INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **23**: 679–692 (2003)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/joc.906

HOMOGENEITY OF 20TH CENTURY EUROPEAN DAILY TEMPERATURE AND PRECIPITATION SERIES

J. B. WIJNGAARD,* A. M. G. KLEIN TANK and G. P. KÖNNEN
Royal Netherlands Meteorological Institute, PO Box 201, 3730 AE De Bilt, The Netherlands

Received 6 August 2002
Revised 17 February 2003
Accepted 17 February 2003

ABSTRACT

Daily European station series (1901–99) of surface air temperature and precipitation from the European Climate Assessment dataset are statistically tested with respect to homogeneity. A two-step approach is followed. First, four homogeneity tests are applied to evaluate the daily series. The testing variables used are (1) the annual mean of the diurnal temperature range, (2) the annual mean of the absolute day-to-day differences of the diurnal temperature range and (3) the wet day count (threshold 1 mm). Second, the results of the different tests are condensed into three classes: 'useful', 'doubtful' and 'suspect'. A qualitative interpretation of this classification is given, as well as recommendations for the use of these labelled series in trend analysis and variability analysis of weather extremes. In the period 1901–99, 94% of the temperature series and 25% of the precipitation series are labelled 'doubtful' or 'suspect'. In the sub-period 1946–99, 61% of the temperature series and 13% of the precipitation series are assigned to these classes. The seemingly favourable scores for precipitation can be attributed to the high standard deviation of the testing variable, and hence the inherent restricted possibilities for detecting inhomogeneities. About 65% of the statistically detected inhomogeneities in the temperature series labelled 'doubtful' or 'suspect' in the period 1946–99 can be attributed to observational changes that are documented in the metadata. For precipitation this percentage is 90%. Copyright © 2003 Royal Meteorological Society.

KEY WORDS: daily dataset; Europe; homogeneity; temperature; precipitation; trend analysis

1. INTRODUCTION

Within the framework of the European Climate Assessment (ECA) project, a dataset of daily observations of temperature (mean, minimum and maximum) and precipitation was compiled (Klein Tank *et al.*, 2002). The dataset covers the period 1901–99, and is meant to analyse trends and variability in weather extremes in the 20th century climate of Europe and the Middle East (Klein Tank and Können, 2003). Clearly, this type of analysis is limited by the degree of inhomogeneity of the data. Inhomogeneities in station data records are often caused by changes in observational routines, among which are station relocations, changes in measuring techniques and changes in observing practices. Exploiting datasets like those in the ECA requires a continuous monitoring of their quality with a well-defined methodology to support the credibility of the conclusions gained from them.

Numerous methods are in use to evaluate the homogeneity of monthly to annual-resolution climatic time series. Peterson *et al.* (1998) gives a comprehensive overview. Szalai (1997) and Szalai *et al.* (1999) summarize the application of different methods that are used in European countries. Generally, a combination of statistical methods and methods relying on metadata information is considered to be most effective to track down inhomogeneities.

* Correspondence to: J. B. Wijngaard, Royal Netherlands Meteorological Institute, PO Box 201, 3730 AE De Bilt, The Netherlands; e-mail: Janet.Wijngaard@knmi.nl

In contrast to monthly or annual-resolution series, well-established statistical methods for testing the homogeneity of daily resolution series are lacking. Given this situation, we developed a hybrid method by compiling from the daily set an annual resolution set of variables representing important characteristics of variation at the daily scale, and then applying to these testing variables an appropriate selection of traditional tests developed for monthly and annual series. The power of this hybrid method in detecting inhomogeneities that originate from the short time scales depends critically on the choice of the testing variables.

In general, it is recommended (e.g. Peterson *et al*., 1998) to apply homogeneity tests relatively, i.e. testing with respect to a neighbouring station that is supposedly homogeneous. If the two series are sufficiently correlated, relative tests are considered to be more powerful than absolute tests, which use only the single station series. Inhomogeneities are thus more easily distinguished from real climate variations. Although relative tests are not able to deal with simultaneous changes in the observational routines of both the test and reference station, relative testing is often the first option. However, for the ECA dataset, relative testing is not appropriate, given the sparse spatial density of the station network. Therefore, we currently restrict the homogeneity analysis to absolute tests, having the additional advantage of being able to deal with simultaneous changes in observational routines.

In the current assessment of the homogeneity of the daily temperature and precipitation series, a two-step approach is followed. First, the hybrid method using four established statistical tests of partly complementary properties is applied to the ECA dataset to identify potential inhomogeneities in annual resolution testing variables that are representative for the daily resolution series. For temperature, two variables related to the diurnal cycle are tested; for precipitation, the wet day count (threshold 1 mm) is tested. Second, the results of the tests for a given series (temperature or precipitation) are grouped in an overall classification of the reliability of that series. Historic metadata information is used to find supporting evidence of changes in observational routines that may have triggered the inhomogeneities detected. We believe that the current evaluation provides insight into the potential of the series. The overall classification forms a fair basis to select series appropriate for trend analysis and variability analysis. The need remains to revisit the series at a later stage with tests dedicated to daily resolution series.

2. STATISTICAL METHODS

The four test methods selected to test the departure of homogeneity in the time series are: the standard normal homogeneity test (SNHT) for a single break (Alexandersson, 1986), the Buishand range test (Buishand, 1982), the Pettitt test (Pettitt, 1979), and the Von Neumann ratio test (Von Neumann, 1941).

All four tests suppose under the null hypothesis that the annual values Y_i of the testing variable Y are independent and identically distributed. Under the alternative hypothesis, the SNHT, the Buishand range and the Pettitt test assume that a step-wise shift in the mean (a break) is present. These three tests are capable of locating the year where a break is likely. In this paper these three tests are referred to as *location-specific* tests. The fourth test, the Von Neumann ratio test, assumes under the alternative hypothesis that the series is not randomly distributed. This test is not location specific, which means that it does not give information on the year of the break.

Although the three location-specific tests have many characteristics in common, they are also different. The SNHT detects breaks near the beginning and the end of a series relatively easily, whereas the Buishand range and the Pettitt test are more sensitive to breaks in the middle of a time series (Hawkins, 1977). The SNHT and the Buishand range test assume that the Y_i values are normally distributed, but the Pettitt test does not. The reason that the Pettitt test does not require such an assumption is that this test is based on the ranks of the elements of a series rather than on the values themselves. The ranking approach of the Pettitt test also implies that it is less sensitive to outliers than the other tests. The non-location-specific Von Neumann ratio test is complementary to the other three tests because of its sensitivity to departures of homogeneity that are of a nature other than strict step-wise shifts (Buishand, 1981, 1982). Appendix A gives the mathematical formulation of the four tests and tables with critical values for their test statistics.

