

## Reference Station Networks for Monitoring Climatic Change in the Conterminous United States

RUSSELL S. VOSE

*National Climatic Data Center, Asheville, North Carolina*

(Manuscript received 7 June 2004, in final form 14 February 2005)

### ABSTRACT

Set cover models are used to develop two reference station networks that can serve as near-term substitutes (as well as long-term backups) for the recently established Climate Reference Network (CRN) in the United States. The first network contains 135 stations distributed in a relatively uniform fashion in order to match the recommended spatial density for CRN. The second network contains 157 well-distributed stations that are generally not in urban areas in order to minimize the impact of future changes in land use. Both networks accurately reproduce the historical temperature and precipitation variations of the twentieth century.

### 1. Introduction

According to the National Research Council (1999), our ability to monitor the climate system is degrading significantly at the same time that many countries are making major policy decisions on climate change. In response, some nations are establishing new observing systems specifically designed and managed for climate monitoring purposes [e.g., the U.S. Climate Reference Network (CRN); Heim (2001)]. In other cases, “reference station networks” consisting of high quality, spatially representative stations from existing networks have been proposed to meet climate monitoring requirements (Collins et al. 1999). New monitoring networks such as CRN have a number of scientific and technical advantages, including the ability to choose pristine station locations, design precision instrumentation suites, and standardize climate observing practices. Unlike reference station networks, however, new observing systems can be extremely expensive to establish, and they may require many years to reach full deployment (CRN, for instance, will not be fully deployed until at least the end of 2007). Furthermore, they cannot provide a historical perspective on climate change until data have been collected for a sufficient period of time.

Given these limitations, the purpose of this paper is to identify two potential reference station networks that could serve as near-term substitutes (as well as long-term backups) for the recently established CRN in the United States. The U.S. Historical Climatology Network (HCN; Easterling et al. 1996) is used as the pool of candidate sites for these reference station networks. The first network is identified by systematically eliminating stations from HCN using a facility location model that minimizes the total number of stations, ensures a fairly uniform coverage of the country, and approximates the planned spatial density of CRN. The second network is identified using a more elaborate location model that also maximizes the number of stations in rural areas. Because the primary intent of CRN is to monitor future climatic change at the national scale, an assessment is made of each network’s ability to capture historical temperature and precipitation trends in the conterminous United States during the period 1911–2000.

### 2. HCN data

HCN contains mean monthly temperature time series for 1221 stations that are distributed in a relatively uniform fashion across the conterminous United States. The network is commonly used to quantify climate change because its temperature records contain adjustments that explicitly account for documented changes in observation time (Karl et al. 1986), station location,

---

*Corresponding author address:* Russell S. Vose, National Climatic Data Center, 151 Patton Ave., Asheville, NC 28801.  
E-mail: Russell.Vose@noaa.gov



FIG. 1. Locations of the 601 HCN stations used in the study.

and instrumentation (Karl and Williams 1987). HCN is a subset of the much larger U.S. Cooperative Observing Network, stations having been chosen based upon their spatial coverage, record length, data completeness, and historical stability (i.e., the number of changes in location, instrumentation, and observing practice). In short, HCN consists of among the highest quality stations in the U.S. historical climate record.

Detailed land use/land cover metadata are available for most HCN stations through a survey conducted in 1990 by the National Climatic Data Center (Gallo et al. 1996). In this survey the observers at each station were asked to classify the land use–land cover types within 100-, 1000-, and 10 000-m radii of the station. If a station's survey indicated that a city with 10 000 or more people was within any of these three radii in 1990, then the station was excluded from this investigation because of the somewhat greater potential for future urbanization. This eliminated 620 HCN stations from consideration for the reference station networks, resulting in the 601-station subset depicted in Fig. 1. In general, this station subnetwork is relatively dense in most of the country, though gaps exist in states such as California, Nevada, and Texas.

In theory, the 601-station subset could itself be a reference station network. For several reasons, however, a smaller network is more compatible with the requirements of CRN. For instance, station history information is necessary to make adjustments for changes in station location, instrumentation, and observing practice, yet it is difficult to acquire such information in near–real time for most stations in the Cooperative Observing Network. Likewise, it would be cumbersome to conduct regular land use/land cover surveys for a national network of 601 stations (in fact, the HCN itself has been surveyed just once). Finally, only about 135 stations are needed to meet the climate monitoring goals of CRN (Vose and Menne 2004).

