Homogeneity, Data Adjustments and Climatic Normals

Nathaniel B. Guttman
National Climatic Data Center
151 Patton Avenue
Asheville, NC 28801-5001
704-271-4479
nguttman@ncdc.noaa.gov

March 1998

1. Introduction

Climatic normals are defined by the World Meteorological Organization (1983) as “period averages [of a climatic element such as temperature or precipitation] computed for a uniform and relatively long period comprising at least three consecutive ten-year periods”. Normals are computed for 30-year periods beginning with each decade (e.g., 1961-1990). If data are not available for thirty years, then it is permissible to compute averages for shorter periods of at least 10 years of record if the data reflect both current observation practices and the climate at the station. In practice, observing sites and instruments are moved, new instrumentation is used, sensor calibration and/or maintenance procedures change, observing methods and codes change, and environmental effects such as vegetation are not constant. Data collected over a long period therefore may not reflect uniform climatic conditions over the period for which normals are to computed. Prior to computing period averages, or normals, the homogeneity of observed data with respect to non-climatic influences must therefore be assessed. If the data are found to be inhomogeneous, then it must be determined if the data can be adjusted so that the adjusted data set will reflect a uniform observing environment for a 30-year period. Heim and Guttman (1997) discuss these issues as applied to the recently instituted automated surface observing system in the U.S.

The broad subjects of data homogeneity, continuity and adjustments have received a lot of attention in the literature and at conferences for more than a decade. A good summary of homogeneity analyses of long-term *in situ* climatic time series and data adjustment methods that are practiced in various countries around the world is given by Peterson *et al.* (1998). Another good summary is by the Hungarian Meteorological Service (1997). Both of these contain numerous references.

Most of the analysis techniques and adjustment methods have been applied to monthly or annual time series. The U.S. climatic normals, however, are (at the time of this writing) likely to be computed from daily data for the period 1971-2000. This paper examines some of the techniques and methods from a statistical viewpoint for the purpose of assessing the magnitude of non-climatic inhomogeneities in the 30-year, daily time series that can be detected and adjusted by the various procedures. It also recommends an alternative, deterministic approach to calculating daily normals.

2. Basic concepts

The most commonly used information about non-climatic influences comes from records of station moves, changes in instrumentation, problems with instrumentation, sensor calibration and maintenance logs, changes in surrounding environmental characteristics and structures, observing practices, and other similar features. The composite of this information is generally referred to as station metadata. The advantage of metadata is that it can provide the analyst with detailed knowledge of when possible non-climatic heterogeneities were introduced into a time series of observations as well as of the exact cause of the heterogeneities. Unfortunately, metadata is often incomplete or erroneous, so that information from other sources, such as statistical analyses, are needed to assess the homogeneity of a time series.

A single time series can be statistically examined for heterogeneities, but it is difficult to determine if changes or lack of changes result from non-climatic or climatic influences. The concept of relative climatic homogeneity was introduced by Conrad and Pollak (1950) to isolate the non-climatic influences. They assume that within a geographical region, climatic patterns will be identical and that observations from all sites within the region will reflect this identical pattern. The data collected at all of the sites within the region should be highly correlated, have similar variability, and differ only by scaling factors and random sampling variability. Data from a site that indicate more than a random departure from the regional climate are considered to be inhomogeneous, and the cause of the inhomogeneity is considered to be non-climatic. If the regional climate is represented by a reference series of data that reflects the characteristics of the regional climate, and if a heterogeneity in the data for a location within the region is determined to be non-climatic, then the heterogeneous data can be adjusted to conform to the regional climate. In hydrologic and water resource circles, these principles are the basis of regional frequency analysis (Hosking and Wallis, 1997).

Three types of adjustment factors are commonly used. First, for elements such as temperature that are assumed to change in parallel to the reference series, the adjustment factor is the average difference between the reference series and the inhomogeneous data. Second, for elements such as precipitation that are assumed to change in a relative manner to the reference series, the adjustment factor is the average ratio of the inhomogeneous data to the reference series. Third, adjustments are based on regression relationships.

There are three basic practical questions that need to be answered before applying one of the above approaches:

1. Is there a non-climatic inhomogeneity in a data series?
2. What is the magnitude of an inhomogeneity that can be detected?
3. How should a regional climate be described?

The following sections will assess some of the statistical techniques that are commonly used to answer these questions.

3. Detecting non-climatic inhomogeneities

If accurate and complete metadata is available, the temporal record of non-climatic or external factors that affect the data is known. Determining if there is an inhomogeneity in the data then becomes a conceptually trivial matter of examining the metadata. However, if metadata are or are suspected of being inaccurate and/or incomplete, the detection of inhomogeneities is not trivial and involves a comparison of the data series to that of the reference series.

