1999

NOTES AND CORRESPONDENCE: The Effects of Data Gaps on the Calculated Monthly Mean Maximum and Minimum Temperatures in the Continental United States: A Spatial and Temporal Study

David E. Stooksbury

Craig D. Idso

Kenneth G. Hubbard

NOTES AND CORRESPONDENCE

The Effects of Data Gaps on the Calculated Monthly Mean Maximum and Minimum Temperatures in the Continental United States: A Spatial and Temporal Study

DAVID E. STOOKSBURY, CRAIG D. IDSO, AND KENNETH G. HUBBARD

High Plains Climate Center, School of Natural Resource Sciences, University of Nebraska, Lincoln, Nebraska

1 December 1997 and 5 June 1998

ABSTRACT

Gaps in otherwise regularly scheduled observations are often referred to as missing data. This paper explores the spatial and temporal impacts that data gaps in the recorded daily maximum and minimum temperatures have on the calculated monthly mean maximum and minimum temperatures. For this analysis 138 climate stations from the United States Historical Climatology Network Daily Temperature and Precipitation Data set were selected. The selected stations had no missing maximum or minimum temperature values during the period 1951–80. The monthly mean maximum and minimum temperatures were calculated for each station for each month. For each month 1–10 consecutive days of data from each station were randomly removed. This was performed 30 times for each simulated gap period. The spatial and temporal impact of the 1–10-day data gaps were compared. The influence of data gaps is most pronounced in the continental regions during the winter and least pronounced in the southeast during the summer. In the north central plains, 10-day data gaps during January produce a standard deviation value greater than 2°C about the ''true'' mean. In the southeast, 10-day data gaps in July produce a standard deviation value less than 0.5°C about the mean. The results of this study will be of value in climate variability and climate trend research as well as climate assessment and impact studies.

1. Introduction

In this paper we discuss the impacts that data gaps have on the calculated mean maximum and minimum monthly temperatures. Missing data is not an appropriate term to describe the situation arising when no data are taken for one or more observation intervals. Missing data implies that the data were first taken and then misplaced or lost. This term is used so widely that we often hear such phrases as ''replacing the missing data with . . . .'' We prefer the term data gaps for data that was never collected. Data gaps are a common problem that plagues climate research. As an example, *Climatological Data*, an official publication of the National Climatic Data Center (Asheville, NC), includes monthly values for stations when daily values are not reported on as many as 9 days (NOAA 1996). Our results will serve as a guide for researchers in determining the potential impact of noncontinuous temperature data in their research.

Researchers have developed many methods to lessen the impact of incomplete data. A common approach is to systematically produce an estimate of the observed value. Some common estimation methods are substitution of the nearest neighbor value, using a regression-derived value, using a kriging method–derived value, or using the mean value. All data estimation methods have advantages and disadvantages. The nearest neighbor method is simple but prone to errors due to microclimate responses. The mean value method is also simple but misses extreme events and tends to artificially reduce variation about the mean. Guttman (1991) has

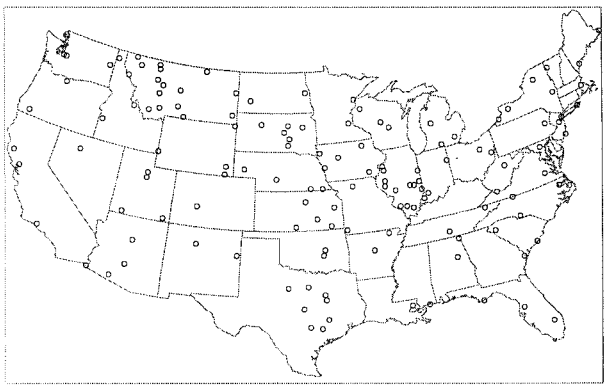

FIG. 1. The 138 climate stations used in this study.