3. OVERALL CLASSIFICATION AND ITS QUALITATIVE INTERPRETATION

The second step in the homogeneity assessment of the ECA dataset is an overall evaluation of the four tests. The outcomes of the four tests for each temperature and precipitation series are grouped together. As in Schönwiese and Rapp (1997), a classification is made depending on the number of tests rejecting the null hypothesis. The following categories are distinguished:

Class 1: ‘useful’ — one or zero tests reject the null hypothesis at the 1% level.
Class 2: ‘doubtful’ — two tests reject the null hypothesis at the 1% level.
Class 3: ‘suspect’ — three or four tests reject the null hypothesis at the 1% level.

For temperature, where two variables are tested, the two categories are calculated separately for each variable. If the results are different, then the highest of the two category values (hence the least favourable) is assigned to the temperature series of the station.

The qualitative interpretation of the categories is as follows:

Class 1: ‘useful’. No clear signal of an inhomogeneity in the series is apparent. Hence, inhomogeneities that may be present in the series are sufficiently small with respect to the inter-annual standard deviation of the testing variable series that they will largely escape detection. The series seems to be sufficiently homogeneous for trend analysis and variability analysis.

Class 2: ‘doubtful’. Indications are present of an inhomogeneity of a magnitude that exceeds the level expressed by the inter-annual standard deviation of the testing variable series. The results of trend analysis and variability analysis should be regarded very critically from the perspective of the existence of possible inhomogeneities.

Class 3: ‘suspect’. It is likely that an inhomogeneity is present that exceeds the level expressed by the inter-annual standard deviation of the testing variable series. Marginal results of trend and variability analysis should be regarded as spurious. Only very large trends may be related to a climatic signal.

It is clear that series labelled class 3 ‘suspect’ lack credibility. These series should not be used in the analysis of trend and variability analysis. Only in exceptional cases, where it is plausible that a real climatic signal rather than an artificial break triggered the test results (see the example in Section 6.3), can the original station series be used in further analysis.

One strategy to improve the classification of a series is to analyse the metadata in order to correct for documented breaks. In our analysis of trends in temperature and precipitation extremes over Europe (Klein Tank and Können, 2003) we did not elaborate upon this strategy, but adopted an alternative strategy instead, i.e. straightforwardly excluding from the analysis all series labelled ‘suspect’.

4. ECA TEST PERIOD AND TESTING VARIABLES

The period of analysis is the complete period of the ECA dataset: 1901–99. However, as the station coverage increases with time (Klein Tank *et al.*, 2002), test results are also presented for the sub-period 1946–99. Three testing variable series are selected and derived from the daily ECA series, each consisting of annual values. For temperature, the testing variables are two characteristics of the diurnal temperature range (DTR): its annual mean (mDTR) and its annual mean of the absolute day-to-day differences (vDTR). For precipitation, the annual number of wet days (threshold 1 mm) is considered.

The motivation for choosing mDTR and vDTR, rather than the underlying annual means of minimum and maximum temperature, is that the DTR variables are often more sensitive to the homogeneity tests. A reason is that breaks due to station relocations and changes in measuring techniques are usually radiation-related, with different effects on minimum and maximum temperature. Consequently, a break that appears clearly in the DTR-based testing variable series may be only weakly apparent in the minimum and maximum temperature series, and might even be totally masked in the annual mean temperature series (see also Sparks (1972) and Heino *et al.* (1999)).

The choice of DTR variables as testing variables instead of minimum, maximum or mean temperature causes our absolute testing to exhibit to a certain extent the properties inherent to relative testing. This includes the improved power of relative testing and its capacity to discriminate between real and spurious climate variations, as well as the well-known weakness of relative tests for simultaneous changes in the observational routines.

In addition to the annual mean DTR (mDTR), we also introduced the annual mean of the absolute day-to-day differences of the DTR (vDTR) as the testing variable for temperature. The motivation is that inhomogeneities may manifest themselves more strongly in a break in a higher order statistical moment, e.g. the standard deviation, than in a break in the mean. We did not use the standard deviation of the DTR as the testing variable but instead chose to use the annual mean of absolute day-to-day differences of the DTR. The latter is preferred because it is less influenced by low-frequency variability than the standard deviation (Karl *et al.*, 1995), and because day-to-day temperature differences are particularly sensitive to inhomogeneities (Moberg *et al.*, 2000). Note that vDTR itself is a testing variable whose nature is linked to the daily character of the underlying series. The precise definition of vDTR is

$$\mathrm{vDTR} = \frac{1}{M-1}\sum_{i=2}^{M}|\mathrm{DTR}_i - \mathrm{DTR}_{i-1}|$$

where DTR_i is the diurnal temperature range for day i in a specific year, and M is the number of days in that year.

The inherently high variability of precipitation makes detection of breaks in precipitation series possible only if the breaks are relatively large. The higher the standard deviation of the testing variable, then the more inhomogeneities will escape detection. However, the other side of the coin is that trends (and long-term variations) in series become harder to detect in more variable signals. In that sense, the outcomes of the homogeneity tests remain consistent with the capabilities of trend detection in such series.

As the testing variable for precipitation, the annual precipitation amount and the number of wet days in a year are potential candidates, whereas the day-to-day variability is not suitable because of the large number of dry days in any series. In the present study, we chose the number of wet days as the testing variable for precipitation, rather than the annual amounts. Wet day counts generally have a lower variability than series of annual amounts, particularly in areas with a large contribution from convective precipitation. As a result, inhomogeneities are easier to detect in wet day count series than annual amount series.

In the present study, a wet day is defined as a day with precipitation above 1 mm. Wet day count series based on threshold values higher than 1 mm usually have a higher variability from year-to-year than wet days with a threshold of 1 mm. On the other hand, thresholds lower than 1 mm can lead to a dominance of inhomogeneities caused solely by errors in measuring very low amounts, including dew. Over-detection of such inhomogeneities should be avoided, as they hardly affect the homogeneity of other aspects of the precipitation series, like extremes or series of cumulative precipitation. This is the reason for our choice of the 1 mm threshold in the definition of wet days as the testing variable for the precipitation series in the ECA.

5. ILLUSTRATION OF THE TEST APPLICATION

5.1. Temperature series of Eelde (The Netherlands)

To illustrate the capacity of the homogeneity tests for temperature series, the results of the four tests applied to the ECA station series of Eelde (The Netherlands) are discussed.