It is important to note the different kinds of effects that could be exhibited in a climatological record from non-climatic inhomogeneities. A step function could result from a recalibration of an instrument. A linear trend could result from a gradual but constant degradation of a sensor, and a non-linear trend could result from vegetative growth around the instruments. A new instrument may respond differently from the old instrument only in some atmospheric conditions but not in others. These examples show the broad range of effects that non-climatic inhomogeneities may have on a data record; the problem of detecting all of these effects is difficult.

Easterling and Peterson (1995) reviewed some of the easily automated detection techniques that involve a comparison between a reference series and the data series that is being examined. Most of the techniques that they reviewed are designed for detecting abrupt discontinuities that are defined as a change in the mean of a time series. These include double mass analysis (Kohler, 1949), concluding that it is impossible to determine which one of the two series being compared contains a heterogeneity, and the parallel cumulative sum (CUSUM) method described by Rhoades and Salinger (1993) that uses comparisons between the data series and several reference series, concluding that the method is subjective, tedious, and has the disadvantage of not immediately providing the means for calculating an adjustment factor. Zhang (1998), in a paper concerning statistical control charts, relates and references studies that show that the CUSUM method is not appropriate for autocorrelated data such as climate data. He also describes the application of CUSUM techniques to residuals from time series models, exponentially weighted moving average (EWMA) charts, and his modifications to the EWMA, and concludes that no process control chart performs well (identifies an inhomogeneity) when mean shifts are small in a nonstationary or near nonstationary process, i.e., in cases similar to the heterogeneities found in climate data. More importantly, he shows that sample sizes from stable processes should be at least 100 for reliable results; this sample size is usually much greater than climatic time series allow.

The Standard Normal Homogeneity Test, or SNHT, is described by Tuomenvirta and Alexandersson (1997). It assumes that a standardized difference or ratio series between the reference data and the data that are being assessed is fairly constant. It is a parametric procedure since it assumes that the difference/ratio series is a standard normal distribution. Early versions of the test looked at sequential segmentation of the series into two parts, and the test statistic compares the mean for each of the two segments before and after the point of segmentation. This version is designed to identify a single shift of the mean. Later versions allow for a double shift of the mean, and a single change in both mean and variance. Although these tests are formulated for identifying abrupt changes in a series, the SNHT can also be formulated to examine linear temporal trends.

Easterling and Peterson (1995) use regression techniques to search for inhomogeneities in temperature data. They fit a simple linear regression to the whole difference series with time as the predictor and differences as the predictand, and then calculate the residual sum of squares. They then calculate, for sequentially increasing segmentation points, the residual sum of squares from a two-phase linear regression. Since they are looking for abrupt changes, the two regression lines are not constrained to meet at the segmentation point. The time for which the residual sum of squares from the two-phase regression is a minimum is the time at which a potential inhomogeneity is thought to occur. The significance of this potential inhomogeneity is determined from a likelihood ratio test given by Solow (1987). They also test the difference in the means between the two segments with a Student’s t-test to insure that a series with a trend will found to be inhomogeneous even if the likelihood ratio is not found to be significant. Once a potential discontinuity is found, the procedure is iterated on the subsets of data both before and after the point of initial segmentation. The process is therefore designed to identify multiple potential discontinuities. The final test of discontinuities involves application of a nonparametric multiresponse permutation procedure that is described by Mielke and Berry (1994, 1997, 1998).

Vincent (1998) applies four regression models to identify periods of homogeneity and inhomogeneity, abrupt changes, trends, and the most probable time of occurrence of an inhomogeneity. First, a simple linear regression is calculated with a reference series (or multiple reference series) as the predictor(s) and the series being examined as the predictand. The data series is considered to be homogeneous if the residuals from the regression line are independent, normal random variables with zero mean and constant variance. The residual pattern is examined visually as well as subjected to the Durbin-Watson test for significance of lag one autocorrelation and the Chatfield assessment of autocorrelation at lags greater than one. If autocorrelation exists, then a second regression is calculated in which a linear trend is included. If autocorrelation in the residuals of this second model exist, then the model is discarded and a third model is examined. This model is the first model with a step function added. The third kind of regression is calculated for sequential increases in the time at which the step can occur. The minimum residual sums of squares from these regressions identifies the time of the discontinuity. If autocorrelations exist in the residuals from regression with the identified step, a fourth model is considered. The last regression is that of trends before and after a step.