### 3. Development of reference networks

As with other atmospheric observing systems, reference station networks are generally designed to meet specified performance requirements (Peterson et al. 1997). Two potential CRN reference station networks, and the performance requirements upon which they are based, are described here. The first network targets baseline CRN needs; in particular, the stations within it are distributed to approximate the planned spatial density (and, thus, the climate monitoring capabilities) of CRN, and the number of stations is minimized to simplify the collection of station metadata and land use/land cover information. The second reference station network goes one step further; in addition to meeting the baseline CRN requirements, the number of stations in rural areas is maximized in an attempt to limit the potential impact of future urbanization.

Both reference station networks are developed using a locational modeling technique known as set covering (Current et al. 2002). Over the past several decades, location models have been employed to determine the optimum siting for a wide variety of “facilities,” such as fire stations (Schilling 1980), brewery depots (Gelders et al. 1987), and rain gauges (Hogan 1990). The general objective of the set cover location model is to site the minimum number of facilities required to “cover” a given set of locations within an area. A location is said to be “covered” if it is within a specified distance of a facility. In this investigation, the specific objective of the set cover location model is to identify the smallest possible reference station network that can cover all of the HCN stations in Fig. 1. An HCN station is said to be covered if it is within a specified distance of any station in the reference station network. For an in-depth discussion of the set cover model (and location models in general), see Daskin (1995).

The length of the “cover” distance has a major impact on the resulting minimum-size network (i.e., larger distances result in networks with fewer stations), and thus the distance threshold is generally chosen to ensure that the final network has a desired characteristic. Because the average great-circle distance between CRN stations is planned to be approximately  $2.5^\circ$  (Vose and Menne 2004), a cover distance of half that size is employed here to develop the reference station networks. In essence, a distance threshold of  $1.25^\circ$  implies that any given station can cover all other locations that fall within an imaginary circle that is  $2.5^\circ$  in diameter and that is centered over the station. Assuming that the original network is sufficiently dense (which is mostly true for the HCN station subset in Fig. 1), a set cover model using this cover distance should produce a

minimum-size network consisting of stations that are spaced at about  $2.5^\circ$  intervals.

A separate set cover model is used to develop each reference station network. Formally, the model for the first network (termed the “baseline” network) can be stated as

$$\min \sum_{i=1}^{601} x_i,$$

where  $x_i$  is assigned a value of 1 if the station is included in the reference station network and 0 if the station is not included. For each station that is not included (i.e., when  $x_i = 0$ ),

$$d_i \leq 1.25^\circ,$$

where  $d_i$  is the distance from station  $i$  to the nearest station that was included in the reference station network. In other words, the goal of the model is to minimize the number of stations, subject to the constraint that each of the original 601 stations must be within  $1.25^\circ$  of a station in the final network.

The set cover model for the second network (termed the “rural” network) is designed to maximize the number of stations in rural areas while minimizing the total number of stations. This is accomplished by assigning each station a “cost” that represents the land use/land cover characteristics around the station (the more rural the environment, the lower the cost, and the greater the likelihood the station is selected for the network). Formally, the model can be stated as

$$\min \sum_{i=1}^{601} c_i x_i,$$

where  $c_i$  is the cost of including station  $i$  in the reference station network. In other words, the goal is to minimize the overall “land use” cost, again subject to the  $1.25^\circ$  cover distance constraint.

The cost for each station is derived from land use/land cover information collected in the HCN metadata survey:

$$c_i = (3 \times l_{100}) + (2 \times l_{1000}) + (1 \times l_{10000}) + 1,$$

where  $l_{100}$ ,  $l_{1000}$ , and  $l_{10000}$  are variables that represent the land use/land cover characteristics within 100, 1000, and 10 000 m radii of the station, respectively. The largest contribution to the overall cost is from  $l_{100}$  and the smallest is from  $l_{10000}$  because land use conditions near a station generally have the greatest impact on observed temperatures (Peterson 2003). Each land use/land cover variable can take on one of three values. If there is a “town” (population between 1000 and 10 000) within the given radius, the land use/land cover variable

