Vol. 6: 33–43, 1996 | CLIMATE RESEARCH Clim Res | Published January 30

Statistical characteristics of U.S. Historical Climatology Network temperature distributions

Nathaniel B. Guttman*

National Climatic Data Center, 151 Patton Avenue, Asheville, North Carolina 28801, USA

ABSTRACT: The U.S. Historical Climatology Network (HCN) adjusted data are designed to aid in the investigation of climate change. The primary goal in the development of the adjustment procedures was to produce temporally homogeneous at-site data, with respect to averages, that could be used in the investigation of regional, long-term trends. The data, however, have been (and likely will continue to be) used for applications other than the detection of climate trends, and some of these other uses involve statistical properties of the data distributions. This study examines the change in distributional characteristics that result from the site location and urbanization adjustments to maximum and minimum temperature data in the HCN. It is shown that at most locations distributional shapes are changed when the data are adjusted. For climate change studies the differences may not be important, but for other studies involving distributional characteristics they could be important. In particular, probabilistic assessments of unusual climatic conditions that are used for design and planning purposes are highly dependent upon the shape of the frequency distributions. Changing the skewness and kurtosis, i.e. the shape of a frequency distribution, often changes the probability densities in the tails of the distribution, and therefore impacts the probabilistic assessments and the subsequent decisions that are based on these assessments.

KEY WORDS: Climatic temperature data · Statistical characteristics · Probability assessments

INTRODUCTION

The U.S. Historical Climatology Network (HCN) dataset has been compiled by the National Climatic Data Center (NCDC) for the purpose of providing an accurate, serially complete, unbiased climatic record that is suitable for detecting and monitoring climatic change over the past 2 centuries. These monthly precipitation and temperature (maximum, minimum and mean) data, according to the associated documentation (Karl et al. 1990), 'represent the best data available from the United States for analyzing long-term climate trends on a regional scale and may be used for studies attempting to determine the climatic impacts of increased concentrations of greenhouse gases.' The dataset is distributed to users by the Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory, Oak Ridge, TN.

The primary goal in the development of the adjustment procedures was to produce temporally homogeneous at-site data, with respect to averages, that could be used in the investigation of regional, long-term trends. The data, however, have been (and likely will continue to be) used for applications other than the detection of climate trends, and some of these other uses involve statistical properties of the data distributions.

For example, Stefanski & Andresen (1991) used the HCN data in the development of synthetic weather generators. The generators were required to produce synthetic data for input into crop yield analysis and management models. Estimators of the parameters of temperature distributions were important in the investigation of anomalous conditions of warmth.

For a second example, Meisner (1992), while investigating the magnitude, duration and spatial variations in weather elements that are directly related to fire severity in the Pacific Southwest, used the joint distributions of at-site HCN temperature and precipitation

*E-mail: nguttman@ncdc.noaa.gov

data. Of particular concern were warm and dry conditions, i.e. high fire potential. Gridded estimates of joint distributions were subjected to a principal component analysis to reveal spatial patterns.

The HCN dataset is currently being updated, and the record is being extended, by the NCDC. More than one version of the HCN temperature data will be available from CDIAC. Each of these versions will comprise the data resulting from different levels of processing, and each will have its own distributional characteristics.

The compilation of the temperature data is being accomplished in a 5-step process. First, the source data are quality assured with statistical outlier checks, areal edits, and manual comparison of the digital data with original manuscript records. Second, the quality-assured data are corrected for a sensor change to the Maximum-Minimum Temperature System (MMTS) instrumentation. Next, they are corrected for biases that result from differing times of observation during a day. The fourth step adjusts the data for known site location changes that have been documented in station history files. Finally, the data are adjusted for urbanization effects. Detailed information about these steps has been described by Quayle et al. (1991), Karl et al. (1986, 1988), and Karl & Williams (1987).

Because the data have been used for purposes other than those for which the dataset was created, i.e. studies other than climate change detection, and because some of these studies depend on the statistical parameters of the data, it is important to document differences in distributional characteristics of the data that result from different levels of processing and adjustment. This study examines changes in distributional characteristics that result from the site location and urbanization adjustments to maximum and minimum temperature data in the HCN.

Changes imparted by the switch to the MMTS sensors were not considered for 2 reasons. First, the switch occurred in the middle to late 1980s and thereby affects only a small portion of the long-term climatic record. Second, and more importantly, the effect of the switch is a known systematic average bias (Quayle et al. 1991) that can be removed from the climatic record. Similarly, the systematic biases induced by varying times of observation are known (Karl et al. 1986) and can be removed by adjusting the data so that they reflect midnight observation times. In addition, the software that applies the time of observation adjustments also includes quality assurance algorithms that evaluate the validity of some of the station history information, and by implication, the quality of the data.

In contrast, the site location change and urbanization adjustments are temporally and spatially dependent. The adjustments for site location changes are based solely on changes in means before and after a move. The adjustment factor is determined by comparing the record at the station that has moved with the most highly correlated records at nearby sites within the network (Karl & Williams 1987). No comparison was made of other distributional properties. The urbanization adjustment as described by Karl et al. (1988) is regression-based with population as the predictor. It therefore depends on time-dependent population changes. In addition, the urbanization adjustment is applied to individual site data in the compilation of the HCN data, but according to Karl et al. (1988), the adjustments are intended to be used with large-scale area temperature averages. The combined site location change and urbanization adjustments modified over 90% of the data.

DATA

Monthly mean maximum and minimum temperature data through 1992 from the HCN dataset that have been quality assured, corrected for MMTS sensor changes, and adjusted to a midnight time of observation are the initial data that were compared to the fully adjusted data, i.e. also adjusted for site location changes and urbanization. Using indicators appended to the data records, the maximum and minimum temperature data were independently constrained in the following sequence:

(1) The monthly mean temperature for a given year and month (year-month) was required to be based on a full month of daily data.

(2) A year-month mean was required to be, according to an indicator in the dataset, not suspect because of inaccurate or incomplete station history information.

(3) If a year-month datum for a site in the initial dataset was missing, the same year-month datum for the same site in the fully adjusted dataset was set to missing. Similarly, if a datum in the fully adjusted dataset was missing, the same year-month datum for the same site in the initial dataset was set to missing.