These techniques for detecting inhomogeneities in a climate record are the common approaches that are used, with some variation, by researchers in many countries. They are intended to detect three kinds of inhomogeneities: a change in the average value between two segments of data, a change in the average value and variance between two segments, and a linear trend. The characteristics of a data series that is being assessed are smooth properties. Non-climatic influences are assumed to affect all the data for a specified time period in a uniform manner; they do not consider non-climatic influences that affect the data only during certain weather patterns. In addition, they do not consider non-linear influences.

The techniques described above are certainly not exhaustive and are only intended to illustrate the kinds of approaches that are commonly used. The literature is replete with techniques, but most of them are similar to or variations of the above methodologies.

4. Magnitude of detectable inhomogeneities

A Monte Carlo study by Easterling and Peterson (1995) compared some of the inhomogeneity detection techniques. A thousand series of 100-year annual mean temperature data with weak autocorrelation were simulated. A single abrupt change of between a half and two standard deviations of the data was introduced at time 65. These data sets were subtracted from corresponding reference series without step functions and the resulting difference series were tested. Results show that the SNHT, two-phase regression, Student's t-test and CUSUM procedures all correctly identified the single step function to within one time interval in over 92 percent of the cases when the magnitude of the step was greater than 1.5 standard deviations. For a small step of 0.5 standard deviations, the SNHT procedure was best but not good; only half the abrupt changes were detected. For a step of 1.0 standard deviations, the SNHT detected about 85 percent of the inhomogeneities, and the two-phase regression detected about 75 percent.

Similar simulations were made for multiple step functions of varying magnitude and time of change. Easterling and Peterson concluded that the SNHT is the best approach for detecting a single abrupt change of small magnitude within 2 time intervals, but that the two-phase regression approach is much more robust at detecting two changes that occur relatively closely in time. The simulation results show that for multiple step changes of at least one standard deviation, the Easterling-Peterson approach correctly identifies the time of the step to within 2 time intervals in more than 85 percent of the cases, but also that the approach either incorrectly identifies the time or identifies non-existent steps in about 20 to 25 percent of the cases.

Bosshard and Baudenbacher (1997) performed similar simulations of 60-month series of monthly temperature. Their evaluation of a single abrupt change showed that all tests correctly identified a change of 0.8°C within 11 time intervals in more than 90 percent of the cases, and that the SNHT test for an abrupt change was able to detect a change of 0.4°C in more than 85 percent of the cases. The SNHT test for a shift in variance was shown to be effective in more than 80 percent of the cases when a change in the standard deviation was greater than about 3.5. Trends indicated by the Easterling-Peterson and the SNHT trend tests were also compared. The SNHT procedure far outperformed the Easterling-Peterson method for trend detection within 2 time intervals and rates of between .008 for ratios and .040°C/year for differences.

Vincent (1998) also tested her procedure with simulated data corresponding to 100-year annual temperature series. She found that homogeneous series were incorrectly identified as being inhomogeneous in 13.6 percent of the cases. Her procedure is able to correctly identify the position of a step function in about 78 percent of the cases when the magnitude of the step is 1.0 standard deviations and in more than about 88 percent of the cases when the magnitude is greater than 1.25 standard deviations. She also shows that steps with a magnitude of 0.2°C can be detected in more than 83 percent of the cases when a segment length before or after the step is at least 10 time intervals and the step change is at least 0.5 standard deviation.

The magnitude of changes that can be detected by some parametric tests for means and residuals from regression models can be easily estimated theoretically as a function of the standard deviation s of the data series being examined. The relationship for tests that are based on the assumption of normality, such as Student's t-test and the SNHT (Alexandersson, 1986), is

$$f(s) = (T_{\mathbf{n},\alpha})s^{\mathbf{2}}$$

where f is the magnitude of detectable change, and T is based on the critical value of the test statistic for a sample size n and significance level α. For Student's t-test, the multiplier of the common standard deviation of the data from which the means of two groups are computed ranges from about 1.5 for total sample sizes n of about 10 to 1.0 for n of about 20 to .5 for n of about 60 when α=.05. For the SNHT test of a single step in the mean, the multiplier ranges from about 0.8 for n of about 25 to 0.6 for n of about 50 when α=.05.

5. Reference series

Comparing a data series to a reference series is a standard methodology in the detection of non-climatic inhomogeneities. The assumption to this approach is that the reference series is homogeneous with respect to climatic variations and is also representative of real climatic variations that are uniform over a geographical region that includes the site whose data are being evaluated. The critical part of the premise is that the reference series be representative of uniform climatic conditions. Under this assumption, the data collected at sites within the region can be considered to be random samples of the climatic conditions that affect the entire region equally. Since the characteristics of a climate are measured at point locations, the definition of a region becomes a collection of point-source data samples, each of which are representative of the same climatic environment. In practice, the "equal" constraint is often not met because of deterministic factors such as the effect of topography on the meteorological variables of interest. The constraint is therefore relaxed so that climatological homogeneity can be defined as assuming that frequency distributions at various sites are identical apart from a scaling factor.