Corresponding author address: Dr. David E. Stooksbury, Georgia State Climate Center, Dept. of Biological and Agricultural Engineering, Driftmier Engineering Center, University of Georgia, Athens, GA 30602-4435.
E-mail: stooks@bae.uga.edu

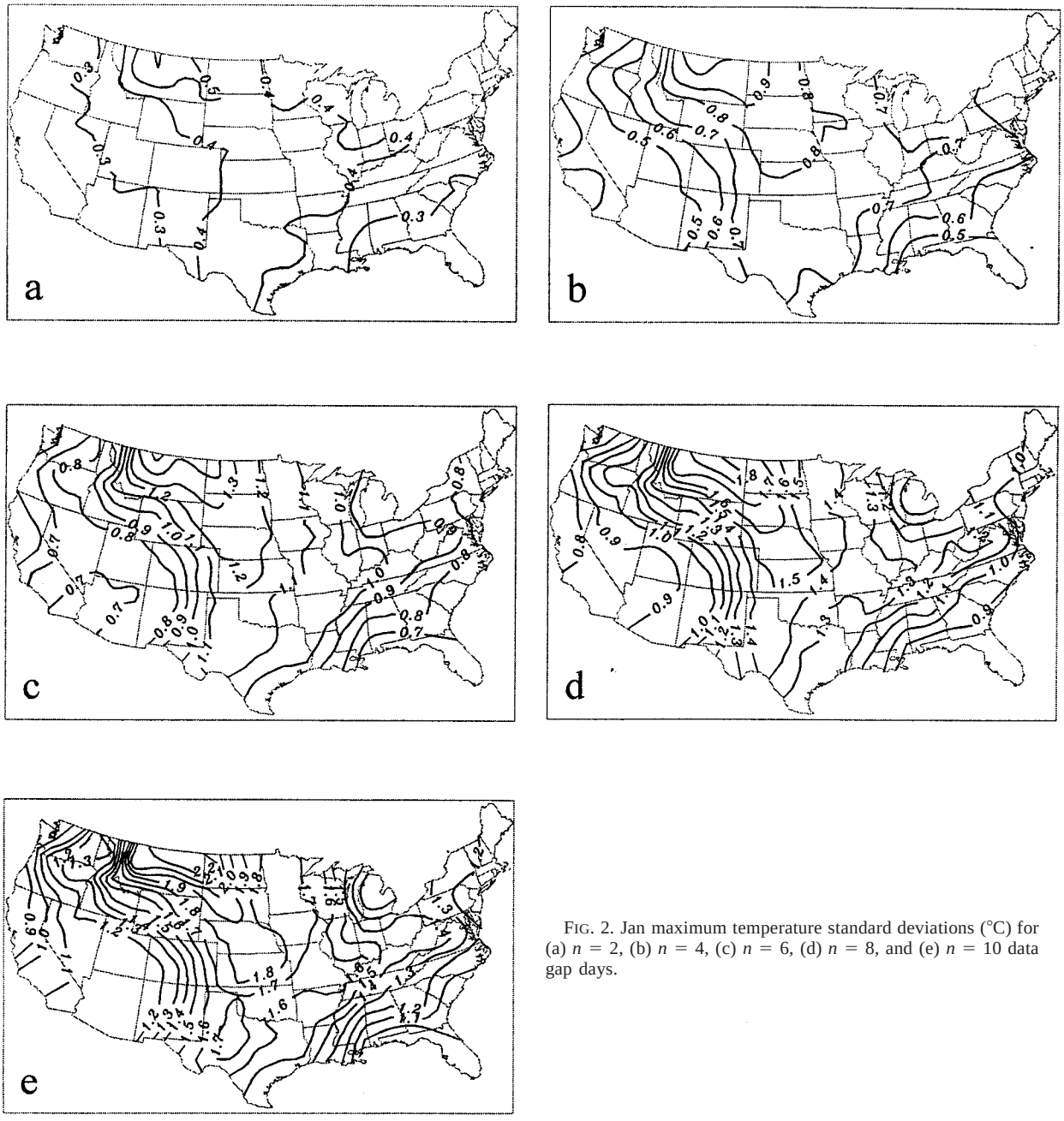

FIG. 2. Jan maximum temperature standard deviations (°C) for (a) $n = 2$, (b) $n = 4$, (c) $n = 6$, (d) $n = 8$, and (e) $n = 10$ data gap days.

shown that even with a record of 112 yr, daily means are not a smooth monotonic curve but contain many ''peaks and valleys.'' Thus, replacing data gaps with the daily means may have other unknown impacts on the research results. The regression method has some robustness with respect to microclimate effects and extreme events, but it is more complicated to perform. Another common method used to deal with data gaps is to ignore them. This method assumes that a data gap is a random event. In some instances researchers simply cannot ignore gaps in data because the analysis they wish to perform requires continuous data.