(4) The year-months of fully adjusted data at a site were required to be temporally coincident with the year-months of initial data at the same site.

(5) At each site, record lengths were required to be at least 60 yr, with no more than 10% of the months nor more than 12 consecutive months missing.

Constraints (1) and (2) are designed to eliminate suspect data as well as to eliminate monthly means that are based on less than a full month of daily data. The next 2 constraints insure that the periods of record of the initial and fully adjusted data were equal and also that a missing (non-missing) year-month datum in the initial dataset was coincidentally missing (non-miss-

ing) in the fully adjusted dataset. The last constraint insures that large temporal blocks are not missing and that record lengths at each site are adequate to calculate stable sample estimates of distributional characteristics (Guttman 1994).

The constrained temperature data were also converted from °F to K to avoid any potential problems that could arise from mixing positive, negative and zero data values in follow-on probability studies.

There are 1221 sites in the HCN. The constraints reduced the number of sites for which the analysis was made to 835 for maximum temperature and 824 for minimum temperature. The average record length is 82 yr for both constrained datasets. The longest length is 120 yr for maximum temperature and 122 yr for minimum temperature. The shortest length is 49 yr for maximum temperature and 52 yr for minimum temperature.

DISTRIBUTIONAL CHARACTERISTICS

Distributional characteristics are often described by moment estimation of a distribution's parameters since

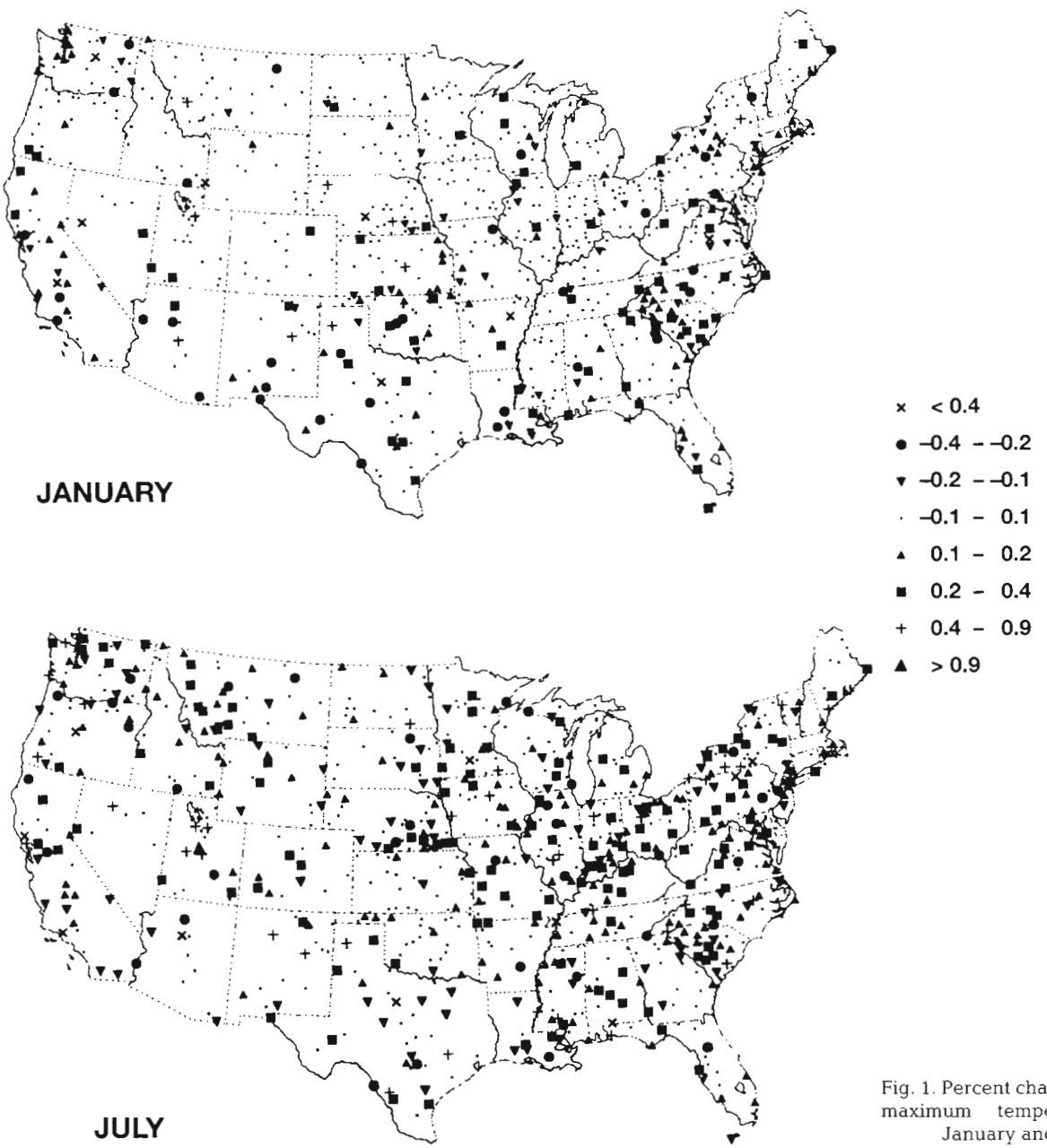

Fig. 1. Percent change in mean maximum temperature for January and July

a given set of moments define a unique probability density function. Estimates of the parameters of a distribution are obtained by equating sample moments with theoretical moments. The resulting nonlinear system of equations are then solved.

Traditionally, theoretical product moments, which are power functions, have been equated to the sample product moments. Another class of moments, called L-moments, can also be used to characterize a density function. L-moments are linear combinations of order statistics. The derivation and properties of L-moments are described by Hosking (1990). Because L-moments involve only linear combinations of the data, and do not require raising the data values to higher powers, they are less sensitive than the conventional product moments to the numerical values of the most extreme observations. This and other advantages of L-moments have been demonstrated by several authors (Hosking 1990, 1992, Royston 1992, Vogel & Fennessy 1993).