Figure 1 shows the annual mean temperature and mDTR at station Eelde. Three well-documented changes in observational routine contaminate the homogeneity of this station series during the last century: the

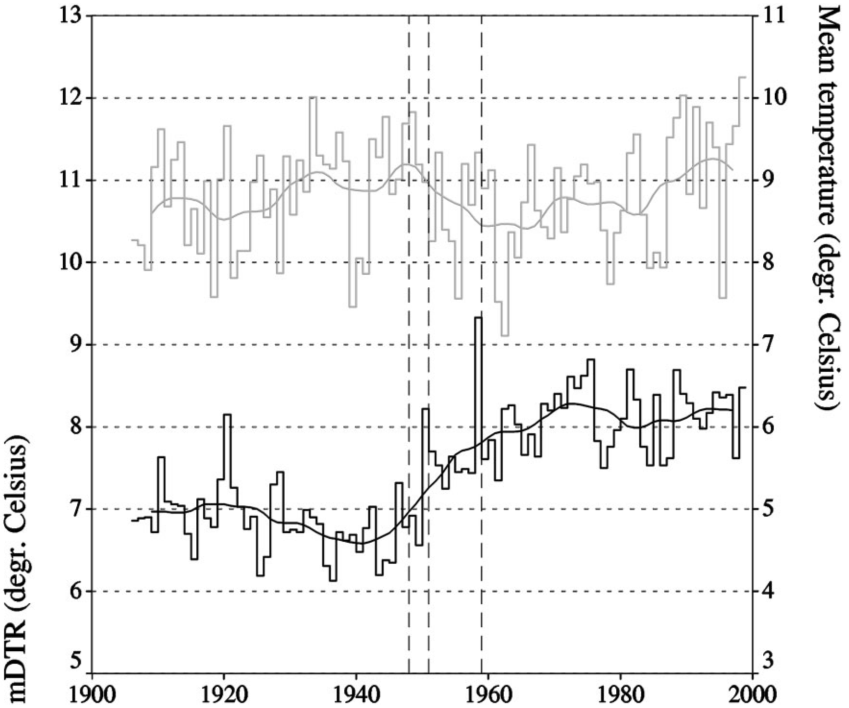

Figure 1. Black: annual mean of the diurnal temperature range (mDTR) at station Eelde (The Netherlands). Grey: annual mean temperature of the station. The smoothed curves are calculated using the Loess smoother (Cleveland, 1979) with a time span of 15 years. Station changes that are documented in the station history are indicated by the vertical dashed lines

introduction of a ventilated observation hut in 1948, a station relocation from the city to the nearby airport in 1951, and a change in sensor height from 2.2 m to 1.5 m in 1959. Figure 1 shows that the contaminations are hardly visible in the annual mean temperature series, whereas they lead to a 1 °C increase of the mDTR in the 1950s.

In Figure 2, the results of the SNHT, the Buishand range and the Pettitt tests applied to the mDTR of Eelde are shown. The test statistic of the SNHT passes an extreme in 1950, indicative of a break around that year. This maximum causes a rejection of the null hypothesis clearly significant at the 1% level. Hence, the alternative hypothesis, which assumes a break, becomes likely. The same conclusion is drawn from the minima in the graphs of the Buishand range and Pettitt test statistics, both having extremes in 1950. The value of the Von Neumann ratio test, 0.55, also clearly indicates an inhomogeneity at the 1% level. Since all four tests reject the null hypothesis at the 1% level, the temperature series of Eelde is assigned

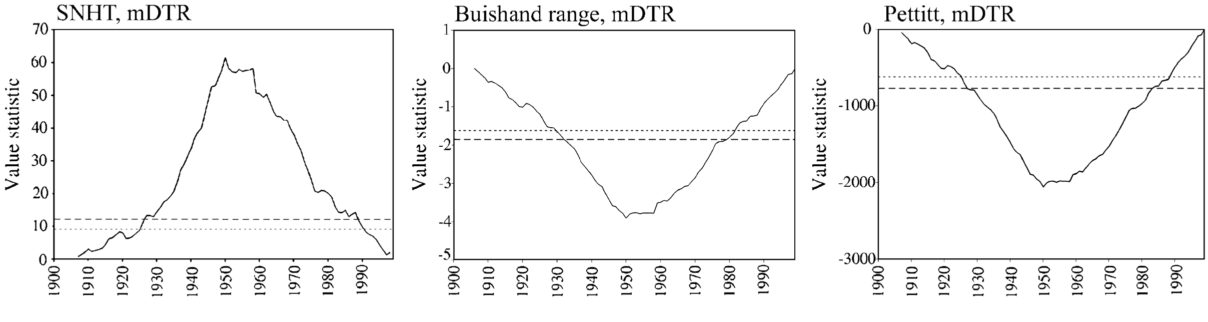

Figure 2. Test results of the SNHT (left), Buishand range (middle) and Pettitt test (right) applied to the mDTR series of station Eelde (The Netherlands). Dashed lines give 1% critical values; dotted lines give 5% critical values

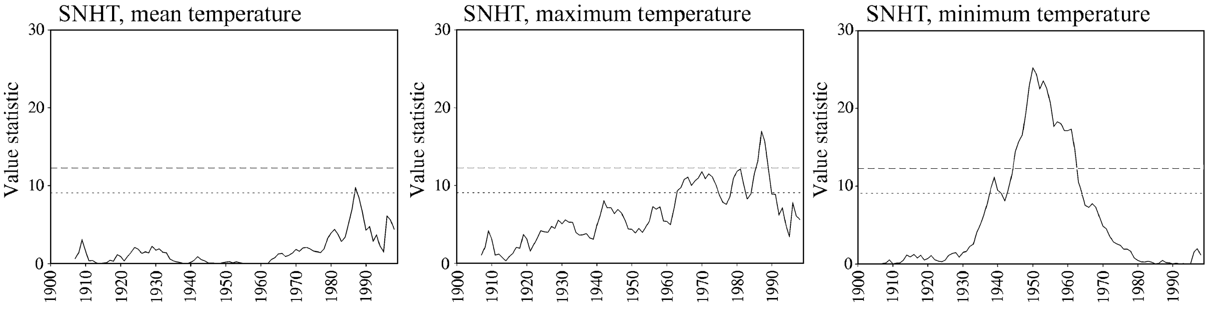

Figure 3. Test results of the SNHT applied to the annual mean (left), maximum (middle) and minimum (right) temperature series of station Eelde (The Netherlands). Dashed lines gives 1% critical values; dotted lines give 5% critical values. The results of the Buishand range test and the Pettitt test (not shown) are similar

to class 3 'suspect'. Similar results are obtained from the vDTR, also indicating that the temperature series of Eelde receives the label 'suspect'. Note that the three changes as documented in the metadata are not separately traced in the three location-specific tests. Since these changes are so close to each other, only the most prominent one is detected. When the 1906–47 sub-period before the documented breaks and the 1960–99 sub-period after the documented breaks are tested separately, no significant breaks are detected (not shown). This is consistent with the metadata, which do not indicate any major changes in these sub-periods.