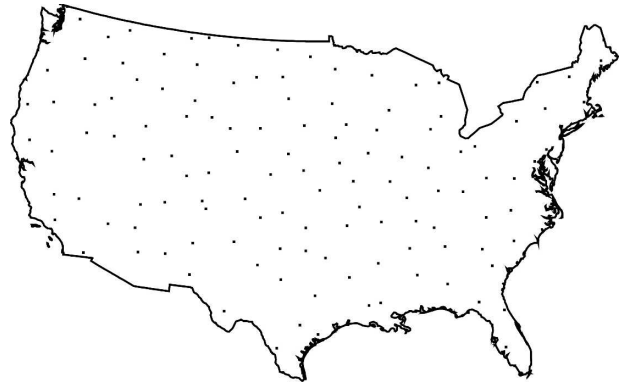


FIG. 2. Locations of the 135 stations in the baseline network.

is assigned a value of 2. If there is a “small town” (population less than 1000) at that radius, the variable is assigned a value of 1. If conditions are “rural” (nonvegetated, coastal, island, forest, farmland, savannah, prairie, or tundra), the variable is assigned a value of 0. In other words, the more rural the environment, the lesser is the cost. For instance, a station surrounded by prairie at all radii would have a cost of 1, whereas a station surrounded by a town at all radii would have a cost of 13.

In theory, each set cover model could be solved by simulating all possible combinations of stations (e.g., all 100-station networks, all 101-station networks) and selecting the smallest solution that meets the  $1.25^\circ$  cover distance constraint. Unfortunately, this approach is not practical here because there is an extremely large number of combinations to enumerate. Consequently, a “very good” solution to each model was obtained with a technique known as linear programming, which is commonly employed in engineering and business for optimization problems (Kuby et al. 1997). For a detailed introduction to linear programming, see Vanderbei (1996). In this investigation, the revised Simplex algorithm (Dantzig 1951) was iterated with a branch and bound algorithm until an optimal solution (i.e., reference station network) was found for each set cover model.

Figure 2 depicts the locations of stations in the baseline network. In general, the distribution of stations is relatively uniform across most of the country, though parts of a few large western states contain no stations. The uniform station distribution is a consequence of the distance-based thinning effect of the set cover model that was employed to define the network; the gaps in the network are a function of gaps in the original pool of candidate sites. In total, the network consists of 135 stations, which is the same number recommended for CRN by Vose and Menne (2004). This result is coinci-

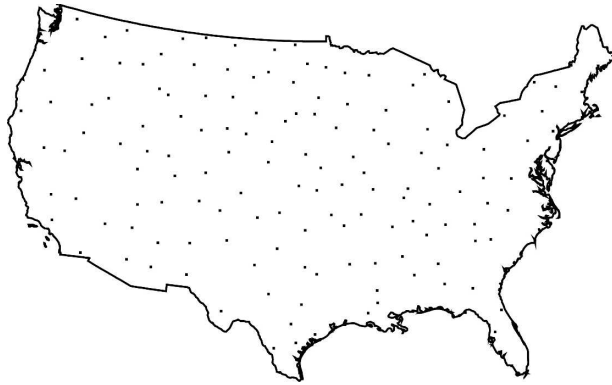


FIG. 3. Locations of the 157 stations in the rural network.

dental, but in general one would expect the totals to be similar because the set cover model for the baseline network was designed to replicate the ideal spatial density for CRN.

Figure 3 depicts the locations of stations in the rural network. Once again, the distribution of stations is relatively uniform except for the aforementioned gaps in the West. In total, the network consists of 157 stations, 22 more than in its baseline counterpart. The increased size results from the addition of the “cost” variable to the second set cover model, which minimized a weighted sum of stations rather than only the total number of stations. As intended, the inclusion of land use/land cover information in the model significantly increased the proportion of rural stations in the resulting network. For instance, only 3% of the stations have a town within a 100-m radius, whereas 24% of the stations in the baseline network have a town within that distance. Over 80% of the stations in the rural network have predominantly rural conditions within a 1000-m radius (versus 44% for the baseline network). Finally, more than 50% of the stations in the rural network are still surrounded primarily by rural environments at a radius of 10 000 m (versus only 24% of the baseline network).