Since they are more advantageous than conventional product moments, sample L-moments are defined as the distributional characteristics of interest in this study. Ordinarily, 4 sample L-moments are calculated. The first L-moment, L-1, is the mean. The second

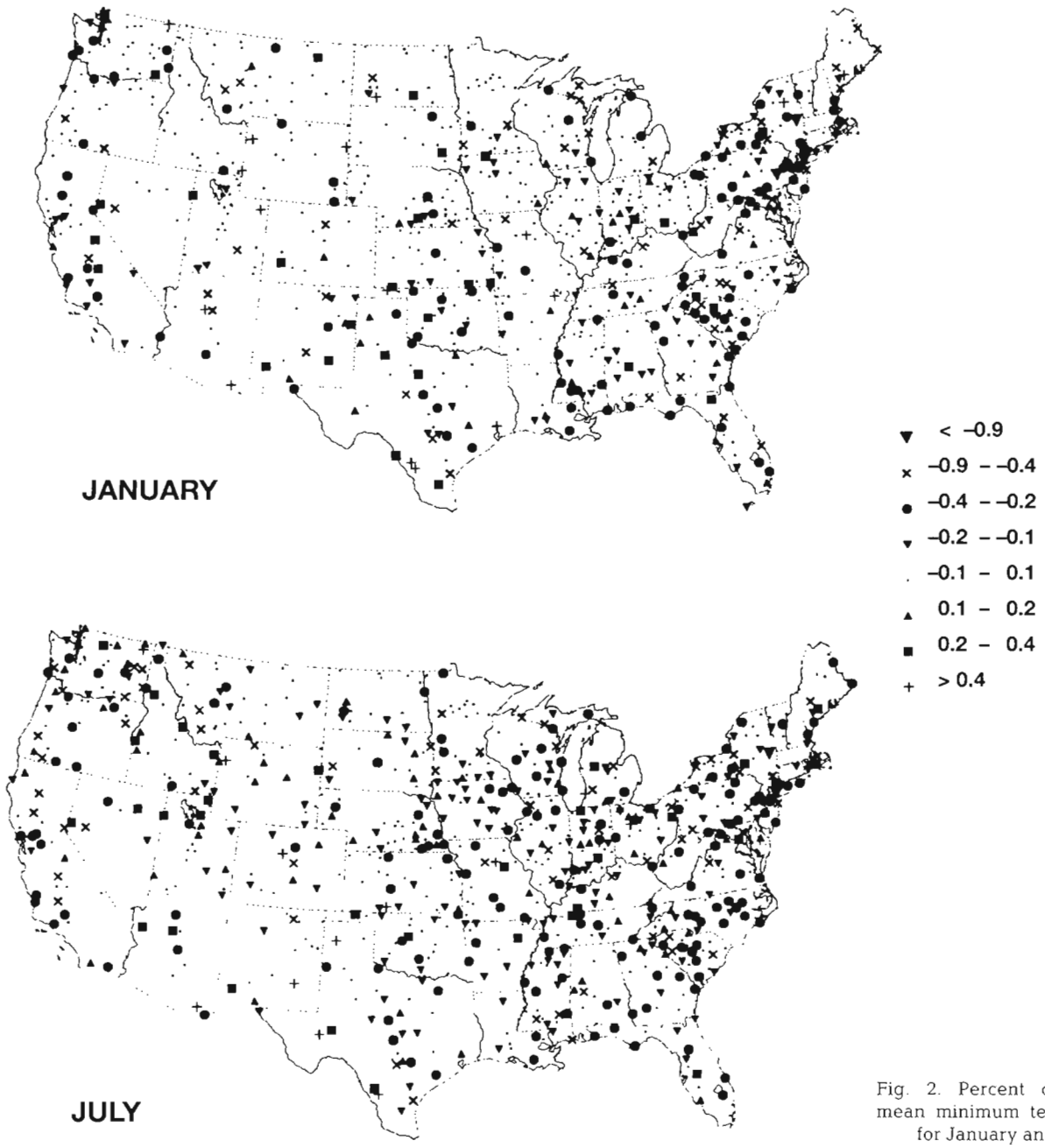

Fig. 2. Percent change in mean minimum temperature for January and July

L-moment, L-2, is a measure of dispersion that is analogous, but not equal to, the standard deviation. Usually, it is divided by the mean and called the L-CV, a quantity that is analogous to the coefficient of variation and represents a standardized measure of dispersion relative to the mean. L-moment measures of skewness and kurtosis, T3 and T4, are analogous to conventional measures of skewness and kurtosis. The change in L-1, L-CV, T3 and T4 from the initial to the fully adjusted January, April, July and October monthly mean maximum and minimum temperatures at each site have been calculated and represented as a percentage. A positive percentage indicates that the adjusted values are higher than the initial data.

RESULTS

The January and July percentage changes in mean maximum temperatures at each site are shown in Fig. 1, and changes in minimum temperatures are shown in Fig. 2. Changes for April and October are not shown. Percentage changes in the mean are small because the unit of temperature is K.