In this paper, as an example of the seriousness of data gaps, we investigate the effect on the calculated monthly mean maximum and minimum temperatures in the continental United States. We will show that errors in the mean temperature due to the influence of data gaps have temporal and spatial patterns. While the scope of this

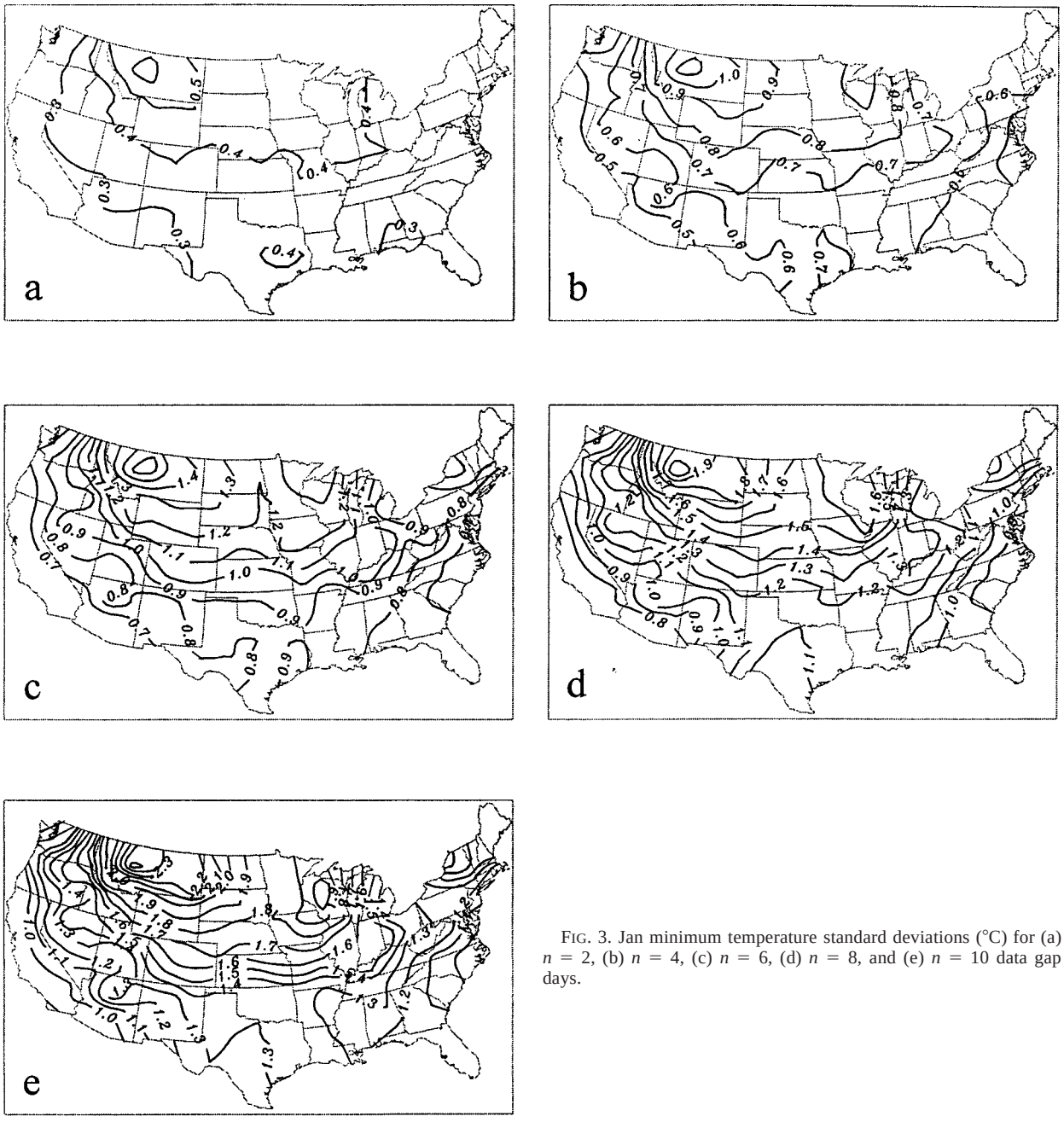

FIG. 3. Jan minimum temperature standard deviations (°C) for (a) $n = 2$, (b) $n = 4$, (c) $n = 6$, (d) $n = 8$, and (e) $n = 10$ data gap days.

paper deals with the seriousness of data gaps, we do not address the effect of data estimation techniques.

