Updated Population Metadata for United States Historical Climatology Network Stations

TIMOTHY W. OWEN

National Climatic Data Center, Asheville, North Carolina

KEVIN P. GALLO

Office of Research and Applications, NOAA/NESDIS, Camp Springs, Maryland

(Manuscript received 11 October 1999, in final form 17 February 2000)

ABSTRACT

The United States Historical Climatology Network (HCN) serial temperature dataset is comprised of 1221 high-quality, long-term climate observing stations. The HCN dataset is available in several versions, one of which includes population-based temperature modifications to adjust urban temperatures for the "heat-island" effect. Unfortunately, the decennial population metadata file is not complete as missing values are present for 17.6% of the 12 210 population values associated with the 1221 individual stations during the 1900–90 interval. Retrospective grid-based populations, within a fixed distance of an HCN station, were estimated through the use of a gridded population density dataset and historically available U.S. Census county data. The grid-based populations for the HCN stations provide values derived from a consistent methodology compared to the current HCN populations that can vary as definitions of the area associated with a city change over time. The use of grid-based populations may minimally be appropriate to augment populations for HCN climate stations that lack any population data, and are recommended when consistent and complete population data are required. The recommended urban temperature adjustments based on the HCN and grid-based methods of estimating station population can be significantly different for individual stations within the HCN dataset.

1. Introduction

The United States Historical Climatology Network (HCN) comprises 1221 high-quality, long-term climate observing stations. The HCN serial temperature dataset (Easterling et al. 1996) is updated periodically and includes monthly temperature data that have been adjusted to remove biases caused by changes in the time of observation, changes in instruments, movement of weather stations, and urbanization. The HCN dataset is available in several versions that include no adjustments, one or more of the adjustments, or all of the adjustments. The HCN dataset is a potent tool for climate researchers, given both its long retrospective data offerings for many stations and these multiple inhomogeneity adjustments that are derived from a rich and extensive station metadata.

Presently, the HCN dataset includes an adjustment for urban heat-island biases using decennial population values for each station. The population values were derived from U.S. Census Bureau tabular data. These population values have been useful in research on the effects of urbanization on temperature data (Karl et al. 1988; Gallo et al. 1999). The population data associated with the stations are also used operationally in the preparation of the HCN dataset for estimating the effect of urbanization on temperature with the methodology of Karl et al. (1988). Unfortunately this population metadata file is not complete, as missing values are present for 17.6% of the 12 210 decennial population values associated with the 1221 individual stations during the 1900–90 interval. The distribution of these missing values by decade indicates a range in the percent of stations with missing values from 34% in 1900 to 12% in 1990 (Table 1).

Throughout the 1900–90 interval a total of 523 HCN stations have at least one missing population value. Several stations displayed more than one missing value during the interval, which resulted in a total of 2144 missing values within the HCN dataset. These missing values are likely the result of unavailable or incomplete population information, particularly in the earlier decades. Preliminary evaluations have revealed that some of these missing values are associated with stations in highly populated metropolitan areas.

Population values for surrounding political jurisdic-

Corresponding author address: Kevin P. Gallo, USGS EROS Data Center, 47914 252nd Street, Sioux Falls, SD 57198.
E-mail: Kevin.P.Gallo@noaa.gov

TABLE 1. Number of original HCN stations, by decade, with a missing population value.

Decade	Number	Decade	Number
1900	413	1950	174
1910	292	1960	171
1920	234	1970	161
1930	209	1980	144
1940	205	1990	141

tions (e.g., town, city, township, county) may not accurately reflect the urban concentration (and associated anthropogenic vs natural land cover) located in the vicinity of a climate observation station. While Karl et al. (1988) demonstrated that population could be used as a surrogate for urban land cover, Gallo et al. (1996) found that urban land cover itself had a statistically significant influence on temperature observations up to 10 km from a station. This paper describes a methodology that (i) facilitates population estimation within fixed distances (Grid cells) centered on HCN stations, and (ii) provides retrospective grid-based estimates of decennial population for the HCN stations. The implications of the use of grid-based estimates for urban temperature adjustments are also presented.