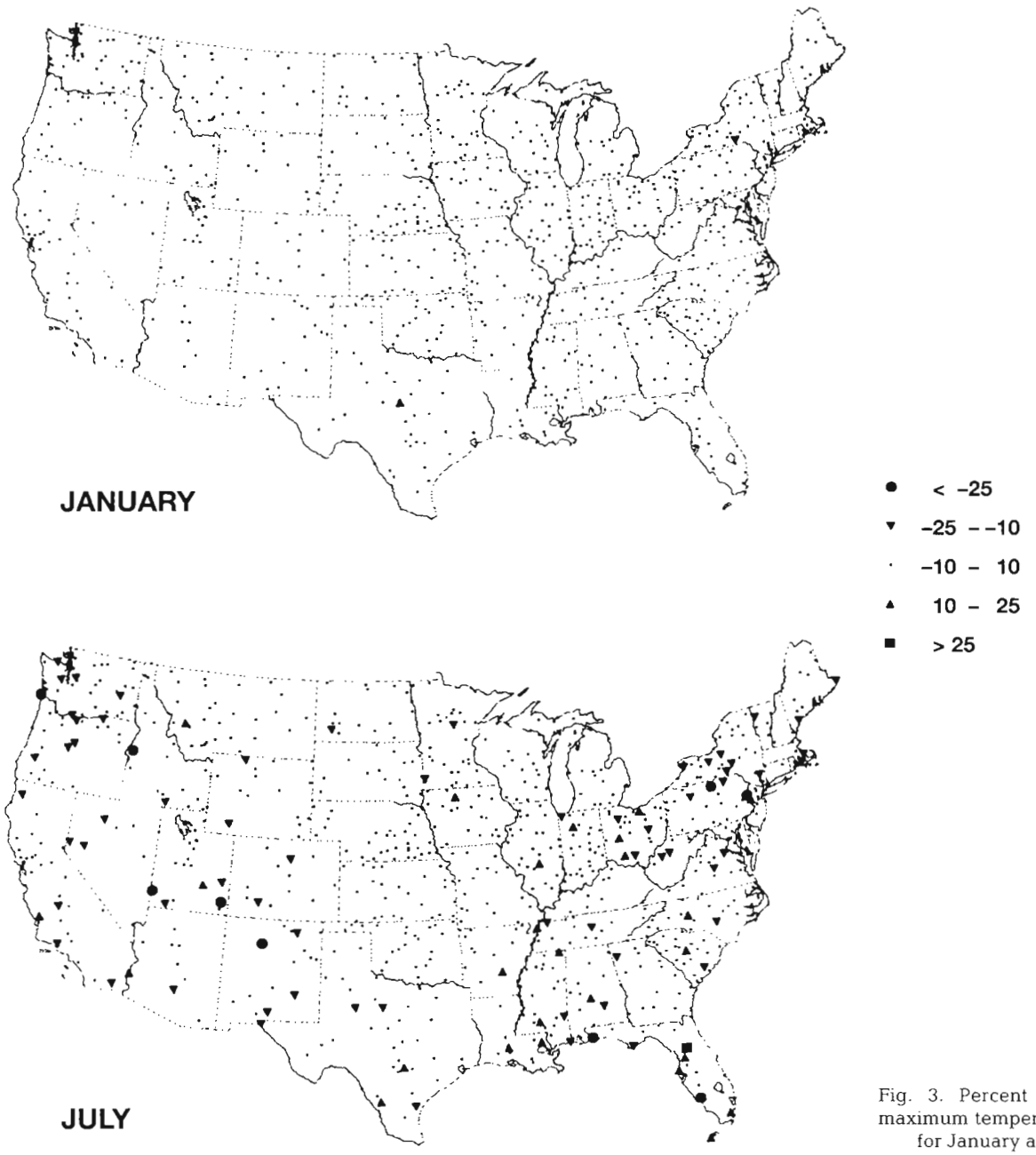

Fig. 3. Percent change in maximum temperature L-CV for January and July

Changes of more than 0.2% of maximum temperature means occur for between 9 and 10% of the sites in January, April and October. In July changes of this magnitude occur for 16% of the sites. For sites with changes of at least 0.2%, the adjustments increase the mean at slightly more than half the locations in January and April, but in July and October, the mean decreases at twice as many sites as those for which the mean increases.

Changes of more than +0.2% in the minimum temperature means occur at less than 4% of the sites in all 4 months. Negative changes of this magnitude, however, occur at about 20% of the sites in January and April and at about 27% of the sites in July and October.

Changes in the measure of relative variability L-CV of maximum and minimum temperatures are shown in Figs. 3 & 4. For maximum temperatures, most of the changes are minor and tend to be decreases. For minimum temperatures, the January changes are minor. In April, small changes are noted along both the East and West Coasts, and most of these changes reflect a decrease in relative variability. July changes are scattered throughout all sections of the U.S. except the Northern Plains. The largest positive changes are in

