, 262–265.

Auer, I., 1992. *Experiences with the Completion and Homogenization of Long-term Precipitation Series in Austria*, Centr. Europ. research. initiative, Proj. Gr. Meteorology, Wp. 1, Vienna.

Auer, I. and Boehm, R. 1994a. 'Recent climatological time series investigations in Austria–an overview', in Heino, R. (ed.), *Climate Variations in Europe*, Proceedings of the European Workshop held in Kirkkonummi (Majvik), Finland 15–18 May 1994, Publications of the Academy of Finland 3/94, 103–117.

Auer, I. and Boehm, R. 1994b. 'Combined temperature-precipitation variations in Austria during the instrumental period', *Theor. Appl. Climatol.*, **49**, 161–174.

Barnett, T.P. 1984. 'Long-term trends in surface temperature over the oceans', *Mon. Wea. Rev.*, **112**, 303–312.

Boehm, R. 1992. *Description of the Procedure of Homogenizing Temperature Time Series in Austria*, Centr. Europ. research. initiative, Proj. Gr. Meteorology, Wp. 2, Vienna.

Bottomley, M., Folland, C.K., Hsiung, J., Newell, R.E. and Parker, D.E. 1990. *Global Ocean Surface Temperature Atlas (GOSTA)*, Joint Meteorological Office and Massachusetts Institute of Technology Project, HMSO, London, 20 + iv pp. and 313 plates.

Bradley, R.S., Kelly, P.M., Jones, P.D., Goodess C.M., and Diaz, H.F. 1985. *A Climatic Data Bank for Northern Hemisphere Land Areas, 1851–1980*, TRO17, Depertment of Energy, Washington, 335 pp.

Caussinus, H. and Lyazrhi, F. 1997. 'Choosing a linear model with a random number of change-points and outliers', *Ann. Inst. Stat. Math.*, **49**, 761–775.

Caussinus, H. and Mestre, O. 1996. 'New mathematical tools and methodologies for relative homogeneity testing', *Proceedings of the Seminar for Homogenization of Surface Climatological Data*, Budapest, 6–12 October, pp. 63–82.