2. Methodology

Two datasets were used for the grid-based estimates of decennial population at the HCN stations. The first is a 1 km × 1 km gridded population dataset developed for the conterminous United States by the Socioeconomic Data and Applications Center (SEDAC 1996). This dataset is based on 1990 U.S. Census Bureau tract-level data, aggregated at a spatial resolution of 1 km × 1 km. This dataset provided direct estimates of the 1990 population associated with the HCN stations.

In order to retrospectively estimate grid-based populations within a fixed distance of an HCN station, historically available (1900–90) tabular U.S. Census county data (U.S. Bureau of Census 1996) were utilized. The tabular county data, however, had to be linked to the spatial population density within each county. This was accomplished in two steps. First, the 1 km × 1 km gridded population density dataset for the conterminous United States was merged with a 1 km^2 gridded dataset that defined the boundaries of all counties. Both gridded datasets were current for 1990. The combination of the gridded spatial population density and county boundary datasets provided for estimation of the spatial distribution of population within each county.

With spatial population densities determined within each county, the decennial tabular county population data for 1900–90 were then allocated for each decade and each 1 km × 1 km pixel. This allocation assumed that population density patterns were constant over the 1900–90 interval (i.e., population centers were constant). Finally, based on the results from the spatial analysis of Gallo et al. (1996), populations for each 1 km × 1 km cell of a 21 km × 21 km grid cell centered on each HCN station were summed to provide grid-based population estimates for each station for each decade from 1900 through 1990.

The initial HCN population dataset was input into operational algorithms used by the National Climatic Data Center (NCDC) to generate HCN population estimates for decades where missing values existed. The NCDC algorithm fills in missing values for any station that has at least one nonmissing decadal population value. Missing population values preceding or succeeding nonmissing values that are not between two nonmissing values are estimated by iteratively multiplying or dividing the nonmissing value by 0.9, respectively, while missing values between two nonmissing values are linearly interpolated. Thus, for the HCN dataset, population values and estimates from the NCDC algorithm (operational HCN) were available to calculate an adjustment for all stations except the 128 stations with no population values at all.

In addition to the population estimates for the original HCN data, a "Hybrid" dataset was created by selectively replacing missing HCN values with the grid-based population data for only the 128 HCN stations with no population values. This Hybrid dataset incorporated grid-based data for the 1280 missing values that remained after the NCDC operational adjustments to the HCN population dataset. Combined with the NCDC operational population estimates, the Hybrid population dataset was serially complete.

Population-based temperature adjustments were then estimated for each (i) the operational HCN, (ii) the Hybrid, and (iii) the grid-based population datasets. The adjustments are computed according to the equations given in Table 2 of Karl et al. (1988), Adjustment = a × (Population)$^{0.45}$, where a is a seasonal-based coefficient.

3. Results and discussion

a. General

Grid-based population estimates were made for the 21 km × 21 km grid cell size for all decades. The

TABLE 2. Mean, standard deviation (s), minimum and maximum differences between grid-based and original HCN populations, coefficient of determination (r^2), and the coefficient of correlation (r) between the original HCN and grid-based population values (n = 10 066).

	Grid-based and original HCN population difference					
Grid cell size	Mean	s	Minimum	Maximum	r^2	r
21 km × 21 km	−1753	293 745	−14 153 929	988 044	0.73	0.85

TABLE 3. Distribution of population values by population class (e.g., rural) used by Karl et al. (1988) for I) operational HCN population values, II) hybrid population values, and III) grid-based population values. Parenthetical percentages are based on individual column totals.

	Full datasets		
Range of estimated population	I. Operational	II. Hybrid	III. Gridded
<2000 (rural)	3860 (35%)	4652 (38%)	2337 (19%)
2000–9999 (small urban)	4534 (42%)	4840 (40%)	4417 (36%)
10 000–100 00 (medium urban)	2176 (20%)	2336 (19%)	4949 (41%)
>100 000 (large urban)	360 (3%)	382 (3%)	507 (4%)
Total	10 930	12 210	12 210

gridded populations for each 21 km × 21 km grid cell, centered on the location of each observation station, were compared with the original HCN (no operational adjustments) population data (Table 2). The 21 km × 21 km grid cell size slightly underestimates population compared to the HCN metadata values (Table 2) with a mean difference of approximately 2000. The relatively large standard deviations of the difference between the grid-based and HCN population values might be expected because of the inherent differences in the two methodologies used to determine the population values.