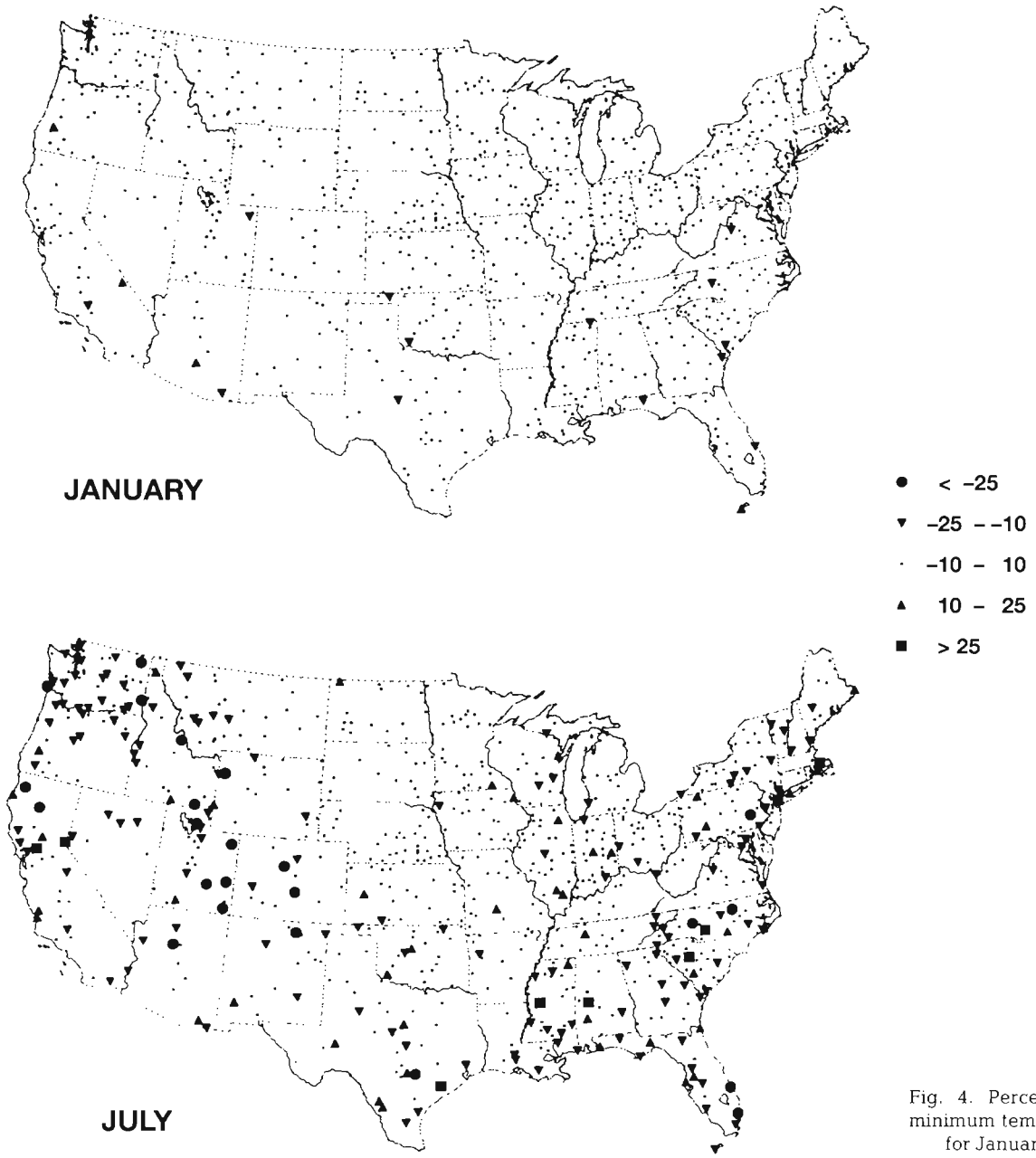

Fig. 4. Percent change in minimum temperature L-CV for January and July

the East, and the largest negative changes are in the West. In October, most of the changes are in the western third of the country and tend to be negative.

Substantial changes in the measure of skewness T3 are illustrated in Figs. 5 & 6. For maximum temperatures, changes of more than 50% occur at about 14% of the sites in January, and in slightly more than one third of the sites in April, July and October. For each of the 4 months, a little more than half of these changes are negative (decrease in skewness). The number of sites with changes in excess of ±50% is higher for minimum temperatures than for maximum temperatures. In January, 20% of the sites exhibit changes of this magnitude, and in the other 3 months, slightly more than half of the sites exhibit these large changes.

Figs. 7 & 8 show changes in the measure of kurtosis T4. For maximum temperature, changes are minimal in January. In April, July and October, the larger changes, which are predominantly positive, are scattered through the Southeast and along the West Coast. More changes in kurtosis occur with minimum temperatures than for maximum temperatures, and they are scattered throughout the country. In all 4 months the largest changes are predominantly positive.

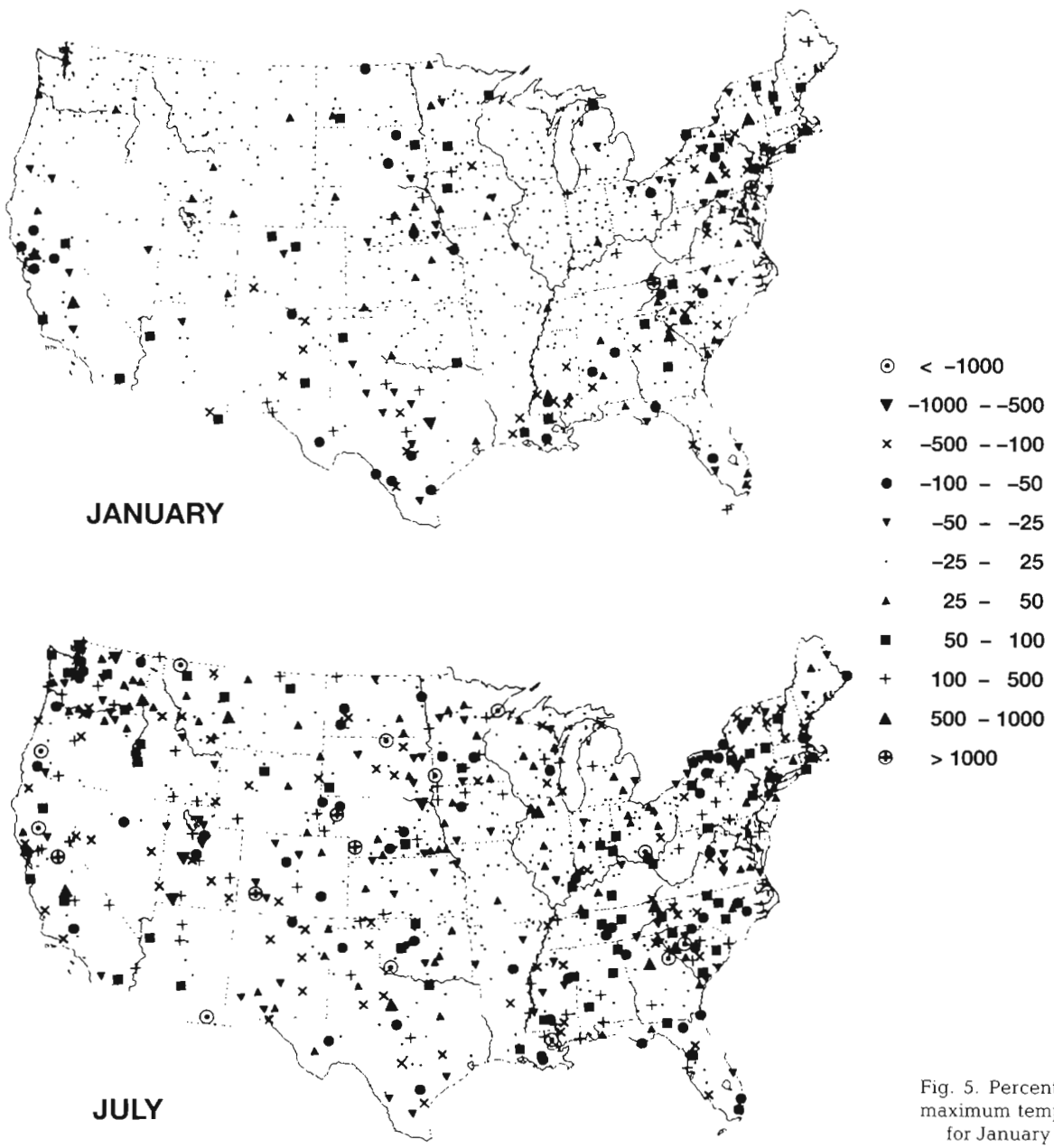

Fig. 5. Percent change in maximum temperature T3 for January and July

DISCUSSION

The results show that for both maximum and minimum temperature, the site location and urbanization adjustments change the distributional characteristics of the data. Changes in the mean are expected because the adjustment factors for a particular datum depend on the dates of site location changes, the rate of population growth, and the relationships among the datum and those at neighboring locations. These dependencies are not constant over the entire length of record at a site, nor are they necessarily the same among sites.