In the climatology community, reference sites are commonly selected on the basis of distance and correlation under the assumption that data from nearby locations should reflect the regional climate and be highly, positively correlated. "Highly, positively correlated" is generally defined as being measured by linear correlation coefficients that exceed about .90. When a composite reference series is constructed from data collected at several sites, distance weighting functions or other optimization functions that minimize characteristics such as a coefficient of variation are used to accommodate climatic gradients.

Peterson and Easterling (1994) report that when considering pairs of possible reference series, discontinuities in one of or both the series that are being compared can drastically alter the correlation coefficient. If discontinuities in both series are of comparable magnitude and direction, then the correlation coefficient is higher than would be expected if one series were homogeneous. Conversely, they found that if the discontinuities are in opposite direction, the correlation is lower than would be expected if one series were homogeneous. They attempted to solve this problem by correlating series of the change of data per unit time rather than the original data series. With this approach, correlation between homogeneous rate of change series was about the same as that for the original series. For original data with a few discontinuities, the correlation between the rate of change series did not mask nor inflate correlations. The use of rate of change series in correlation analyses is now being adopted by many analysts (Peterson, personal communication, 1998).

After identifying highly correlated series that could be used to construct a reference series, Peterson and Easterling (1994) perform an additional test to decide if the high correlations could have resulted from chance. The exact distribution of the correlation coefficient for normally distributed data is known, but the authors prefer to use a nonparametric permutation test so that *a priori* distribution assumptions are unnecessary.

Rather than examining covariances as in the correlation coefficient approach, a second method, that has been used for many years in the hydrological community, examines frequency distributions of the data collected at sites within a region. It is assumed that in a climatologically homogeneous region, each frequency distribution is a random sample of a common regional distribution from which the observed data are generated. The region can be described by frequency distributions which are, after appropriate scaling, the same for all sites within the region. The measures used to describe the frequency distributions are those of location, scale and shape (the parameters of the distributions).

Hosking and Wallis (1997) present a complete description of the regional frequency analysis methodology that is based on L-moments rather than conventional moments for characterizing frequency distributions. The authors found that for small and moderate samples, the use of L-moments yields efficient and computationally convenient estimates of parameters and quantiles. A key part of the methodology is to use site characteristics such as geographical location, elevation and other physical and deterministic properties associated with each site. Potentially homogeneous regions are constructed by combining sites with similar vectors of site characteristics through standard multivariate cluster analysis or other similar techniques. The homogeneity of the potential region is then tested by using the L-moments at each site. This concept has the advantage of testing for homogeneity on the basis of characteristics that are not used in the clustering of sites.

The hypothesis of homogeneity is that the frequency distributions at each site are the same except for a site-specific scale factor. The average regional coefficient of L-variation (L-CV) is compared to the L-CV that would be expected in a homogeneous region in order to assess homogeneity; the expectation is determined by simulation. The test assumes that the frequency distributions at all sites represent the same regional climate; discordancy among the sites is measured in terms of L-moments.

This second method of constructing reference series has intuitive appeal in that regionalization is performed on the basis of physical rather than strictly statistical properties and that the functional form of frequency distributions does not have be fixed in advance. There are, however, two potential problems in its application to the construction of reference series. First, the discordancy measure is not likely to be useful for less than 7 sites in a region. Second, according to Hosking and Wallis (1997), sample sizes of at least 20 are needed to calculate reasonably unbiased estimates of the population L-moments, and Guttman (1994) recommends even larger sample sizes.

6. Confidence intervals

Decisions regarding inhomogeneities are generally based on test statistics that are expected to occur with a predetermined and constant degree of confidence. For a given significance level α, the critical value upon which a decision is based is a function of the sample size. Therefore, the confidence interval or range of values for which a null hypothesis is accepted or rejected with a specified α is also a function of the sample size.

Curves showing the relationship between critical values for α=.05 and sample size for some of the commonly used tests are shown in figure 1. They are intended as examples to show the effect of sample size on the width of confidence intervals. Flat segments of the curves indicate ranges of n for which a confidence interval is approximately constant, and sloping segments indicate the ranges of n for which the intervals vary. The curve for Student's t-test is for the hypothesis that the means of two normally distributed groups, each of size n / 2 and common standard deviation s, are equal. The bound represents the multiplier of s. The bound for the two-phase regression is based on an $F_{3,n-4}$ distribution of the test statistic (Solow, 1987). The curve for the ratio was adapted from Alexandersson (1986), and the curve for the rank correlation was plotted from a table in Snedecor and Cochran (1967). The confidence bounds for most tests are related to sample size in a manner that is similar to these examples.