The greatest grid-based underestimation (14 153 929) was for the 1990 population associated with the New York—Central Park observation station. The grid cell estimate for the population of the station was 3 933 071, compared to 18 087 000 based on the HCN population data. The greatest grid-based overestimation (988 044) was for the 1970 population associated with the College Park, Maryland, station. The grid-based estimate for the population of the station was 1 014 200 compared to 26 156 based on the HCN population data.

The comparisons include populations within a consistently defined grid cell size to that of an often irregularly shaped boundary associated with a city. Additionally, several stations are located in small suburbs (e.g., College Park, Maryland) that are surrounded by a densely populated region. For example, the 1990 HCN listed for the station located in College Park, Maryland, is 21 927, while the 21 km × 21 km estimated population for the station is 1 001 896. The Washington D.C.–Baltimore consolidated metropolitan statistical area included a population of 6 726 395 in 1990 (U.S. Bureau of Census 1997a). In contrast to the College Park, Maryland, example, many of the HCN stations located in large urban areas generally would have the population of the entire urban area designated as their HCN population (e.g., New York—Central Park). The estimates from the grid-based methodology would include only the population within the consistently defined 21 km × 21 km grid cell that surrounds the station.

The impact of the significant population underestimation in the College Park, Maryland, example is reflected in the urban-related temperature adjustments. The greater population for College Park associated with the grid-based dataset results in an urban-adjusted decrease in the maximum temperature and an increase in the minimum and average temperatures, compared with the adjustments based on the HCN population values. With the HCN population value of 21 927, the 1990 annual maximum, minimum, and mean adjustments are −0.035°, 0.3243°, and 0.1635°C, respectively. The grid-based 1990 annual maximum, minimum, and mean adjustments, based on a population for College Park of 1 001 896, are −0.1956°, 1.8108°, and 0.9129°C. Opposite examples also appear in comparisons of the HCN and grid-based datasets.

Grid-based population estimates for the 2144 missing values in the original HCN metadata ranged from 0 (417 stations) to 1 211 516 (Chestnut Hill, Massachusetts). The distribution of these grid-based populations were grouped by the population categories used by Karl et al. (1988) to identify the number of stations associated with rural, small urban, medium urban, and large urban environments. Most of the missing population values were estimated as less than 10 000, however, 18% were greater than or equal to 10 000 and nearly 3% were greater than 100 000. Table 3 shows the distribution of I) operational HCN population values (including estimates using the "NCDC estimate" as described in section 2), II) the "hybrid" dataset, and III) the grid-based population dataset.

The decadal distribution of urban temperature adjustments are shown in Fig. 1 for the three population datasets. The bias of the adjustment is slightly less amplified for the hybrid datasets for all decades, compared with the operational HCN or gridded datasets. This is due in part to the large number of low population values being introduced to the hybrid dataset when the operational HCN displayed missing values (86% of the population values, replacing missing HCN population values, are less than 10 000). This suggests that many HCN values were initially missing because of their siting in rural locations, where population statistics were more difficult to ascertain (especially in the early decades of the twentieth century). In contrast, the gridded dataset is uniformly more amplified than the HCN dataset. This is due to the more urban (small, medium, or large urban) composition of the gridded dataset (81%) compared to the HCN (65%) or Hybrid (62%) datasets.

b. Case studies

Five stations in the northeast region of the United States required population estimates for all or most of

the 1900–90 interval. Three stations in the vicinity of Boston, Massachusetts, and their grid-based populations for the 1990–90 interval were compared, as were two additional cities for which HCN population values were available. The three stations that required population estimates included Bedford, Blue Hill, and Chestnut Hill, Massachusetts (Fig. 2 and Table 4). The two additional stations examined included Providence, Rhode Island, and Durham, New Hampshire (Fig. 2 and Table 5).