Changes in the L-CV indicate that there is a seasonal cycle to the number of sites with material changes in relative variability. The effects of the adjustments on both maximum and minimum temperature are minimal in January and maximal in July. Most, but certainly not all of the changes, are negative. The L-CV was expected to be reduced by the adjustments since variability is usually reduced by homogenizing the temporal record at a site.

In all 4 months for both variables, the skewness changes the most. The shapes of the frequency distributions are therefore altered. A change in skewness also implies that the distribution of data values that are

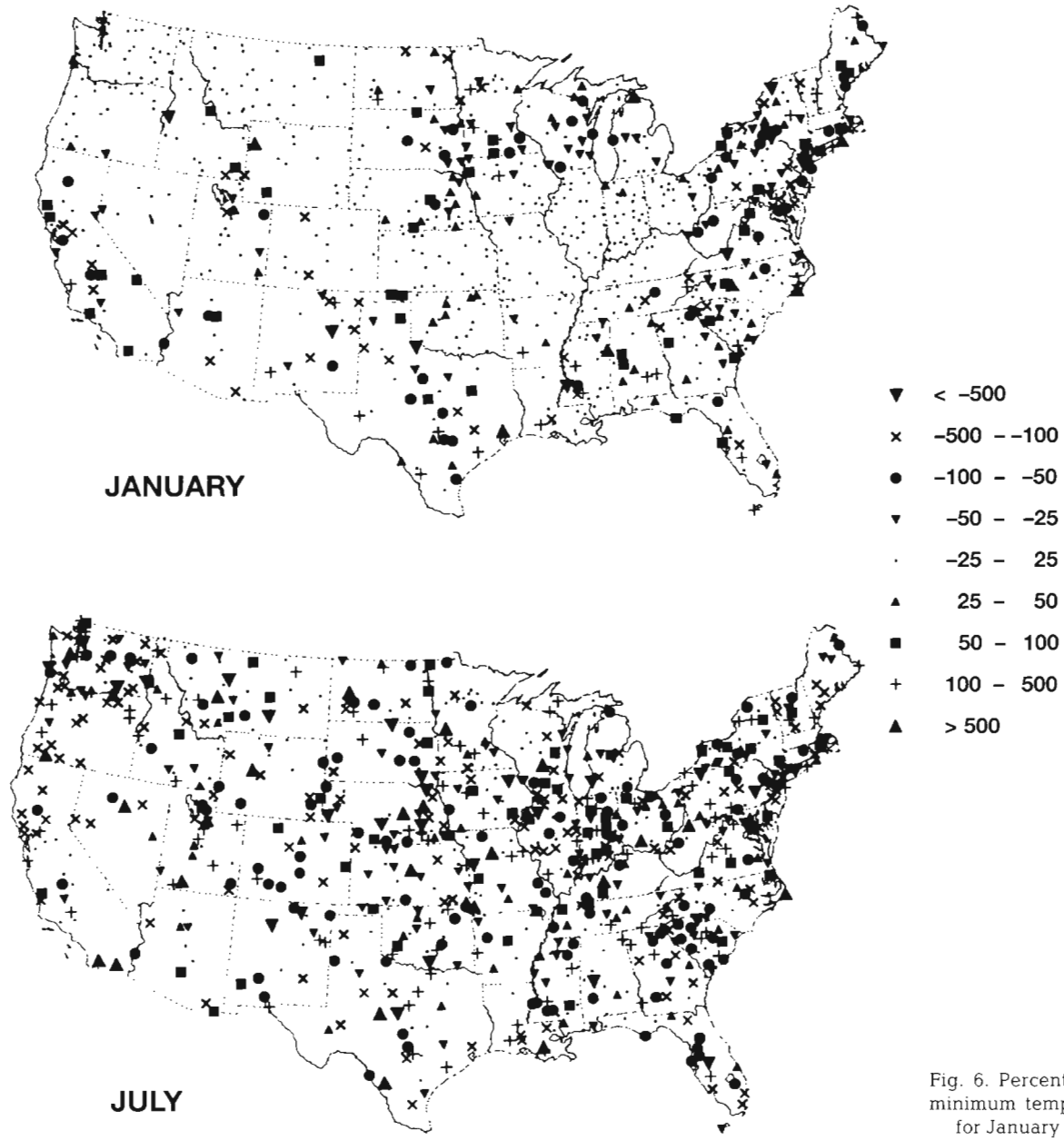

Fig. 6. Percent change in minimum temperature T3 for January and July

separated from the mean (anomalous events) is changed. In view of the magnitude of the changes, probability assessments based on the adjusted data for values in the tails of the distributions are likely to be markedly different from those based on the initial data.

Changes in the measure of kurtosis T4 also support the altering of the shape of the frequency distributions. It is readily apparent from a comparison of the T4 maps that the adjustment processes have more of an effect on the minimum than on the maximum temperature distributions. Since the urban adjustment affects the minimum temperatures much more than the maximum temperatures (Karl et al. 1988), this adjustment could be causing the patterns depicted on the maps.

The impact of using the fully adjusted data, rather than the initial data, in the Stefanski & Andresen (1991) study involving weather generators is likely to be minimal. Cumulative frequency distribution curves were used to differentiate among 3 scenarios: above normal, normal, and below normal. The 30th and 70th percentiles determined the threshold criteria for the scenarios. In this application, a broad central tendency was used so that small differences in the mean or median were likely to be unimportant. Also, the tails of the distributions (less than the 30th and more than the 70th percentile) were not important so that changes in skewness and kurtosis were likely to be insignificant for the purposes of the study.