The curves define sample sizes for which decisions are spatially comparable when a given test is applied to data collected at several sites. Sample sizes corresponding to a flat part of a curve result in confidence intervals that are of equal length for the whole range of n that results in the flat portion, whereas sample sizes corresponding to the sloping part of a curve result in differing confidence lengths as n changes. In the former case, decisions made for multiple sites in a geographical area are based on the same set of conditions, but in the latter case, they are not since the widths of the confidence intervals vary.

7. Relevance to calculating normals

Plans are currently being made to calculate climatic normals for the period 1971-2000 (see, for example, Heim and Guttman, 1997). The goal is to construct a data set that fulfills the World Meteorological Organization criteria as well as accommodates some of the recommendations of the American Association of State Climatologists (Kunkel and Court, 1990). These recommendations include adopting the median as a measure of central tendency for variables, such as total precipitation, which in many locations do not follow a Gaussian or symmetrical distribution; calculating a measure of variability; and describing extremes. Prior calculations of normals used monthly data as a basic time unit from which summaries of longer and shorter time intervals were derived. Current plans, however, include using daily data as a basic time unit since summaries for any desired time interval that is longer than a day can be calculated, a daily time unit is not constrained by non-climatic time entities such as arbitrarily defined months or seasons, a comparison between daily normals calculated from a spline fit of monthly data and those calculated from daily data showed significant differences (Guttman and Plantico, 1987), and electronic computing capabilities have increased dramatically.

The purposes of climatic normals are to allow comparisons of a value of a climatic variable to a reference value (normal), and to allow spatial comparisons. Comparisons among different sets of normals can lead to suggestions of long-term climate change, trends or stationarity, and changes in spatial patterns. Prediction is not one of the purposes of normals, but, as noted by Guttman (1989) and Kunkel and Court (1990), it is in fact a use of normals. These two, along with the international criteria and recommendations by the State Climatologists, serve to define the steps to be taken for calculating normals:

1. Determine a period of record for which data at a site are homogeneous with respect to climatic influences.

2. Adjust non-climatically inhomogeneous data to the latest period of homogeneous data if the causes and effects of the inhomogeneities can be completely described.

3. Describe the frequency distribution (central tendency, variability and extremes) of the homogeneous data for the period of record (1971-2000) in a manner that allows for consistent interpretation both spatially and temporally.

Central to the *statistical* determination of homogeneity, without consideration of metadata, is the construction of a homogeneous, regional reference series against which a data series with potential non-climatic heterogeneities can be compared. Most of the techniques utilize linear correlation among data sets to determine a climatologically coherent region. Highly correlated data sets, where "highly" is defined as a linear correlation that is greater than a threshold value, are assumed to be random samples of the same climatic regime. Although assessment of regional homogeneity with the L-moment frequency analysis approach is more robust than the correlation statistic assessment (in that the discordancy and homogeneity tests consider all of the frequency descriptors), application of these tests requires more data than are generally available for analysis. The discordancy test requires data from at least 7 sites to be useful, and the homogeneity test becomes better at indicating heterogeneity as the number of sites increases. Additionally, both tests require the number of observations at each site to be at least 20 in order to obtain good estimates of the population L-moments. Although preferable to the linear correlation approach, the lack of data necessary to obtain reasonable estimates of the frequency distribution parameters precludes the use of the approach.

The correlation approach can be used in the construction of a reference series, but it also has some important drawbacks. The use of a threshold value for making a decision does not take into account the sampling characteristics and confidence interval of the correlation statistic that is used for making the decision. As an example, assuming normally distributed data, the sampling characteristics of the linear correlation coefficient are known (e.g., van der Waerden, 1969), and the confidence interval varies considerably with sample size n. Figure 2 shows the α=.05 interval as a function of n for a correlation coefficient of .90, a value commonly used as a threshold. The figure shows that for small sample sizes, one cannot be confident that a calculated sample value of .90 is indicative of a high correlation in the population thereby leading to the construction of a reference series that does not in fact represent a homogeneous regional climate.

Another problem with correlation statistics is that they do not consider either nonlinear or nonuniform effects. This problem is especially critical when daily data are being analyzed. Daily observations of a climatic element are much more reflective of the fluctuations of weather patterns passing over a site than are monthly, seasonal or annual data. The daily data may reflect more of a mixture of populations, and are more likely to be affected by nonlinear and nonuniform weather events than data that are averaged over a longer time interval, so that dependence on a correlation coefficient is likely to be suspect when constructing daily reference series.