Figure 3 shows the results of the SNHT test applied to the annual mean, minimum and maximum temperature series of Eelde instead of the DTR testing variable series. The results of the other two location-specific tests are similar. Contrary to the mDTR and vDTR and minimum temperature series, there are hardly any indications for a break around 1950 in the annual series of mean or maximum temperature. Despite the fact that the mean temperature passes the tests successfully in this period, the failure of mDTR and vDTR to pass the tests reminds one to remain suspicious, even with respect to the homogeneity of the daily mean temperatures. This is of particular importance if trends in daily extremes are studied.

Figure 3 shows that the SNHT indicates a homogeneity break in the mean and maximum temperature series around 1987, which is not apparent from the DTR testing variables (Figure 2). This break lacks documentation from the station history and is likely to be unrelated to changes in observational routines. This result, combined with the fact that the homogeneity tests of other Dutch stations show the same feature, indicates that a climate variation rather than homogeneity break triggered the tests.

Summarizing, the Eelde example illustrates two main advantages of the DTR testing variables, namely their power to detect inhomogeneities as indicated by the metadata of the station and their discriminating power between real climate variations and breaks, which is especially important given the strong warming in Europe in the 1990s.

5.2. Precipitation series of Putten (The Netherlands)

Figure 4 shows the annual precipitation amount and the number of wet days (threshold 1 mm) at station Putten (the Netherlands) and Figure 5 shows the results of the Buishand range test applied to the precipitation testing variable: number of wet days. In Figure 5, three definitions of a wet day are used: threshold 0.1, 1 and 10 mm.

In 1950, the height of the Putten gauge was lowered from 1.50 m to 0.40 m. This change is not clearly visible from the annual amounts and the wet day counts (threshold 1 mm) in Figure 4. However, the Buishand range test results for the latter variable (Figure 5, middle panel) clearly detect an inhomogeneity at the 1% level around 1950. As the results for the other three tests are also significant at the 1% level, the series would be labelled 'suspect'. For the 0.1 and 10 mm threshold values, all four tests also indicate inhomogeneity at the 1% level. For all these threshold values, the three location-specific tests point to the correct year for

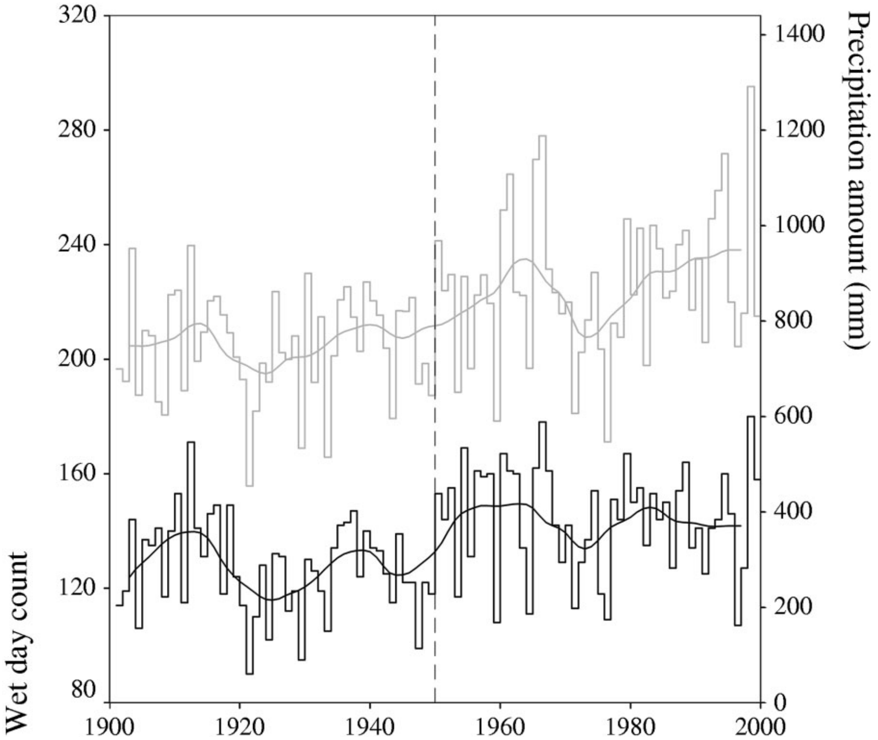

Figure 4. Black: wet day (threshold 1 mm) count at station Putten (The Netherlands). Grey: annual precipitation amount at the station. The smoothed curves are calculated using the Loess smoother (Cleveland, 1979) with a time span of 15 years. The 1950 change in height from 1.50 m to 0.40 m is indicated with the vertical dashed line

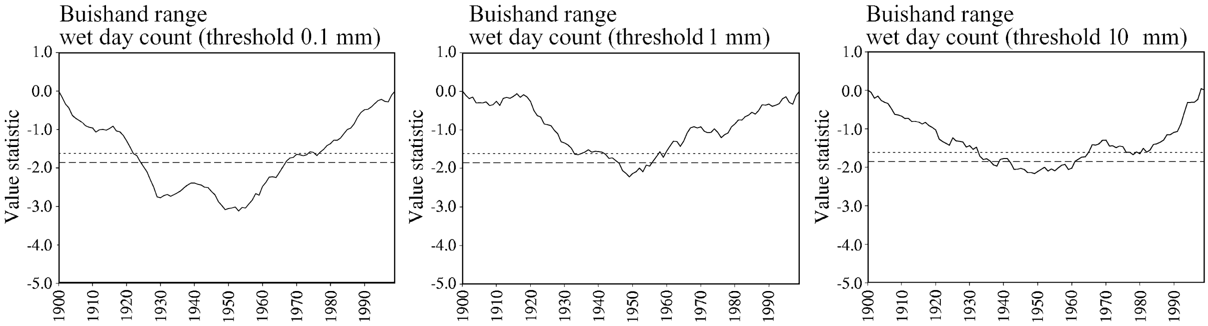

Figure 5. Test results of the Buishand range test applied to the annual wet day count series at station Putten, The Netherlands, with thresholds of 0.1 (left), 1 (middle) and 10 mm (right). Dashed lines give 1% critical values; dotted lines give 5% critical values