Our principal result is that the calculated monthly average maximum and minimum temperatures can differ by more than ±1°C with a 3-day gap in the data. The severity of the error, the size of the standard deviation, is a function of the time of year and location. The largest error occurs in the winter and at interior continental locations.

2. Data

The data used in this study are from the United States Historical Climatology Network Daily Temperature and Precipitation Data (HCN/D) set (Hughes et al. 1992). The HCN/D dataset is available from the Carbon Dioxide Information Analysis Center at Oak Ridge National Laboratory (Oak Ridge, TN). The time period selected for analysis was 1951–80. This

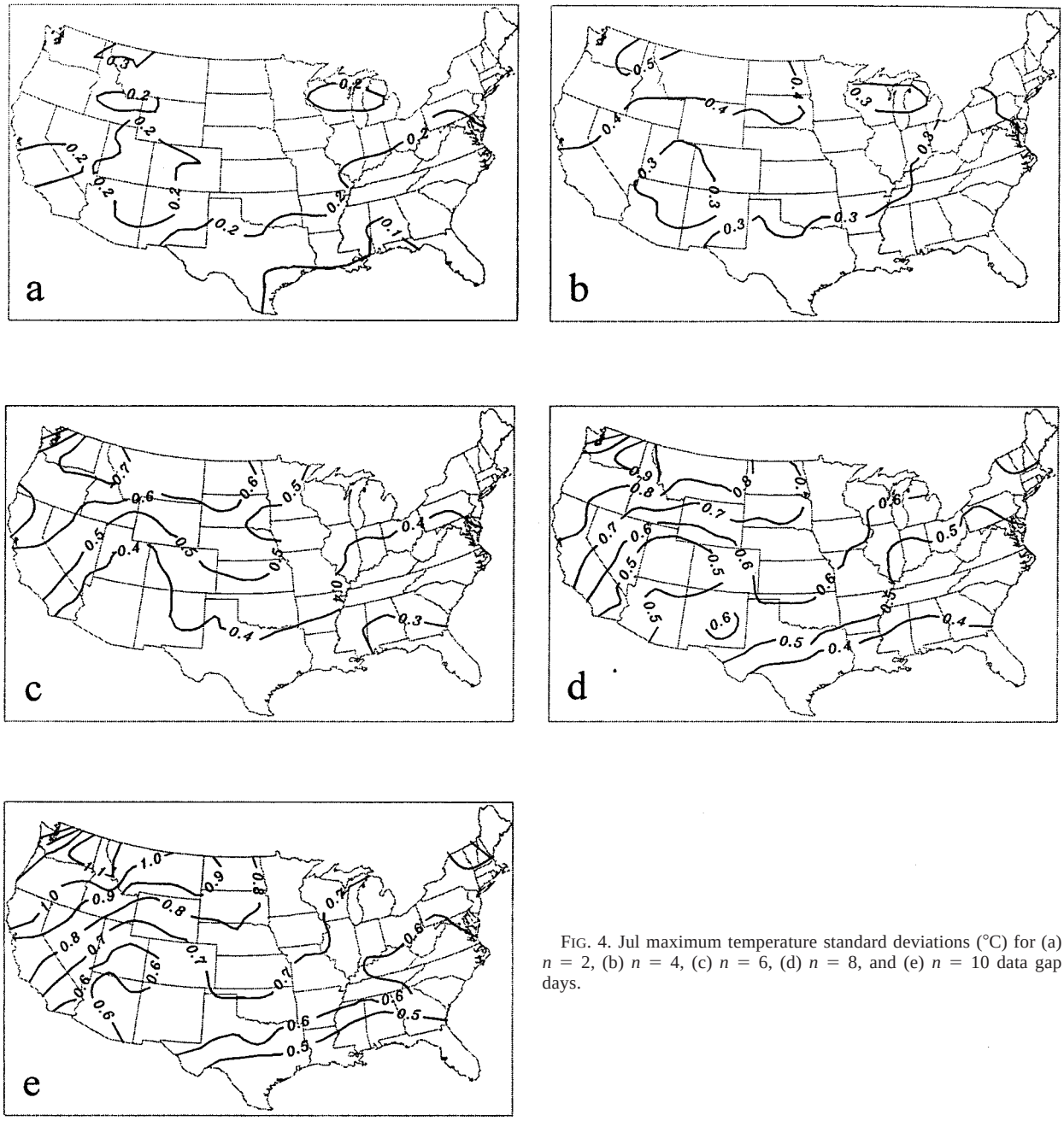

FIG. 4. Jul maximum temperature standard deviations (°C) for (a) $n = 2$, (b) $n = 4$, (c) $n = 6$, (d) $n = 8$, and (e) $n = 10$ data gap days.

analysis requires the use of stations that contain no gaps in data. Because of this restriction, we were limited to 138 stations. The station locations are plotted in Fig. 1.

3. Analysis

We used observed daily maximum and minimum temperature values (1951–80) to calculate the mean maximum and mean minimum temperatures for each day of the year. For each station, we calculated the monthly mean maximum temperatures and the monthly mean minimum temperatures using the daily mean maximum and minimum values. The mean temperatures calculated with no gaps in data will be referred to as the ''true'' means.

To determine the effects of ignoring data gaps, we randomly selected a day on which data gaps of 1–10 consecutive days would be assumed, even though all observations were actually available. Consecutive days are a better representation of the data gaps that occur