The area associated with the 21 km × 21 km grid cell sample of the Chestnut Hill station is displayed in Fig. 2. The Chestnut Hill station is located in a region of Boston with a high population density, compared to the Blue Hill or Bedford stations. Thus, the estimated population values of the Chestnut Hill station, as expected, are greater than the other two stations throughout the examined interval (Table 4). The Blue Hill station is located in a transition zone of population densities that include the densities greater than 1000 km^{-2} while the 21 km × 21 km grid cell sample associated with Bedford includes densities less than 100 km^{-2}. This resulted in population estimates for the Blue Hill station that are minimally twice as great as the Bedford station throughout the 1900–90 interval (Table 4).

The current HCN and grid-based populations for Providence, Rhode Island, present some interesting differences (Table 5). The two populations are similar from 1900–40, after which the HCN population values increase considerably. The increase in the HCN population was not necessarily due to a sudden influx of individuals into the Providence region, but more likely due to a change in the boundaries (and area) associated with the HCN population. The term "standard metropolitan area" (SMA) was introduced by the Bureau of the Budget (now known as the U.S. Office of Management and Budget) in 1949. The SMA boundaries are often defined to include the county in which the largest city resides, as well as adjacent counties.

At least three population estimates are available from the U.S. Bureau of Census for use with the HCN station associated with Providence, Rhode Island. The 1990 HCN value (Table 5) was the value provided for the Providence–Warwick–Pawtucket, Rhode Island, New England county metropolitan area (U.S. Bureau of Census 1997b). Additionally, a value of 160 728 is available for "Providence city" (U.S. Bureau of Census 1997b) and a value of 1 134 350 is available for the Providence–Fall River–Warwick, Rhode Island–Massachusetts, metropolitan statistical area (U.S. Bureau of Census 1997a). The grid-based estimates of the population associated with the HCN station located in Providence, would appear to offer a more consistent measure of the population.

The Durham, New Hampshire, example suggests that the original HCN and grid-based measures of population can be incompatible. The original HCN values of population are missing from 1900 through 1940. While grid-