the break, i.e. 1950. As Figure 5 (left panel) shows, the 0.1 mm threshold counts point towards a possible second inhomogeneity in 1930, which is hardly visible for the other thresholds. This second inhomogeneity is likely to be caused by changes in how precisely very low amounts of precipitation are measured. As the main purpose of the present study is to assess the quality of the ECA precipitation data for investigating trends in wet extremes, day counts over the threshold of 0.1 mm are less suitable as a testing variable, being oversensitive to breaks in the low amounts. On the other hand, a 10 mm threshold seems too high, as it decreases the sensitivity of the tests (see Figure 5, right panel) because of the increased rarity of such days. In conclusion, the 1 mm threshold is a good compromise for detecting breaks in the ECA precipitation series, as illustrated in the Putten example.

6. TEST RESULTS FOR THE ECA SET

6.1. ECA temperature series

Figure 6 shows, separately for the mDTR and the vDTR testing variables, for each ECA station the number of tests that rejected the null hypothesis at the 1% level for the period 1901–99 and its sub-period 1946–99. The figure shows that the breaks in mDTR (left) are not always accompanied by breaks in vDTR (right) or *vice versa*. In the 1946–99 (1901–99) period 20% (65%) of the series are classified 'suspect' (at least three tests significant) on the basis of both the mDTR and vDTR, whereas 23% (17%) of the series are classified 'suspect' only on the basis of the mDTR and 11% (10%) of the series are classified 'suspect' only on the basis of the vDTR. The series with detected breaks are from stations scattered all over Europe, but for mDTR a concentration in central Europe is apparent. Table I summarizes the statistical results according to the classification system outlined in Section 3, in which mDTR and vDTR results are merged. The full 1901–99 period shows a very high percentage of 'suspect' station series, but for the 1946–99 sub-period the situation improves. Both in the full period and the sub-period, only a few series are assigned to the 'doubtful' class. Apparently, the four tests agree on the class in most cases. Agreement also exists on the year of the break for the three location-specific tests. For 73% (79%) of the 'suspect' series the year of the detected break is within three (six) years (not shown).

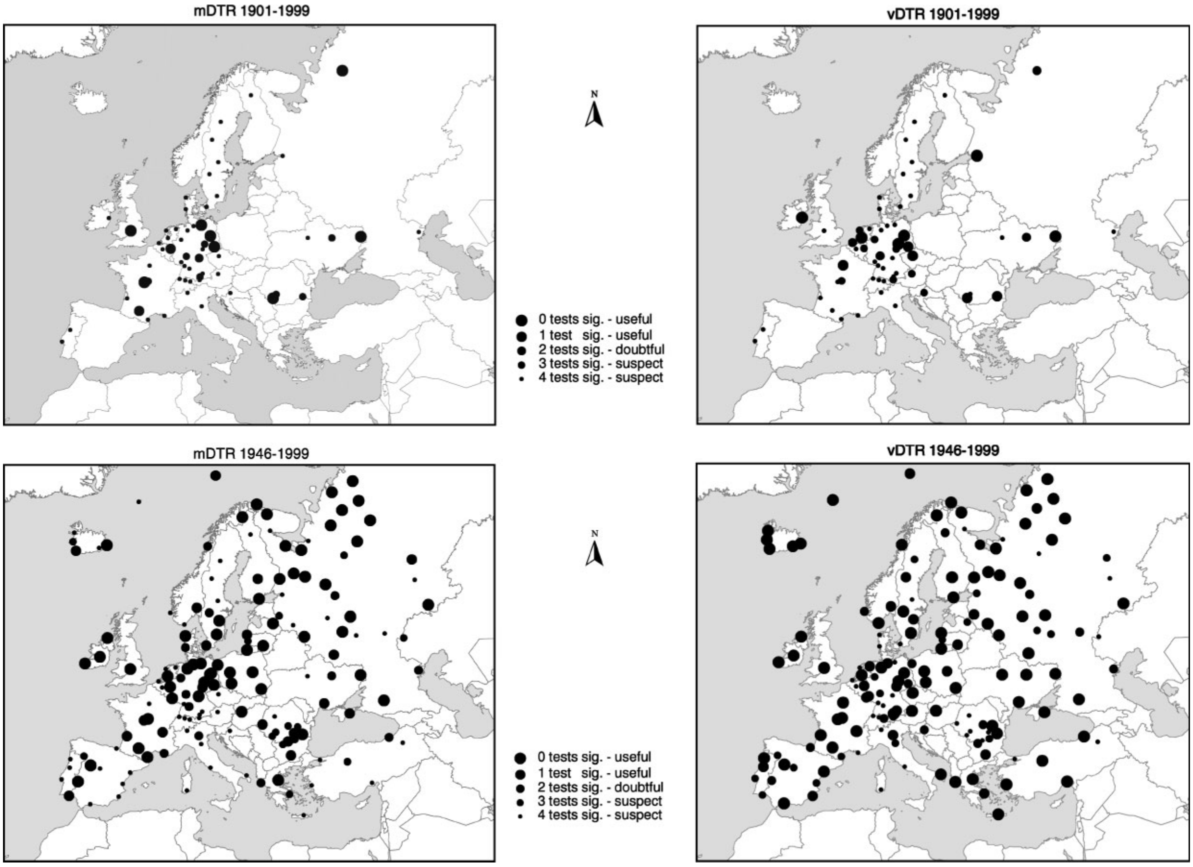

Figure 6. Test results for the ECA temperature series. Station dots are scaled with the number of tests that detected breaks in temperature series at the 1% significance level. The four tests applied to the mDTR series (left) and the vDTR (right) are: the SNHT, the Buishand range, the Pettitt, and the Von Neumann ratio test. The classification is indicated in the legend. Top: 1901–99 (stations with at least 79 observation years). Bottom: 1946–99 (stations with at least 43 observation years)

Table I. Number of temperature series (percentage) in the ECA set in the categories ‘useful’, ‘doubtful’ and ‘suspect’. This classification is based on four tests, applied for the period 1901–99 (stations with at least 79 observation years) and the sub-period 1946–99 (stations with at least 43 observation years). If the test results are different for the two temperature testing variables (mDTR and vDTR), then the highest of the two categories is assigned to the daily series. This applies to 28% and 41% of the cases respectively for the 1901–99 and 1946–99 periods. For more details see Section 3