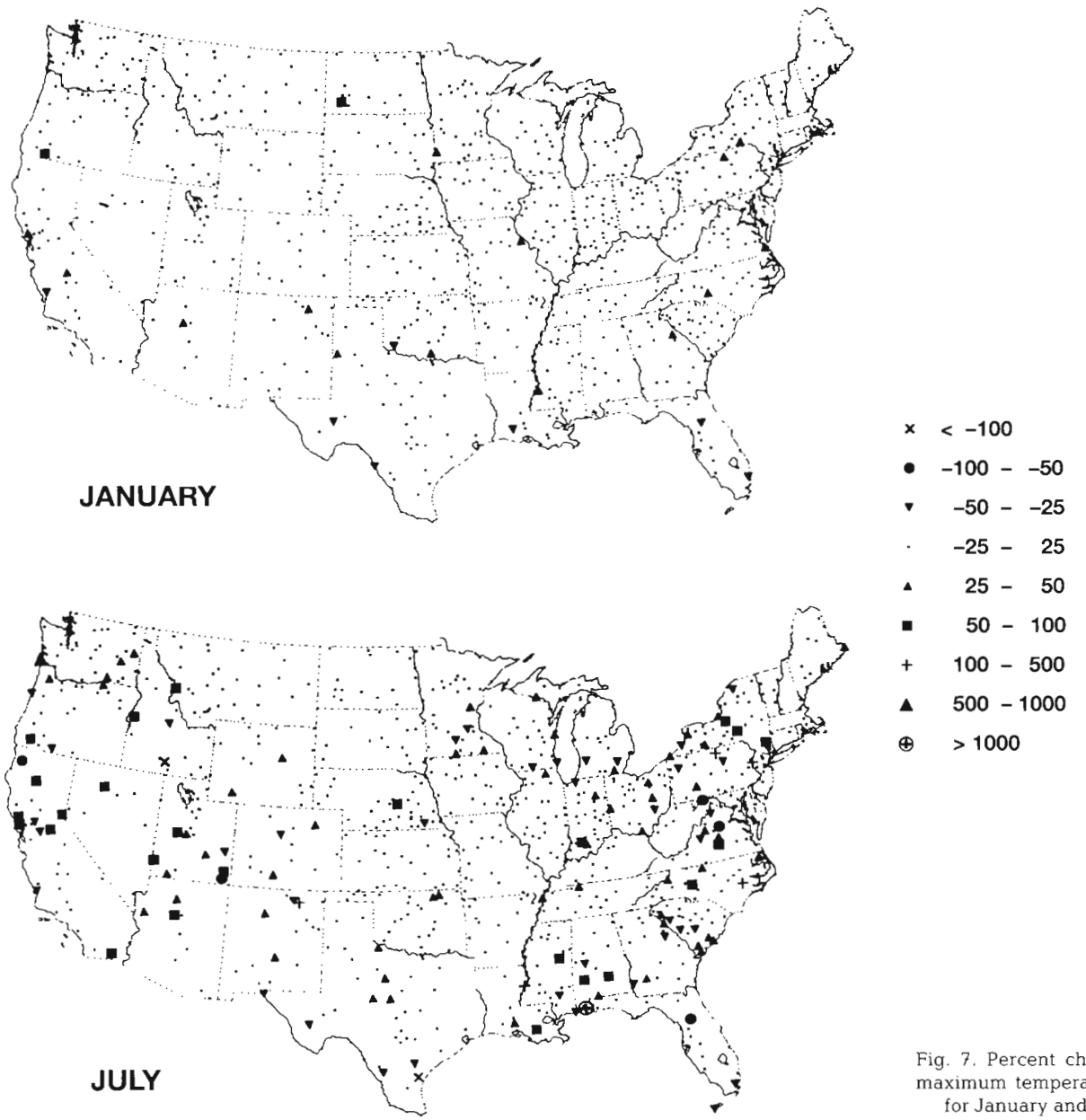

Fig. 7. Percent change in maximum temperature T4 for January and July

The Meisner (1992) study used the initial data. Temperature data were grouped into above- and below-median categories as well as into quartile categories. In this application, skewness, and by implication, asymmetry of the distribution, is important since it affects the value of the median. If the fully adjusted data were used, different results may have been obtained because it is likely that the median values (and perhaps the quartile values) would have shifted thereby altering the frequencies in each category.

In applications-oriented analyses of unusual climatic events, the tails of the probability distributions are of primary concern, and it is in these areas of the distributions that changes of skewness and kurtosis have the most impact. The HCN data could be used, for example, by the natural gas industry for assessing the expected frequency of occurrence of prolonged, severe, hot and cold waves. Since temperature extremes increase the demand for natural gas, suppliers must plan for adequate storage as well as for adequate distribution of the gas in order to meet the increased demand. Knowledge of the expectation of unusual events is important in this planning process; probabilistic assessments that are in error could lead to an imbalance between supply and demand of natural gas.

Another example concerns the use of water. During droughts, high temperatures over an extended period of time increase the demand for water for irrigation, horticultural activities, watering lawns and golf courses, and residential uses. Knowing the risks asso-