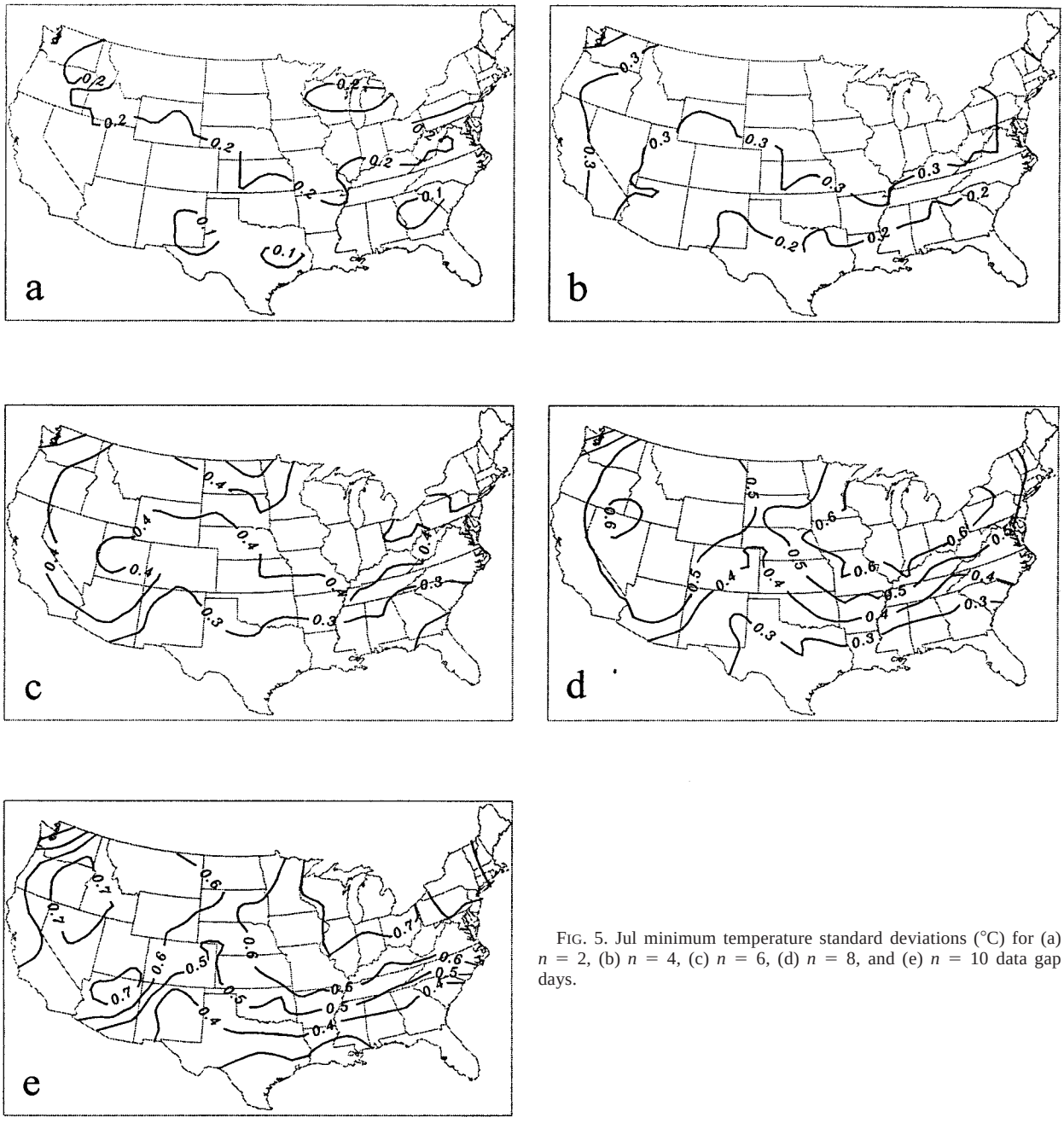

FIG. 5. Jul minimum temperature standard deviations (°C) for (a) $n = 2$, (b) $n = 4$, (c) $n = 6$, (d) $n = 8$, and (e) $n = 10$ data gap days.

due to instrument failure, illness, or absence of the cooperative observer.

A random number generator (Press et al. 1992) was used to determine the starting day for each gap. For each month we repeatedly assigned gaps in data and calculated the associated means. These ''representative'' means, resulting from the simulation of data gaps, were generated 30 times for each month. We then calculated the departures of the representative means from the true means and the standard deviation of the departures. This analysis was repeated for data gaps of 1–10 days in length.

The maps were prepared using the Surfer (Golden Software Inc., Golden, CO) graphic package. The maps are a Lambertian projection. We selected the kriging option to produce contour lines on all maps.

4. Results

a. Spatial distribution—January

The effects of data gaps are most severe in the interior continental regions but there are noticeable effects in

TABLE 1. List of selected stations and Köppen classifications.

Station name, state	Köppen classification	Köppen climate symbol
Olga 2SE, WA	Temperate oceanic	Do
Ukiah, CA	Subtropical dry summer	Cs
Yuma, AZ	Desert	BWh
Bridgeport, NE	Steppe	BSk
Fairbury, NE	Temperate continental	Dca
Cloquet, MN	Temperate continental	Dcb