In general, both networks match their original design specifications (e.g., a uniform distribution of stations, a large proportion of rural stations). However, both networks also have several practical limitations. First, they are based on land use/land cover data collected more than a decade ago. It is entirely possible that environmental conditions around some stations have changed, and thus a new survey of stations should be conducted periodically to assess recent development. Second, stations in the networks will not be collocated with CRN stations (the latter will be arranged in a quasi-uniform fashion, and there is not usually an HCN station that coincides with each grid node). Consequently, the data

from the reference station networks can only serve as a complement to CRN on the national scale; individual stations in the networks may not be able to act as “backups” for individual CRN stations. Finally, station history information is required to make adjustments for changes in location, instrumentation, and observing practice. As a result, it will necessary to update this information with greater regularity than is currently the practice for HCN if the networks are to be used for real-time climate monitoring.

#### 4. Evaluation of reference networks

The primary goal of CRN is to monitor future climate change at the national scale. Consequently, the ability of each reference station network to capture historical U.S. temperature and precipitation trends is evaluated here. The assessment is accomplished by computing an annual U.S. temperature and precipitation time series for the period 1911–2000 for both the baseline network and the rural network, then comparing each time series to the comparable series computed using the full HCN. In the case of temperature, the analysis is based on HCN data that have been adjusted for changes in observation time, station location, and instrumentation (but not urbanization). The first step in calculating each U.S. time series involved converting the annual value for each variable at each station to an anomaly from its mean during the period 1961–90. For temperature, anomalies were computed by subtracting the 30-yr mean from the annual average; for precipitation, anomalies were computed by dividing the 30-yr mean into the annual total (i.e., by calculating the “percent of normal”). The annual anomalies were then interpolated to the nodes of a  $0.25^\circ \times 0.25^\circ$  latitude–longitude grid, and the grid points were area weighted into a mean anomaly for the conterminous United States for each year. Interpolation was performed using the inverse distance weighting model of Willmott et al. (1985), which accounts for the sphericity of the earth when computing the distance between each station and grid point and permits extrapolation beyond the range of data values in the neighboring stations. Finally, the time series for each reference station network was compared with the full HCN series over rolling 30-yr periods (e.g., 1911–40, 1912–41, and so on up through 1971–2000). Three statistics were used to describe the similarity between the time series: the coefficient of determination ( $r^2$ ), the mean absolute difference, and the difference in decadal trend.

The results of this analysis indicate that both reference station networks replicate the full HCN with a high degree of fidelity. For example, the lowest  $r^2$  pro-



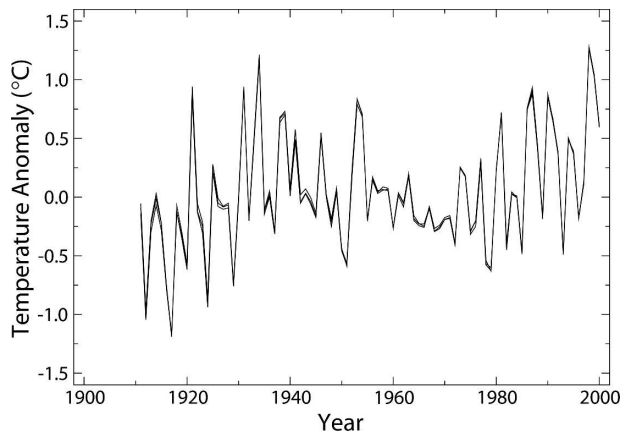


FIG. 4. Annual temperature anomalies in the conterminous United States from 1911 to 2000. The three networks (the full HCN, the baseline network, and the rural network) are all plotted as thin black lines. They are nearly indistinguishable on this scale.

duced by either reference station network for any 30-yr period was 99% for temperature and 95% for precipitation. Similarly, the largest mean absolute difference for either reference station network for any 30-yr period was  $0.030^{\circ}\text{C}$  for temperature and 3% for precipitation. Both networks were also effective in quantifying climatic change; each on average had a temperature trend that was within  $0.005^{\circ}\text{C decade}^{-1}$  of the full HCN and a precipitation trend that was within  $0.25\%$   $\text{decade}^{-1}$  of the full HCN. Furthermore, the “worst” temperature and precipitation trends for either network were within  $0.025^{\circ}\text{C decade}^{-1}$  and  $0.85\%$   $\text{decade}^{-1}$  of the full HCN, respectively. Not surprisingly, the rural network usually had a slightly lower temperature trend than the full HCN (by about  $0.005^{\circ}\text{C decade}^{-1}$ ). However, as depicted in Fig. 4, the high degree of similarity between the three temperature time series (the full HCN, the baseline network, and the rural network) suggests that land use change may have a relatively minor impact on twentieth century temperature trends (or that HCN in general is not significantly impacted by such changes).