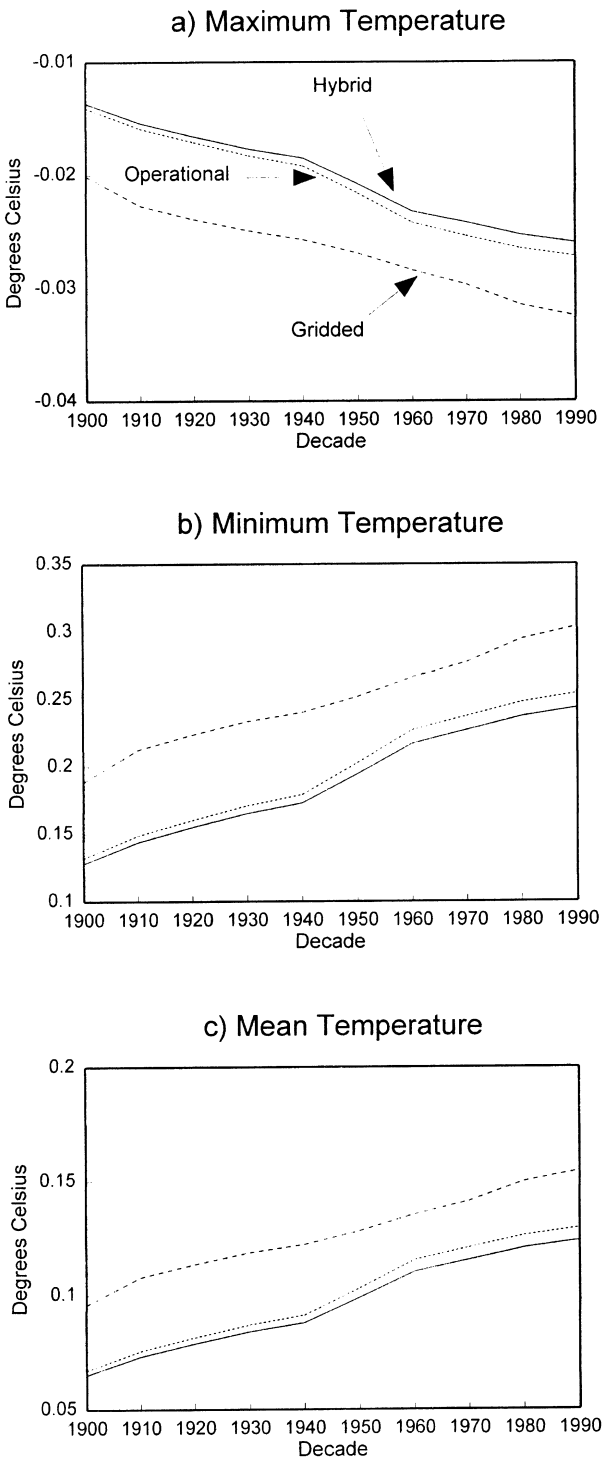

FIG. 1. Urban temperature adjustments for (a) maximum, (b) minimum, and (c) mean temperatures based on operational HCN, hybrid, and gridded population data input.

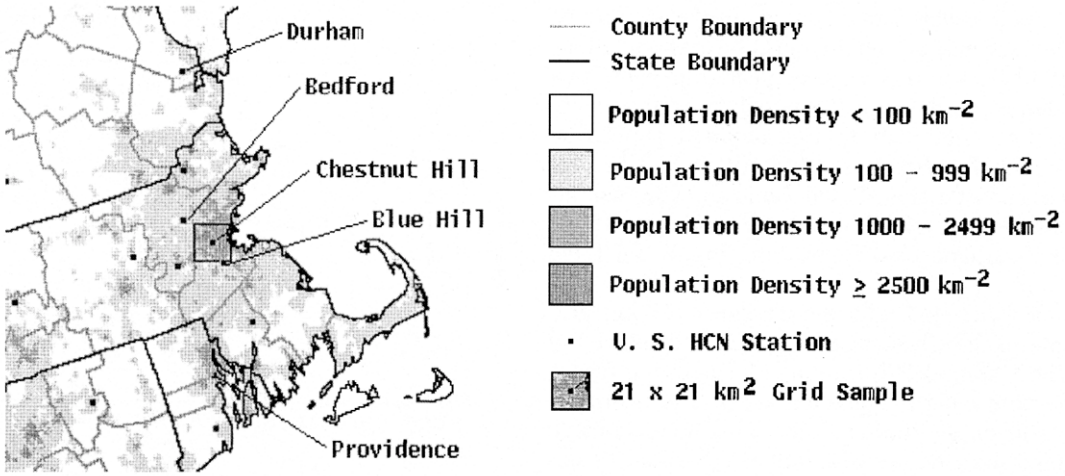

FIG. 2. The HCN stations within the Massachusetts, Rhode Island, New Hampshire region overlaid on population density within the region. The 21 km × 21 km grid cell for the Chestnut Hill, Massachusetts, station is identified.

based values are available for these years the value for 1950 is about six times as great as the HCN value. Thus, use of grid-based values (e.g., the use of gridded values for Durham during 1900–40) to replace only the individual missing values that may exist in the HCN record of population values is not recommended. However, the use of gridded values to selectively replace missing values for those HCN stations without *any* population values may be appropriate for researchers who require population estimation for all stations in the HCN network. For this reason, the hybrid dataset, along with the gridded and original HCN population datasets, are planned to be made available on the World Wide Web in the near future.[1]

The Providence, Rhode Island, and Durham, New Hampshire, examples appear to be representative of the differences in the HCN and grid-based populations for many of the HCN stations. Generally, the current HCN population for stations associated with large urban cities (e.g., Providence) will be greater than the grid-based population. In contrast, the HCN population for medium urban cities (e.g., Durham, New Hampshire, or College Park, Maryland) will generally be less than the grid-based population.