The ability to construct realistic daily reference series is therefore highly questionable. However, even if a reasonable series could be constructed, the determination of relative homogeneity is problematic. Most of the techniques summarized in the preceding sections assess relative homogeneity in terms of abrupt shifts in means, departures from a linear regression model, or a simple combination of shifts in means and linear trends, and they are used on monthly or annual data. They are capable of detecting, to within a couple of time units, shifts in the mean of about 10 percent for ratios and 3/4 to 1 standard deviation for differences. Fairly small trends can also be detected to within a couple of time units. (Unfortunately, the techniques also falsely detect changes at a high rate that is on the order of 20 percent or more.) Assuming that the magnitude of the levels of change that can be detected in monthly or annual data also apply to daily data, the inherently high variability of the daily data leads to rather large ranges of undetectable inhomogeneities.

For example, in the very simple case where climate is defined only by an average, assume, for a given calendar day, that 30 values of a well behaved element such as temperature are independently and identically normally distributed with variance σ^2, and that a step function is introduced by a non-climatic effect midway in the period of record. The magnitude of differences between the means before and after the step that could be detected with a confidence of, for example, .95, is approximately .7σ. Since the standard deviation of daily temperatures at most locations in the U.S. is of the order of 10 to 12 ^{o}F in winter and 4 to 5^{o}F in summer (Heim, personal communication, 1998), then the magnitude of detectable shifts in a mean are about 7 to 8 ^{o}F in winter and 3 to 4 ^{o}F in summer. For monthly data, with the same assumptions as for the daily data, the standard deviations are 5 to 6 ^{o}F in winter and 2 to 3 ^{o}F in summer so that the magnitude of detectable shifts are about 4 ^{o}F in winter and near 2 ^{o}F in summer. In this ideal example, the confidence intervals are too broad to provide the precision desired for and implied by climatic normals. In the real world, the test assumptions made for this example will, for most daily data sets, be violated and the magnitude of detectable effects of non-climatic inhomogeneities is likely to be even greater.

Test assumptions are critical to the application of the parametric tests upon which the homogeneity decisions are based. Student's t-test, the likelihood ratio tests in the SNHT procedure, and other parametric tests all have assumptions which must be valid (or approximately valid) if the test is to be used properly. Most tests assume that the data are distributed according to a known function such as the normal distribution and that the data are independent and identically distributed. Tests that compare means often assume that variances of subsets of the data are equal. Daily climatological data are generally non-normal and often not identically distributed so that the assumptions of most parametric tests are not met, and decisions based on the tests are questionable. The nature of even the monthly and annual data may be a major contributor to the high false detection and error rate noted earlier in the use of parametric tests; results for daily data are likely to be worse.

Nonparametric tests such as the permutation tests used described by Mielke (*op.cit*), the Mann-Whitney-Wilcoxon test, and Spearman's rank correlation test, are based solely on the ordering of the data and not on any definite distribution function. Since they are distribution-free, they are more attractive than parametric tests in the analysis of climatological data. An objection to the use of permutation tests, however, is that the tests are strictly data-dependent. All the information that is available is contained in the sample data so that different random samples of the same population may lead to different test results. In an assessment of climatological homogeneity, replication is not an option since only one observed sample of data is available for analysis, and therefore the effect of the data-dependency on test results cannot be determined.

Both the parametric and nonparametric tests that are described above and commonly used in the assessment of regional and site homogeneity look at only one or a few of the characteristics of a frequency distribution. These characteristics therefore are the defining qualities of homogeneity. They do not include nonlinear effects nor do they consider non-climatic influences that affect data in a nonuniform manner, such as only during certain weather events, seasons, etc. Since the goal of the daily normals is to describe the complete frequency distribution of daily data, any assessment of homogeneity that is based solely on the parametric and/or nonparametric tests will be incomplete.

None of the commonly used statistical techniques can adequately identify climatologically homogeneous periods of record. Since homogeneous periods cannot be established, there is very little basis upon which to either adjust data or completely describe the central tendencies, variability and extremes of frequency characteristics, i.e., climatic normals. When only these statistical techniques are used, confidence that could be placed on calculated normals would be minimal.

If metadata is available from which *a prior* determinations can be made of the introduction of heterogeneities, and if the metadata is suspected of containing errors or of not containing important information, the problem of identifying periods of homogeneity becomes the more restrictive problem of the statistical testing for homogeneity of the *a priori* identification of a potentially homogeneous period.