Period	Class 1 ‘useful’	Class 2 ‘doubtful’	Class 3 ‘suspect’	Total number of station series
1901–99	4 (6%)	1 (2%)	55 (92%)	60 (100%)
1946–99	61 (39%)	12 (7%)	85 (54%)	158 (100%)

6.2. ECA precipitation series

Figure 7 shows the test results for precipitation. In both test periods, only for Eastern Europe does a considerable part of the station series suffer from significant breaks. These breaks are mostly detected by all four tests and are mainly located around 1950 and 1965. Making a distinction between the different seasons (not shown), it appears that the winter season is more strongly affected by breaks. Regarding Europe in its entirety, Table II shows that the majority of precipitation series are labelled ‘useful’. Like temperature, the highest number of precipitation series labelled ‘suspect’ appear in the full period. Consistent with the higher year-to-year variability of the precipitation testing variable, the percentage of stations labelled ‘suspect’ is much smaller for precipitation than for temperature. For 59% (69%) of the precipitation series, the year of the break detected by the location-specific tests agrees within three (six) years.

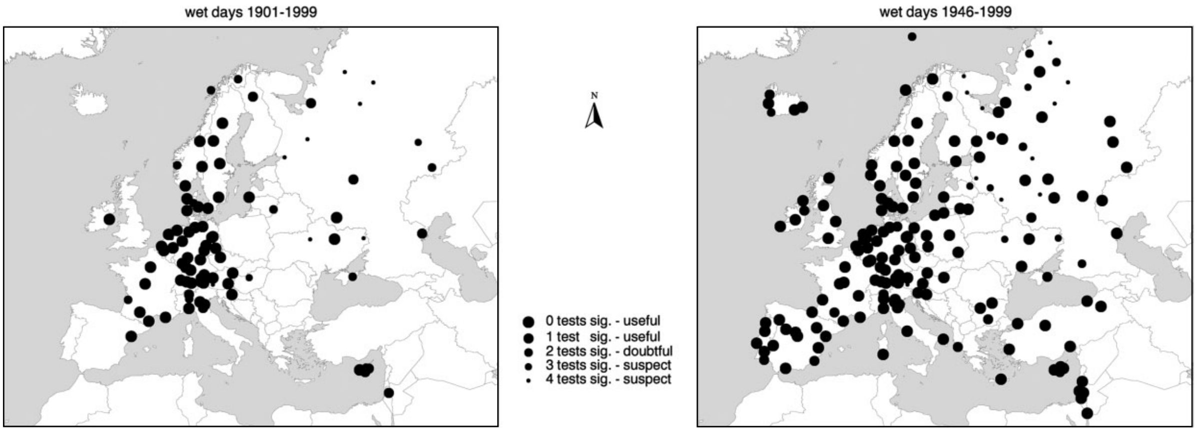

Figure 7. Test results for the ECA precipitation series. Station dots are scaled with the number of tests that detected breaks in precipitation series at the 1% significance level. The four tests applied to the annual number of wet days (threshold 1 mm) are the SNHT, the Buishand range, the Pettitt, and the Von Neumann ratio test. The classification is indicated in the legend. Left: 1901–99 (stations with at least 79 observation years). Right: 1946–99 (stations with at least 43 observation years)

Table II. As Table I, but for precipitation. The testing variable is the wet day count (threshold 1 mm)

Period	Class 1 ‘useful’	Class 2 ‘doubtful’	Class 3 ‘suspect’	Total number of station series
1901–99	66 (75%)	10 (11%)	12 (14%)	88 (100%)
1946–99	156 (87%)	7 (4%)	17 (9%)	180 (100%)

6.3. Comparison with metadata

For temperature, the cause of about 65% of the breaks detected in station series labelled class 2 'doubtful' or class 3 'suspect' in the period 1946–99 can be traced back from historic metadata information. About 15% of the breaks appear to be related to known climate variations, whereas the remaining 20% of the breaks could not be explained from metadata or climate variations.

The most frequent change in observational routine documented in the metadata is station relocation, but changes in observing practice and measuring technique (including changing instrument exposure) do occur as well. Although we did not include these series in our first analysis (Klein Tank and Können, 2003), they are good candidates for correction and hence promotion to a more reliable category.

The breaks that are triggered by true local climate variations, rather than by inhomogeneities, are identified on the basis of the expert judgment of the participants in the ECA project (Klein Tank *et al.*, 2002). These series can be considered for studying trends and variability in weather extremes, despite their poor classification. A prominent example is the temperature series of the coastal station Oksøy-Fyr in Norway, which was assigned to the class 'suspect' (all four tests agreed at the 1% level for the mDTR), even though this station was not subject to any change in observational routine. The temperature series of this station is very sensitive to the statistical tests because of its very low temperature variability. Another example refers to the clustering of stations labelled 'suspect' in central Europe, an effect that is likely caused by a real regional climate variation in that area (e.g. Brázdil *et al.* (1996)).

For precipitation, about 90% of the breaks detected in station series labelled class 2 or 3 in the period 1946–99 can be traced back from historic information. No explanation could be found for the remaining 10%. The breaks detected around 1950 and 1965 in the station series of the eastern part of Europe can be understood from the analysis of the metadata of these stations as performed by Groisman *et al.* (1991). The most important observational changes in the former Soviet Union station series they mention is the introduction of a new wind shield around 1950 and a change in 'wetting' correction in 1966–67. Consistent with their findings, we find that the series are more strongly affected in the winter season than in the summer season, as solid precipitation gives rise to larger inhomogeneities than liquid precipitation.

7. DISCUSSION

We were able to detect inhomogeneities in the ECA daily temperature and precipitation series effectively by testing series of two DTR variables and of the number of wet days (threshold 1 mm). The test results were then condensed in a three-class system with categories termed 'useful', 'doubtful' and 'suspect'. So far, for the series labelled 'doubtful' or 'suspect', about 65% of the statistically detected inhomogeneities in the temperature series and about 90% in the precipitation series could be explained by metadata. Although the hybrid test method described in this paper may not account for all kinds of inhomogeneities, the most severe step-wise discontinuities in the temperature series are recognized. Nevertheless, care has to be taken when doing trend analysis on station series that are prone to urbanization effects, i.e. gradual changes due to non-climatic causes. For precipitation, with its higher variability, fewer breaks can be detected. In trend studies, this shortcoming is compensated by the fact that possibilities for trend detection in these series are also reduced.