Conrad, V. and Pollak, C. 1950. *Methods in Climatology*, Harvard University Press, Cambridge, MA, 459 pp.
Craddock, J.M. 1979. 'Methods of comparing annual rainfall records for climatic purposes', *Weather*, **34**, 332–346.
Easterling, D.R., Karl, T.R., Mason, E.H., Hughes, P.Y. and Bowman, D.P. 1996a. *United States Historical Climatology Network (U.S. HCN) Monthly Temperature and Precipitation Data*, Environmental Sciences Division, Pub. No. 3404, Carbon Dioxide Information and Analysis Center, Oak Ridge National Laboratory, 262 pp.
Easterling, D.R., Peterson, T.C. and Karl, T.R. 1996b. 'On the development and use of homogenized climate data sets', *J. Climate*, **9**, 1429–1434.
Easterling, D.R. and Peterson, T.C. 1995a. 'A new method for detecting and adjusting for undocumented discontinuities in climatological time series', *Int. J. Climatol.*, **15**, 369–377.
Easterling, D.R. and Peterson, T.C. 1995b. 'The effect of artificial discontinuities on recent trends in minimum and maximum temperatures', *Atmos. Res.*, **37**, 19–26.
Folland, C.K. and Salinger, M.J. 1995. 'Surface temperature trends and variations in New Zealand and the surrounding ocean, 1871–1993', *Int. J. Climatol.*, **15**, 1195–1218.
Folland, C.K. and Parker, D.E. 1995. 'Correction of instrumental biases in historical sea surface temperature data', *Q. J. R. Meteorol. Soc.*, **121**, 319–367.
Folland, C.K., Parker, D.E. and Kates, F.E. 1984. 'Worldwide marine surface temperature fluctuations 1856–1981', *Nature*, **310**, 670–673.
Førland, E.J. 1994. 'Trends and problems in Norwegian snow records', in Heino, R. (ed.), *Climate Variations in Europe*, Proceedings of the European Workshop held in Kirkkonummi (Majvik), Finland 15–18 May 1994. Publications of the Academy of Finland 3/94, 205–215.
Førland, E.J., Allerup, P., Dahlstrøm, B., Elomaa, E., Jonsson, T., Madsen, H., Per, J., Rissanen, P., Vedin, H. and Vejen, F. 1996. *Manual for Operational Correction of Nordic Precipitation Data*, DNMI-Reports 24/96 KLIMA, 66 pp.
Frich, P., Alexandersson, H., Ashcroft, J., Dahlström, B., Demarée, G., Drebs, A., Van Engelen, A., Førland, E.J., Hanssen-Bauer, I., Heino, R.T., Jonasson, K., Keegan, L., Nordli, P.Ø., Schmith, T., Steffensen, P., Tuomenvirta, H., and Tveito, O.E., 1996. *North Atlantic Climatological Dataset (NACD Version 1)–Final Report*, Danish Meteorological Institute, Scientific Report, 96, 147 pp.
Golubev, V.S., Koknaeva, V.V., Simonenko, A. Yu. 1995. 'Results of atmospheric precipitation measurements by national standard gauges of Canada, USA, and Russia', *Rus. Meteorol. Hydrol*, **2**, 102–110.
Groisman, P. Ya and Legates, D.R. 1994. 'The accuracy of United States precipitation data', *Bull. Am. Meteorol. Soc.*, **75**, 215–227.
Groisman, P. Ya, Easterling, D.R., Quayle, R.G., Golubev, V.S., Krenke, A.N., and Mikhailov, A. Yu. 1996. 'Reducing biases in estimates of precipitation over the United States: Phase 3 adjustments', *J. Geophys. Res.*, **101**, 7185–7195.
Gullett, D.W., Vincent, L. and Sajecki, P.J.F. 1990. *Testing for Homogeneity in Temperature Time Series at Canadian Climate Stations*, CCC Report No. 90-4, Atmospheric Environment Service, Downsview, Ontario. 43 pp.