## 5. Summary and conclusions

This paper described two reference station networks for monitoring climate change in the United States. The networks were developed by applying set cover models to a subset of stations in HCN. The baseline network, which was designed to approximate the spatial density of CRN, contained 135 stations that were distributed in a relatively uniform fashion across the country. The rural network, which was designed to minimize the potential impact of future urbanization, contained 157 sta-

tions that were likewise well distributed throughout the country. Both networks were capable of accurately reproducing U.S. temperature and precipitation trends over the period 1911–2000 as depicted by the full HCN. Consequently, the networks should be useful in the detection of future climate change in the United States, and they could readily serve as a complement to CRN.

*Acknowledgments.* Many thanks to Randy Cerveny, Dave Easterling, Mike Kuby, and the anonymous reviewers, whose comments and suggestions substantially improved this manuscript. Partial support for this work was provided by the Office of Biological and Environmental Research, U.S. Department of Energy, and the NOAA Office of Global Programs, Climate Change Data and Detection Element.

## REFERENCES

- Collins, D. A., S. Johnson, N. Plummer, A. K. Brewster, and Y. Kuleshov, 1999: Re-visiting Tasmania’s reference climate stations with a semi-objective network selection scheme. *Aust. Meteor. Mag.*, **48**, 111–122.
- Current, J., M. Daskin, and D. Schilling, 2002: Discrete network location models. *Facility Location: Applications and Theory*, Z. Drezner and H. Hamacher, Eds., Springer-Verlag, 81–118.
- Dantzig, G., 1951: Maximization of a linear function of variables subject to linear inequalities. *Activity Analysis of Production and Allocation*, T. C. Koopmans, Ed., Wiley, 339–347.
- Daskin, M. S., 1995: *Network and Discrete Location: Models, Algorithms and Applications*. Wiley, 520 pp.
- Easterling, D. R., T. R. Karl, E. H. Mason, P. Y. Hughes, D. P. Bowman, R. C. Daniels, and T. A. Boden, 1996: United States Historical Climatology Network (U.S. HCN) monthly temperature and precipitation data. Carbon Dioxide Information and Analysis Center Publ. 4500, Oak Ridge National Laboratory, Oak Ridge, TN, 83 pp.
- 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.
- Gelders, L. F., L. M. Pintelon, and L. N. vanWassenhove, 1987: A location-allocation problem in a large Belgian brewery. *Euro. J. Operation Res.*, **28**, 196–206.
- Heim, R. R., 2001: New network to monitor climate change. *Eos, Trans. Amer. Geophys Union*, **82**, 143.
- Hogan, K., 1990: Reducing errors in rainfall estimates through rain gauge location. *Geogr. Anal.*, **4**, 258–266.
- Karl, T. R., and C. Williams, 1987: An approach to adjusting climatological time series for discontinuous inhomogeneities. *J. Climate Appl. Meteor.*, **26**, 1744–1763.
- , C. N. Williams Jr., P. J. Young, and W. M. Wendland, 1986: A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States. *J. Climate Appl. Meteor.*, **25**, 145–160.
- Kuby, M. J., R. S. Cerveny, and R. I. Dorn, 1997: A new approach to paleoclimatic research using linear programming. *Palaeo*, **129**, 251–267.
- National Research Council, 1999: *Adequacy of Climate Observing Systems*. National Academy Press, 51 pp.

- Peterson, T. C., 2003: Assessment of urban versus rural in situ surface temperatures in the contiguous United States: No difference found. *J. Climate*, **16**, 2941–2959.
- , H. Daan, and P. D. Jones, 1997: Initial selection of a GCOS surface network. *Bull. Amer. Meteor. Soc.*, **78**, 2145–2152.
- Schilling, D. A., 1980: Dynamic location modeling for public sector facilities: A multi-criteria approach. *Decision Sci.*, **11**, 714–724.
- Vanderbei, R. J., 2001: *Linear Programming: Foundations and Extensions*. International Series in Operations Research and Management Science, Vol. 37, Springer, 472 pp.
- Vose, R. S., and M. J. Menne, 2004: A method to determine station density requirements for climate observing networks. *J. Climate*, **17**, 2961–2971.
- Willmott, C. J., C. M. Rowe, and W. D. Philpot, 1985: Small-scale climate maps: A sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. *Amer. Cartogr.*, **12**, 5–16.