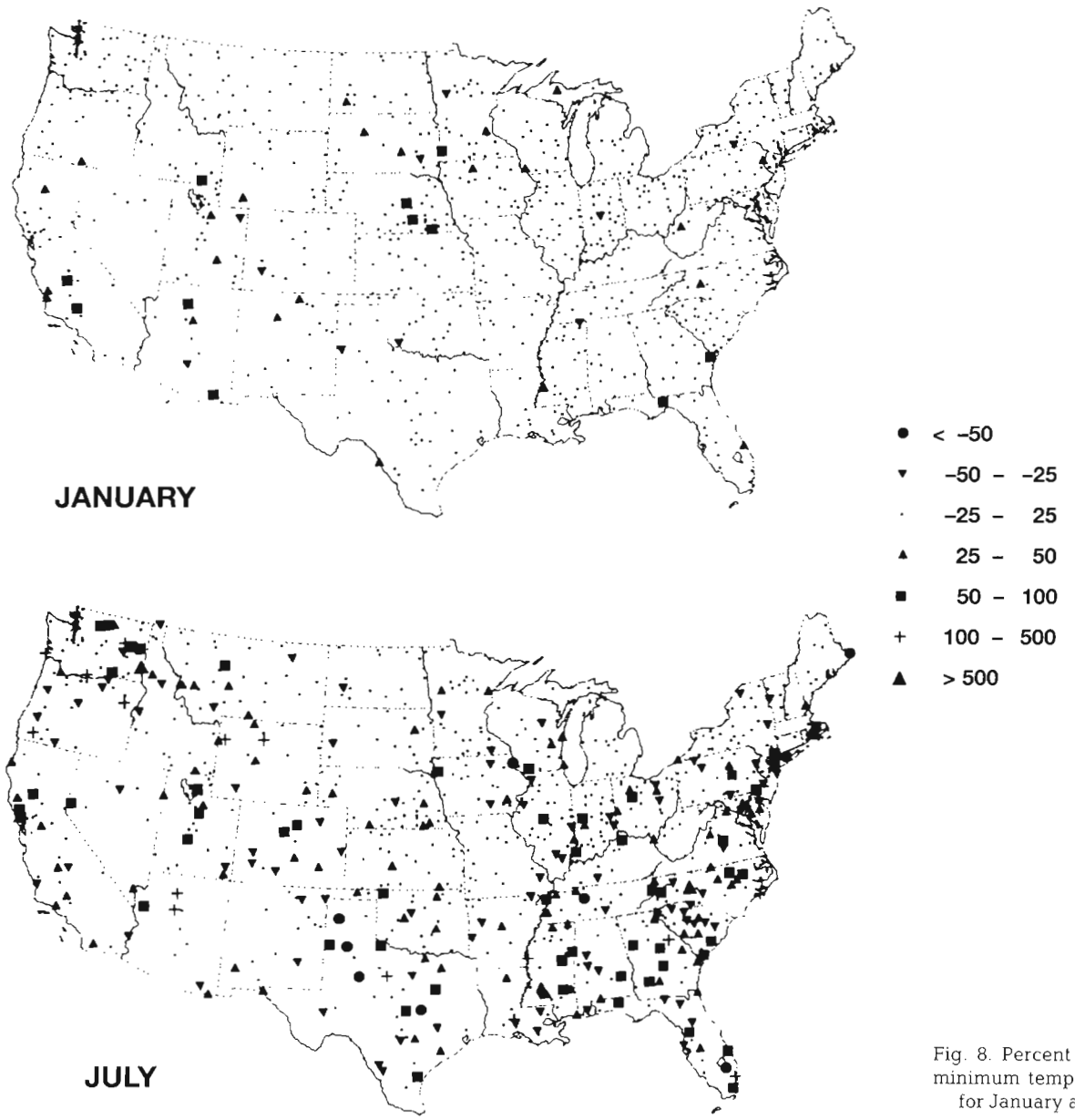

Fig. 8. Percent change in minimum temperature T4 for January and July

ciated with critical temperatures is useful in the design of water storage facilities as well as in determining thresholds at which water conservation measures should be implemented.

CONCLUSION

This study has shown that at most locations, distributional shapes are changed by applying the site location move and urbanization adjustments to the HCN data. For climate change studies, the compilers of the HCN dataset assume that the differences may not be important, but for other studies involving distributional characteristics, these differences could be important. In particular, risk analyses of climatic data, which concern probabilistic assessments of unusual conditions that are used for design and planning purposes, are highly dependent upon the shape of the frequency distributions. Changing the skewness and kurtosis, i.e. the shape of a frequency distribution, often changes the probability densities in the tails of the distribution, and therefore impacts the probabilistic assessments and the subsequent decisions that are based on these assessments.

It is therefore suggested that, if researchers plan to use the HCN data for studies dependent on the characteristics of temperature frequency distributions, care should be exercised in the selection of the version of the data that will be analyzed. In terms of the objectives and methodology of the research, factors to consider include whether (1) the adjustments reflect the real-world temperature distributions, (2) the temporal and spatial assumptions of the adjustment procedures are reasonable, (3) the adjustment methodology is appropriate, and (4) the changes in the moments are significant.

These 4 factors can be generalized in the sense that the caveats are applicable to any secondary user of a dataset. It is the responsibility of the dataset compilers to document the details of their methodologies. However, it is the responsibility of the user to use the documentation to determine if the dataset is sufficient for his intended analyses so that valid conclusions can be reached.

Editor: V. Meentemeyer, Athens, Georgia, USA

Acknowledgements. This work was supported by the National Climatic Data Center and the U.S. Department of Energy through Interagency Agreements DE-AI05-90ER60952 and DE-AI05-900R21956. The enlightening discussions with Drs Tom Peterson and David Easterling of the National Climatic Data Center regarding the HCN data processing are appreciated.

LITERATURE CITED

Guttman NB (1994) On the sensitivity of sample L-moments to sample size. J Clim 7:1026–1029

Hosking JRM (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. J R Stat Soc B 52:105–124

Hosking JRM (1992) Moments or L-moments? An example comparing two measures of distributional shape. Am Stat 46:186–189

Karl TR, Diaz HF, Kukla G (1988) Urbanization: its detection and effect in the United States Climate Record. J Clim 1: 1099–1123

Karl TR, Williams CN Jr (1987) An approach to adjusting climatological time series for discontinuous inhomogeneities. J Clim appl Meteorol 26:1744–1763

Karl TR, Williams CN Jr, Quinlan FT, Boden TA (1990) United States Historical Climatology Network (HCN) serial temperature and precipitation data. ORNL/CDIAC-30, NDP-019/R1. Carbon Dioxide Information Analysis Center, Oak Ridge National Lab, Oak Ridge, TN

Karl TR, Williams CN Jr, Young PY, Wendland WM (1986) A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States. J Clim appl Meteorol 25: 145–160

Meisner BN (1992) Spatial and temporal variations in fire climate. In: Proc 5th International Meeting on Statistical Climatology, Toronto. Atmospheric Environment Service, Downsview, Ontario, p 529–532

Quayle RG, Easterling DR, Karl TR, Hughes PY (1991) Effects of recent thermometer changes in the Cooperative Station Network. Bull Am Meteorol Soc 72:1718–1723

Royston P (1992) Which measure of skewness and kurtosis are best? Stat Med 11:333–343

Stefanski RJ, Andresen JA (1991) Development and analysis of climate scenario statistics for several sites throughout the US Cornbelt. In: Proc 7th Conf Applied Climatology. American Meteorological Society, Boston, MA, p 43–46

Vogel RM, Fennessy NM (1993) L-moment diagrams should replace product-moment diagrams. Water Resour Res 29: 1745–1752

Manuscript first received: June 16, 1995
Revised version accepted: September 30, 1995