Nonparametric tests are more suited for these homogeneity tests than parametric tests because frequency distributions do not have to specified. One disadvantage is that the probability of a calculated test statistic depends on the number of possible permutations of the data and therefore may be a function of sample size. Most tests, however, are equally sensitive to small and large samples. Another disadvantage, as already noted, is that they test only some of the characteristics of a frequency distribution and not the collective of characteristics.

If the collective is not tested, then the complete frequency distributions for different periods of record cannot be compared. The collective of characteristics summarizes the contribution of each data value to the whole frequency distribution. Unless all of the characteristics can be compared, the impact of inhomogeneities that affect data nonuniformly and/or nonlinearly are difficult to determine with any reasonable degree of certainty. Also, it is difficult to separate with certainty the climatic effects from non-climatic effects unless a comparative reference series can be constructed, and, as shown above, construction of an adequate daily or monthly reference series is highly questionable. Therefore, comparisons among frequency distributions for differing periods of record that are determined from incomplete or suspect *a priori* information is not viable, and, according to the recommendations of the World Meteorological Organization (1989), data for one period of record should not be adjusted to the data for another period of record.

Fortunately, for the U.S., there is reasonably complete and accurate metadata for U.S. weather stations that contains information about station moves, instrument changes and other non-climatic factors that can be used to identify climatically homogeneous periods of record. Metadata is simply a collection of information about the data; interpretation of the information is necessary in making the homogeneity decisions. Assessments must be made as to what the impacts are of a specific change in siting environment, instrumentation, observing practices, etc. As examples, it should be determined how much of a horizontal or vertical change in station location is needed to put the station into a different climate, what the response of a new sensor to various weather events is compared to the response of the old sensor, and what the effect of coding changes is on the data representation of the climate. These assessments are deterministic in the sense that they are made from the viewpoint of analyses of the physical processes that control climate and its measurement. Once the assessments are made, decision trees can be constructed that relate the metadata to the physically-based, deterministic impacts of an observing change on the measurement of the climate.

8. Conclusions

In summary, climatic normals are intended to serve as a baseline for both spatial and temporal comparison. They are values derived from the frequencies of identically distributed observations over a 30-year period of record. At most locations, however, non-climatic influences such as station relocations, siting changes, instrument changes and recalibrations, etc. preclude obtaining a climatically homogeneous record of daily observations for 30 years. The statistical problem of detecting the full range of these inhomogeneities from the observational record is currently intractable. However, in the U.S., official, reasonably complete, and accurate metadata can be used to identify the dates of non-climatic changes to the observational record, and physically based decisions can be made to determine the impacts of these changes on climatic homogeneity. It is therefore recommended that periods of climatic homogeneity be determined solely by the use of metadata and not by statistical models that have yet to adequately describe observed processes, and that normals for a station should be computed for an observational record corresponding to the homogeneous period ending in the year 2000. In accordance with international guidelines, it is further recommended that normals be computed only if 10 or more observations are available in the homogeneous period of record, and that the set of normals include the mean and median as measures of central tendency; the standard deviation, mean deviation from the median, and quartiles or quintiles as measures of variability; and the highest and lowest 5 percent of the of the observations as indicators of extremes.

The deterministic rather than statistical approach to assessing homogeneity involves analyses of the processes that control climate at each individual site. Universal decision rules cannot be constructed because of the localized climatic effects of, for example, topography, proximity to water sources, and/or site environment (land use, instrument exposure, etc.). Network density is also a factor in that several sites in a local area provide more comparative information than only a couple of sites. The approach also allows for the possibility of data adjustments to lengthen a period of record if the effects of a non-climatic change can be analytically described with certainty.

It is recognized that these recommendations do not fulfill the goal of providing a set of summary statistics that are designed for temporal and spatial comparisons. The lengths of record will vary among sites so that comparisons cannot be made for a period of record that is uniform at all locations. However, normals that are calculated as recommended do provide a set of summary statistics that are based on unadulterated, official observational records, metadata, and deterministic physically-based reasoning. They are therefore uncontaminated by any potentially inaccurate assumptions about the climate that the data reflect or by any manipulations of the data that potentially result in an inaccurate description of the climate; all of the current, strictly statistical procedures that have been developed for homogeneity testing and data adjustment can lead to contamination since they are based on assumptions that cannot be validated.

The basic question is whether summary statistics of uncontaminated observations for varying periods of record are better than summary statistics of contaminated observations for a constant period of record. The answer depends on the use of the statistics. For the original purpose of the normals, i.e. comparisons of climatically homogeneous data for a uniform period record, neither set of statistics is palatable--one set describes climatically homogeneous data for varying time periods and the other set describes data for a uniform time period but for which climatic homogeneity cannot be reasonably ascertained. The above recommendations are therefore not based on the comparison use, but, instead, on the prediction use of the normals. Although it could easily be argued that normals should not be used for prediction, as Guttman (1989) and Kunkel and Court (1989) relate, they are in fact used for this purpose in, among others, the energy, agribusiness, construction, and insurance sectors of the economy. These authors also reference several studies showing that the latest period of between 10 and 20 years is optimal for predicting the climate of the next few years. These considerations tipped the balance in favor of the summary statistics of uncontaminated data with varying record lengths of at least 10 years.