Gullett, D.W., Vincent, L. and Malone, L.H. 1991. *Homogeneity Testing of Monthly Temperature Series. Application of Multiple-Phase Regression Models with Mathematical Changepoints*, CCC Report No. 91–10. Atmospheric Environment Service, Downsview, Ontario. 47 pp.
Hanssen-Bauer, I., Førland, E.J. and Nordli, P.Ø. 1991. *Homogeneity Test of Precipitation Data. Description of the Methods Used at DNMI*, DNMI-Report KLIMA13/91, 23 pp.
Hanssen-Bauer, I., and Førland, E.J. 1994. 'Homogenizing long Norwegian precipitation series', *J. Climate*, **7**, 1001–1013.
Hanssen-Bauer, I., Førland, E.J. and Tveito, O.E. 1995. *Trends and Variability in Annual Precipitation in Norway*, DNMI-Report KLIMA27/95, 25 pp.
Hanssen-Bauer, I., Førland, E.J., Tveito, O.E. and Nordli, P.Ø. 1997. 'Estimating regional precipitation trends—comparison of two methods', *Nordic Hydrol.*, **28**, 21–36.
Heino, R. 1994. *Climate in Finland During the Period of Meteorological Observations*, Finnish Meteorological Institute Contributions, 12, 209 pp.
Huovila, S., Elomaa, E., Leminen, K., Tammelin, B. and Tuominen, A. 1988. *Comparison of Snow Gauges Used in Nordic Countries-Contribution of Finland to WMO Solid Precipitation Measurement Intercomparison, Part I: System Description*. Finnish Meteorological Institute, Helsinki, 61 pp.
Jones, P.D., Raper, S.C.B., Santer, B., Cherry, B.S.B., Goodess, C., Kelly, P.M., Wigley, T.M.L., Bradley, R.S., and Diaz, H.F., 1985. *A Grid Point Surface Air Temperature Data Set for the Northern Hemisphere*, TRO22, Department of Energy, Washington, 251 pp.
Jones, P.D., Raper, S.C.B. Bradley, R.S. Diaz, H.F. Kelly, P.M. and Wigley, T.M.L. 1986a. 'Northern Hemisphere Surface Air Temperature Variations: 1851–1984', *J. Climate Appl. Meteorol.*, **25**, 161–179.
Jones, P.D., Wigley, T.M.L. and Wright, P.B. 1986b. 'Global temperature variations between 1861 and 1984', *Nature*, **322**, 430–434.
Jones, P.D., Raper, S.C.B., Goodess, C.M., Cherry, B.S.G., and Wigley, T.M.L. 1986c. *A Grid-point Surface Air Temperature Data Set for the Southern Hemisphere*, TRO27, Department of Energy, Washington, 73 pp.
Jones, P.D. 1994. 'Hemispheric surface air temperature variations: A reanalaysis and an update to 1993', *J. Climate*, **7**, 1794–1802.
Karl, T.R., Williams, C.N. Jr., Young, P.J. and Wendland, W.M. 1986. 'A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States', *J. Climate Appl. Meteorol.*, **25**, 145–160.
Karl, T.R., and Williams, C.N. Jr., 1987. 'An approach to adjusting climatological time series for discontinuous inhomogeneities', *J. Climate Appl. Meteorol.*, **26**, 1744–1763.
Karl, T.R., Diaz, H.F. and Kukla, G. 1988. 'Urbanization: Its detection and effect in the United States climate record', *J. Climate*, **1**, 1099–1123.
Kohler, M.A. 1949. 'Double-mass analysis for testing the consistency of records and for making adjustments', *Bull. Amer. Meteorol. Soc.*, **30**, 188–189.
Lanzante, J.R. 1996. 'Resistant, robust and nonparametric techniques for the analysis of climate data. Theory and examples, including applications to historical radiosonde station data', *Int. J Climatol.*, **16**, 1197–1226.
Large, W.G. and Pond, S. 1982. 'Sensible and latent heat flux measurements over the ocean', *J. Phys. Oceanogr.*, **12**, 464–482.