Because of the sparse density of the ECA station network, absolute tests were applied rather than relative tests, i.e. testing station series relative to neighbouring baseline series. For the precipitation series, this seems a good choice, given the fact that the detected inhomogeneities are mostly caused by simultaneous changes in the observational network for which relative tests are insensitive. For the temperature series, absolute tests are less preferable as they tend not to discriminate breaks from climate variations. We found that we could overcome this problem by choosing DTR-related parameters as testing variables rather than mean temperature. However, when more station series become available in the ECA dataset, relative tests should then be able to play an increasingly important role in the homogeneity evaluation as an intermediate step towards the application of methods developed for testing daily resolution temperature and precipitation series.

The most frequently encountered causes for the breaks detected in the temperature series were station or instrument relocations, changing observing practices and measuring techniques. Our station example shows that such discontinuities are difficult to detect in the annual means of temperature. It was found that the DTR variables are capable of detecting these inhomogeneities in temperature series. For precipitation, the main causes of the breaks detected are simultaneous changes in the measuring technique. Our method, with testing variable the wet day count (threshold 1 mm), effectively detected these breaks. The choice of the threshold value could be the subject of debate. We adopted the 1 mm value since (a) within the main ECA goal of studying wet extremes, detection of inhomogeneities due to very low daily amounts is not pursued, and (b) a threshold much higher than 1 mm results in a testing variable series having too few annual counts.

Making adjustments to inhomogeneous series could be a method of improving the series in the ECA dataset. However, we did not try to adjust the daily series for the inhomogeneities detected. Adjusting on a daily basis is not straightforward, and needs a very careful approach. The development of adjustment techniques, as well as robust homogeneity tests for daily series, is still in its infancy. Promising techniques may include information from parallel measurements and other weather variables, such as cloud cover and wind data (Brandsma, 2001). All these efforts should be carefully evaluated when considered as part of the future work following our explorative study.

Historic metadata support is essential for evaluating the breaks detected and for any future attempt to correct series for these breaks. Unfortunately, these metadata are not always readily available. Effort should be put into constructing an extensive metadata file to enhance further the usefulness of the ECA dataset and similar datasets for climate research. This file should at least comprise the most important information on observational changes. In fact, responsible national institutes are increasingly coming to realize that conserving and disclosing the practices of observation are very important in climate research.

Although the need for high-quality meteorological observations for climate analysis is increasingly being recognized, the quality of the recent observations in the ECA series is not superior to the early observations. Tuomenvirta (2001) even reported in Nordic series as many breaks in recent decades as in earlier ones. Given the introduction of automated weather stations throughout Europe, it is likely that the Nordic situation is typical of other European countries. This implies that continuing efforts are necessary to maintain and improve the operation of observational networks and that care has to be taken with automation of observations with respect to time series homogeneity (WMO, 2002).

ACKNOWLEDGEMENTS

The authors thank the participants of the ECA project for their cooperation (see http://www.knmi.nl/samenw/eca). We also thank our colleagues, who provided useful comments.

APPENDIX A

Y_i (i is the year from 1 to n) is the annual series to be tested, $\overline{Y}$ is the mean and s the standard deviation.

Standard normal homogeneity test

Alexandersson (1986) describes a statistic $T(k)$ to compare the mean of the first k years of the record with that of the last $n-k$ years:

$$T(k) = k\overline{z}_1^2 + (n-k)\overline{z}_2^2 \quad k = 1, \ldots, n$$

where

$$\overline{z}_1 = \frac{1}{k}\sum_{i=1}^{k}(Y_i - \overline{Y})/s \quad \text{and} \quad \overline{z}_2 = \frac{1}{n-k}\sum_{i=k+1}^{n}(Y_i - \overline{Y})/s$$

If a break is located at the year K, then $T(k)$ reaches a maximum near the year $k = K$. The $T(k)$ is depicted in the graphs representing the results of this test. The test statistic T_0 is defined as:

$$T_0 = \max_{1 \le k < n} T(k)$$

The test has further been studied by Jarušková (1994). The relationship between her test statistic $T(n)$ and T_0 is

$$T_0 = \frac{n(T(n))^2}{n - 2 + (T(n))^2}$$

The null hypothesis will be rejected if T_0 is above a certain level, which is dependent on the sample size. Critical values are given in Table III.

Buishand range test

In this test, the adjusted partial sums are defined as

$$S_0^* = 0 \text{ and } S_k^* = \sum_{i=1}^{k} (Y_i - \overline{Y}) \quad k = 1, \ldots, n$$

When a series is homogeneous the values of S_k^* will fluctuate around zero, because no systematic deviations of the Y_i values with respect to their mean will appear. If a break is present in year K, then S_k^* reaches a maximum (negative shift) or minimum (positive shift) near the year $k = K$. The $(\mathrm{S}_k^*/s)/\sqrt{n}$ is depicted in the graphs representing the results of this test. The significance of the shift can be tested with the 'rescaled adjusted range' R, which is the difference between the maximum and the minimum of the S_k^* values scaled by the sample standard deviation:

$$R = (\max_{0 \le k \le n} S_k^* - \min_{0 \le k \le n} S_k^*)/s$$

Buishand (1982) gives critical values for $R/\sqrt{n}$ (see Table IV).

Pettitt test

This test is a non-parametric rank test. The ranks $r_1, \ldots, r_n$ of the $Y_1, \ldots, Y_n$ are used to calculate the statistics:

$$X_k = 2\sum_{i=1}^{k} r_i - k(n+1) \quad k = 1, \ldots, n$$

The X_k is depicted in the graphs representing the results of this test.

If a break occurs in year E, then the statistic is maximal or minimal near the year $k = E$:

$$X_E = \max_{1 \le k \le n} |X_k|$$

The significance level is given by Pettitt (1979). Critical values for X_E are given in Table V.