4. Summary

Missing population values represent over 17% of the decennial population values available for the original HCN stations. Additionally, the HCN populations can vary as definitions of the area (boundaries) of a city change. Compared to the original HCN population values the grid-based populations for HCN stations, that are based on samples of gridded population distribution and historical county-based population data, present decennial values derived from a consistent methodology. The operational HCN, hybrid, and grid-based population values are planned to be placed on the National Climatic Data Center's World Wide Web location. The recommended urban-related temperature adjustments

[1] Interested users are encouraged to visit the U.S. HCN Web site at http://www.ncdc.noaa.gov/ol/climate/research/ushcn/ushcn.html.

TABLE 4. Grid-based populations for Bedford, Blue Hill, and Chestnut Hill, Massachusetts, where original and operational HCN-based population values were missing.

Year	Bedford	Blue Hill	Chestnut Hill
1900	94 952	267 374	757 386
1910	112 445	322 687	904 512
1920	130 646	371 169	1 038 315
1930	156 927	425 376	1 140 708
1940	163 048	433 283	1 141 688
1950	178 688	476 901	1 211 516
1960	207 923	503 910	1 190 926
1970	234 531	534 585	1 206 125
1980	229 457	508 961	1 119 975
1990	234 733	518 189	1 144 097

TABLE 5. Original HCN and grid-based populations for Providence, Rhode Island, and Durham, New Hampshire.

	Providence		Durham	
Year	HCN	Grid-based	HCN	Grid-based
1900	175 597	142 914	na*	21 298
1910	224 326	182 483	na	21 175
1920	237 595	203 035	na	21 008
1930	252 981	236 962	na	21 082
1940	253 504	246 632	na	23 731
1950	737 203	273 228	4172	28 089
1960	777 597	303 893	4688	32 971
1970	855 495	336 760	7221	39 392
1980	865 771	344 697	8448	48 415
1990	916 270	360 226	9236	59 400

* Here na indicates that the original HCN population values were not available (missing).

based on the operational HCN (or hybrid dataset) and grid-based methods of estimating station population can be significantly different for individual stations.

Acknowledgments. The authors wish to thank Dr. Hendrik Meij of SEDAC and Paul Sutton of the University of California, Santa Barbara, for their assistance in acquiring the gridded 1990 population density data, and David Bowman and Jay Lawrimore of the National Climatic Data Center for their assistance with documentation of the HCN population data set. This study was partially supported by the NOAA Office of Global Programs and NASA.

REFERENCES

Easterling, D. R., T. R. Karl, E. H. Mason, P. Y. Hughes, and D. P. Bowman, 1996: United States Historical Climatology Network (U.S. HCN) monthly temperature and precipitation data. ORNL/CDIAC-87. NDP-019/R3, Carbon Dioxide Information Analysis Center. Oak Ridge National Laboratory. [Available from National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161.]

Gallo, K. P., D. R. Easterling, and T. C. Peterson, 1996: The influence of land use/land cover on climatological values of the diurnal temperature range. *J. Climate,* **9,** 2941–2944.

——, T. W. Owen, D. R. Easterling, and P. F. Jamason, 1999: Temperature trends of the U.S. Historical Climatology Network based on satellite-designated land use/land cover. *J. Climate,* **12,** 1344–1348.

Karl, T. R., H. F. Diaz, and G. Kukla, 1988: Urbanization: Its detection and effect in the United States climate record. *J. Climate,* **1,** 1099–1123.

SEDAC, 1996: Archive of Census Related Products. Center for International Earth Sciences Information Network (CIESIN)/SEDAC, Palisades, New York. [Available online at http://sedac.ciesin.org/plue/cenguide.html.]

U.S. Bureau of Census, 1996: *Population of States and Counties of the United States: 1790–1990.* U.S. Bureau of Census, 226 pp. [Available from National Technical Information Service, 5285 Port Royal Road, Springfield, VA, 22161.]

——, 1997a: MA-98-5 Metropolitan area and central city population estimates for July 1, 1988 and revised April 1, 1990 census population counts. [Available online at http://www.census.gov/population/estimates/metro-city/ma98-05.txt.]

——, 1997b: MA-96-9 New England County metropolitan area and central city population estimates for July 1, 1988 and revised April 1, 1990 census population counts. [Available online at http://www.census.gov/population/estimates/metro-city/ma98-09.txt.]