Larson, L.W. and Peck, E.L. 1974. 'Accuracy of precipitation measurements for hydrologic modelling', *Water Resour. Res.*, **10**, 857–863.
Mekis, E, and Hogg, W.D. 1997. *Rehabilitation and analysis of Canadian daily precipitation time series*, Proceedings of the 10th Conference on Applied Climatology, Reno, Nevada, October, 300–308.
Mielke, P.W. 1991. 'The application of multivariate permutation methods based on distance functions in the earth sciences', *Earth-Sci. Rev.*, **31**, 55–71.
Nicholls, N., Tapp, R., Burrows, K. and Richards, D. 1996. 'Historical thermometer exposures in Australia', *Int. J. Climatol.*, **16**, 705–710.
Nordli, P.Ø., Alexandersson, H., Frich, P., Førland, E.J., Heino, R., Jonsson, T., Steffensen, P., Tuomenvirta, H. and Tveito, O.E. 1997. 'The effect of radiation screens on Nordic time series of mean temperature', *Int. J. Climatol.*, **17**, 1667–1681.
Nordli, P.Ø., Hanssen-Bauer, I. and Førland, E.J. 1996. *Homogeneity Analyses of Temperature and Precipitation Series from Svalbardog and Jan Mayen*, DNMI-Report KLIMA16/96 ISSN 0805–9918, 41 pp.
Panofsky, H.A. and Brier, G.W. 1968. *Some Applications of Statistics to Meteorology*, Pennsylvania State University, University Park, 224 pp.
Parker, D.E., Folland, C.K. and Jackson, M. 1995. 'Marine surface temperature: Observed variations and data requirements', *Climatic Change*, **31**, 559–600.
Peterson, T.C., and Easterling, D.R. 1994. 'Creation of homogeneous composite climatological reference series', *Int. J. Climatol.*, **14**, 671–679.
Peterson, T.C. and Griffiths, J.F. 1996. 'Colonial era archive data project', *Earth Syst. Monit.*, **6**, 8–16.
Peterson, T.C. and Vose, R.S. 1997. 'An overview of the Global Historical Climatology Network temperature data base', *Bull. Amer. Meteorol. Soc.*, **78**, 2837–2849.
Plummer, N., Lin, Z. and Torok, S. 1995. 'Trends in the diurnal temperature range over Australia since 1951', *Atmos. Res.*, **37**, 79–86.
Potter, K.W. 1981. 'Illustration of a new test for detecting a shift in mean in precipitation series', *Mon. Wea. Rev.*, **109**, 2040–2045.
Quayle, R.G., Easterling, D.R.. Karl, T.R. and Hughes, P.Y. 1991. 'Effects of recent thermometer changes in the cooperative station network', *Bull. Amer. Meteorol. Soc.*, **72**, 1718–1724.
Riesbol, H.S. 1938. 'Results from the experimental rain gages at Coshocton, Ohio', *Trans. AGU Hydrol.*, 542–550.
Rhoades, D.A., and Salinger, M.J. 1993. 'Adjustment of temperature and rainfall records for site changes', *Int. J. Climatol.*, **13**, 899–913.
Salinger, M.J., Basher, R.E., Fitzharris, B.B., Hay, J.E., Jones, P.D., MacVeigh, J.P. and Schmidelly-Leleu, I. 1995. 'Climate trends in the South West Pacific', *Int. J. Climatol.*, **15**, 285–302.
Salinger, M.J., Allan, R., Bindoff, N., Hannah, J., Lavery, B., Lin, Z., Leleu, I., Lindesay, J., MacVeigh, J.P., Nicholls, N., Plummer, N. and Torok, S. 1996. 'Observed variability and change in climate and sea-level in Oceania', in Bouma, W.J. and Pearman, G. and Manning, M. (eds), *Greenhouse: Coping with Climate Change*, pp. 100–126.
Sevruk, B. 1982. *Methods of Correction for Systematic Error in Point Precipitation Measurements for Operational Use*, Oper. Hydrol. Rep. 21, Publ. 589, World Meteorological Organization, Geneva, 91 pp.
Seward, D. and Rhoades, D.A. 1986. 'A clustering technique for fission track dating of fully to partially annealed minerals and other non-unique populations', *Nucl. Tracks. Radiat. Meas.*, **11**, 259–268.
Siegel, S. and Castellan, N. 1988. *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York, 399 pp.
Solow, A. 1987. 'Testing for climatic change: an application of the two-phase regression model', *J. Climate Appl. Meteorol.*, **26**, 1401–1405.
Szentimrey, T. 1994. 'Statistical problems connected with the homogenization of climatic time series', in Heino, R. (ed.), *Climate Variations in Europe*, Proceedings of the European Workshop held in Kirkkonummi (Majvik), Finland 15–18 May 1994. Publications of the Academy of Finland 3/94, pp. 330–339.
Szentimrey, T. 1995. 'General problems of the estimation of inhomogeneities, optimal weighting of the reference stations', *Proceedings of the 6th International Meeting on Statistical Climatology*, Galway, Ireland, pp. 629–631.
Szentimrey, T. 1996. 'Statistical procedure for joint homogenization of climatic time series', *Proceedings of the Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary, pp. 47–62.
Torok, S. and Nicholls, N. 1996. 'An historical annual temperature dataset for Australia', *Aust. Meteorol. Mag.*, **45**, 251–260.
Tuomenvirta, H. and Heino, R. 1996. 'Climatic changes in Finland—recent findings', *Geophysica*, **32**, 61–75.
Tuomenvirta, H. and Alexandersson, H. 1995. 'Adjustment of apparent changes in variability of temperature time series', *Proceedings of the Sixth International Meeting on Statistical Climatology*, Galway, Ireland, 19–23 June 1995, pp. 443–446.
Tuomenvirta, H. and Drebs, A. 1994. 'Homogeneity testing and management of metadata in Finland', in Heino, R. (ed.), *Climate Variations in Europe*, Proceedings of the European Workshop held in Kirkkonummi (Majvik), Finland 15–18 May 1994, Publications of the Academy of Finland 3/94, pp. 321–329.
Vincent, L. 1998. 'A technique for the identification of inhomogeneities in Canadian temperature series', *J. Climate*, **11**, 1094–1104.
Vincent, L. 1990. *Time Series Analysis: Testing the Homogeneity of Monthly Temperature Series*, Survey Paper No. 90-5, Department of Mathematics and Statistics, York University, Ontario, 50 pp.
Young, K.W. 1993. 'Detecting and removing inhomogeneities from long-term monthly sea level pressure time series', *J. Climate*, **6**, 1205–1220.
Zurbenko, I, Porter, P.S., Rao, S.T., Ku, J.Y., Gui, R., Eskridge, R.E. 1996. 'Detecting discontinuities in time series of upper air data: Development and demonstration of an adaptive filter technique', *J. Climate*, **9**, 3548–3560.