Von Neumann ratio

The von Neumann ratio N is defined as the ratio of the mean square successive (year to year) difference to the variance (Von Neumann, 1941):

$$N = \sum_{i=1}^{n-1} (Y_i - Y_{i+1})^2 / \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Table III. 1% critical values for the statistic T_0 of the single shift SNHT as a function of n (calculated from the simulations carried out by Jarušková (1994)) and the 5% critical value (Alexandersson and Moberg, 1997)

n	20	30	40	50	70	100
1%	9.56	10.45	11.01	11.38	11.89	12.32
5%	6.95	7.65	8.10	8.45	8.80	9.15

Table IV. 1% and 5% critical values for$R/\sqrt{n}$ of the Buishand range test as a function of n (Buishand, 1982); the value of $n = 70$ is simulated

n	20	30	40	50	70	100
1%	1.60	1.70	1.74	1.78	1.81	1.86
5%	1.43	1.50	1.53	1.55	1.59	1.62

Table V. 1% and 5% critical values for X_E of the Pettitt test as a function of n; values are based on simulation

n	20	30	40	50	70	100
1%	71	133	208	293	488	841
5%	57	107	167	235	393	677

Table VI. 1% and 5% critical values for N of the Von Neumann ratio test as a function of n. For $n \leq 50$ these values are taken from Owen (1962); for $n = 70$ and $n = 100$ the critical values are based on the asymptotic normal distribution of N (Buishand, 1981)

n	20	30	40	50	70	100
1%	1.04	1.20	1.29	1.36	1.45	1.54
5%	1.30	1.42	1.49	1.54	1.61	1.67

When the sample is homogeneous the expected value is $N = 2$. If the sample contains a break, then the value of N tends to be lower than this expected value (Buishand, 1981). If the sample has rapid variations in the mean, then values of N may rise above two (Bingham and Nelson, 1981). This test gives no information about the location of the shift. Table VI gives critical values for N.

REFERENCES

Alexandersson H. 1986. A homogeneity test applied to precipitation data. *Journal of Climatology* **6**: 661–675.

Alexandersson H, Moberg A. 1997. Homogenization of Swedish temperature data. Part 1: homogeneity test for linear trends. *International Journal of Climatology* **17**: 25–34.

Bingham C, Nelson LS. 1981. An approximation for the distribution of the Von Neumann ratio. *Technometrics* **23**: 285–288.

Brandsma T. 2001. Weather type dependent homogenization of the daily Zwanenburg/De Bilt temperature series. In *Proceedings of the Third Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, Hungary, 25–29 September 2000*, Szalai S (ed.). WMO-TD: Geneva, Switzerland.

Brázdil R, Budíková M, Auer I, Böhm R, Cegnar T, Faško P, Lapin M, Gajić-Čapka M, Zaninović K, Koleva E, Niedźwiedź T, Ustrnul Z, Szalai S, Weber RO. 1996. Trends of maximum and minimum daily temperatures in central and southeastern Europe. *International Journal of Climatology* **16**: 765–782.

Buishand TA. 1981. The analysis of homogeneity of long-term rainfall records in the Netherlands. KNMI Scientific Report WR 81-7, De Bilt, The Netherlands.

Buishand TA. 1982. Some methods for testing the homogeneity of rainfall records. *Journal of Hydrology* **58**: 11–27.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**: 829–836.

Groisman PYA, Koknaeva VV, Belokrylova TA, Karl TR. 1991. Overcoming biases of precipitation measurement: a history of the USSR experience. *Bulletin of the American Meteorological Society* **72**: 1725–1733.

Hawkins M. 1977. Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association* **72**: 180–186.

Heino R, Brázdil R, Førland E, Tuomenvirta H, Alexandersson H, Beniston M, Pfister C, Rebetez M, Rosenhagen G, Rösner S, Wibig J. 1999. Progress in the study of climatic extremes in northern and central Europe. *Climatic Change* **42**: 151–181.

Jarušková D. 1994. Change-point detection in meteorological measurement. *Monthly Weather Review* **124**: 1535–1543.

Karl TR, Knight RW, Plummer N. 1995. Trends in high-frequency climate variability in the twentieth century. *Nature* **377**: 217–220.

Klein Tank AMG, Können GP. 2003. Trends in extreme indices of temperature and precipitation in Europe, 1946–1999. *Journal of Climate*: accepted.

Klein Tank AMG, Wijngaard JB, Können GP, Böhm R, Demarée G, Gocheva A, Mileta M, Paschiardis S, Hejkrlik L, Kern-Hansen C, Heino R, Bessemoulin P, Müller-Westermeier G, Tzanakou M, Szalai S, Pálsdóttir T, Fitzgerald D, Rubin S, Capaldo M, Maugeri M, Leitass A, Bukantis A, Aberfeld R, van Engelen AFV, Førland E, Mietus M, Coelho F, Mares C, Razuvaev V, Nieplova E, Cegnar T, López AJ, Dahlström B, Moberg A, Kirchhofer W, Ceylan A, Pachaliuk O, Alexander LV, Petrovic P. 2002. New daily dataset of surface air temperature and precipitation observations for European Climate Assessment. *International Journal of Climatology* **22**: 1441–1453.

Moberg A, Jones PD, Barriendos M, Bergström H, Camuffo D, Cocheo C, Davies TD, Demarée G, Martin-Vide J, Maugeri M, Rodriguez R, Verhoeve T. 2000. Day-to-day temperature variability trends in 160- to 275-year-long European instrumental records. *Journal of Geophysical Research* **105**: 22 849–22 868.

Owen DB. 1962. *Handbook of Statistical Tables*. Addison Wesley: Reading, UK.

Peterson TC, Easterling DR, Karl TR, Groisman P, Nicholls N, Plummer N, Torok S, Auer I, Böhm R, Gullett D, Vincent L, Heino R, Tuomenvirta H, Mestre O, Szentimrey T, Salinger J, Førland EJ, Hanssen-Bauer I, Alexandersson H, Jones P, Parker D. 1998. Homogeneity adjustments of *in situ* atmospheric climate data: a review. *International Journal of Climatology* **18**: 1493–1517.

Pettitt AN. 1979. A non-parametric approach to the change-point detection. *Applied Statistics* **28**: 126–135.

Schönwiese CD, Rapp J. 1997. *Climate Trend Atlas of Europe Based on Observations 1891–1990*. Kluwer: Dordrecht, The Netherlands.

Szalai S (ed.). 1997. *Proceedings of the First Seminar for Homogenization of Surface Climatological Data*. HMS Publication: Budapest, Hungary.

Szalai S, Szentimrey T, Szinell C (eds). 1999. *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data*. WMO-TD 962. WMO: Geneva, Switzerland.

Sparks WR. 1972. The effect of thermometer screen design on the observed temperature, WMO: 315. WMO, Geneva, Switzerland.

Tuomenvirta H. 2001. Homogeneity adjustments of temperature and precipitation series — Finnish and Nordic data. *International Journal of Climatology* **21**: 495–506.

Von Neumann J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* **13**: 367–395.

WMO. 2002. *Guide to Climatological Practices*, fourth edition. WMO 100. WMO: Geneva, Switzerland.