9. References

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *J. Climatol.*, **6**, 661-675.

Bosshard, W. and M. Baudenbacher, 1997: Evaluation of various homogeneity tests by simulation of climatological time series. *In:* Proceedings of the First Seminar for Homogenization of Surface Climatological Data. Budapest, 6-12 October 1996, Hungarian Meteorological Service, 19-34.

Conrad, V. and L.W. Pollak, 1950: *Methods in Climatology*. Harvard Univ. Press. 459pp.

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

Guttman, N.B., 1994: On the sensitivity of sample L moments to sample size. *J. Climate*, **7**, 1026-1029.

Guttman, N.B., 1989: Statistical descriptors of climate. *Bull. Amer. Meteor. Soc.*, **70**, 602-607.

Guttman N.B. and M.S. Plantico, 1987: Climatic temperature normals. *J. Climate and Appl. Meteor.*, **26**, 1428-1435.

Heim, R.R, Jr. and N.B. Guttman, 1997: On computing 1971-2000 climate normals in the ASOS era. Proc. 10th Conf. On Appl. Meteor., Reno, 20-23 Oct., 1997, American Meteorological Society, 171-175.

Hosking, J.R.M. and J.R. Wallis, 1997: *Regional Frequency Analysis*. Cambridge Univ. Press, 224pp.

Hungarian Meteorological Service, 1997: Proceedings of the First Seminar for Homogenization of Surface Climatological Data. Budapest, 6-12 October 1996. 144pp.

Kohler, M.A., 1949: Double-mass analysis for testing the consistency of records for making adjustments. *Bull. Amer. Meteor. Soc.*, **30**, 188-189.

Kunkel, K.E. and A. Court, 1990: Climatic means and normals--A statement of the American Association of State Climatologists (AASC). *Bull. Amer. Meteor. Soc.*, **71**, 201-204.

Mielke, P.W., Jr. and K.J. Berry, 1998: Multivariate tests for correlated data in completely randomized designs. *J. Educational Behavioral Statistics.*, in press.

Mielke, P.W., Jr. and K.J. Berry, 1997: Permutation covariate analyses of residuals based on Euclidean distance. *Psychological Reports,***81**, 795-802.

Mielke, P.W., Jr. and K.J. Berry, 1994: Permutation tests for common locations among samples with unequal variances. *J. Educational Behavioral Statistics***, 81**, 795-802.

Peterson, T.C. and D.R. Easterling, 1994: Creation of homogeneous composite climatological reference series. *Intl. J. Climatol.*, **14**, 671-679.

Peterson, T.C. *et al.,* 1998: Homogeneity adjustments *in situ* atmospheric climate data: A review. *Intl. J. Climatol.*, in review.

Rhoades, D.A and M.J. Salinger, 1993: Adjustment of temperature and rainfall records for site changes. *Intl. J. Climatol.*, **13**, 399-913.

Snedecor, G.W. and W.G. Cochran, 1967: *Statistical Methods* (6th ed.). Iowa State Univ. Press., 593pp.

Solow, A.R., 1987: Testing for climate change: An application of the two-phase regression model. *J. Climate Appl. Meteor.*, **26**, 1401-1405.

Tuomenvirta, H. and H. Alexandersson, 1997: Review on the methodology of the Standard Normal Homogeneity Test (SNHT). *In:* Proceedings of the First Seminar for Homogenization of Surface Climatological Data. Budapest, 6-12 October 1996, Hungarian Meteorological Service, 35-45.

van der Waerden, B.L., 1969: *Mathematical Statistics*. Springer-Verlag. 367pp.

Vincent, L., 1998: Technique for the identification of inhomogeneities in annual temperature series. *J. Climate,* in press.

World Meteorological Organization, 1989: Calculation of monthly and annual 30-year standard normals. WMO-TD/No. 341. Geneva. 11pp.

World Meteorological Organization, 1983: Guide to climatological practices. WMO-No. 100. Geneva.

Zhang, N.F., 1998: A statistical control chart for stationary process data. *Technometrics*, **40**, 24-38.

Figure 1. Relationship between confidence interval bounds (α = .95) and sample size.

Figure 2. Confidence bounds (α = .95) as a function of sample size for a sample correlation coefficient of .90