Albany, TX	Steppe	BSh
Talladega, AL	Subtropical humid	Cf
Belle Glade, FL	Tropical wet–dry	Aw

coastal regions (Figs. 2 and 3). There is a gradient between the central plains and both the Atlantic and Pacific coasts. The Pacific Ocean had a slightly greater modifying influence than the Atlantic Ocean. Figures 2a and 3a show for a 2-day gap that the value for one standard deviation ranges from ±0.3°C in the desert Southwest to ±0.5°C in the high plains. When 10-day gaps were simulated, the range of one standard deviation increases from ±1.1°C in Florida and along the Pacific coast to ±2.3°C in the high plains of Montana (Figs. 2e and 3e).

b. Spatial distribution—July

As with January, the data gap effect is strongest in the interior continental regions (Figs. 4 and 5). However, the effect is much less pronounced in July. In July the observed effect is a general north–south gradient most pronounced in the northern plains and weakest near the Gulf of Mexico region. The Atlantic Ocean had a slightly greater modifying influence than the Pacific Ocean. The standard deviation for a 2-day gap ranges from ±0.1°C in the South to ±0.2°C across the remainder of the nation (Figs. 4a and 5a). The standard deviation for a 10-day gap ranges from ±0.4°C in Florida to ±1.1°C across eastern Washington (Figs. 4e and 5e).

c. Temporal distribution—January

We selected 10 climate stations that represent the Köppen climate classifications (Trewartha and Horn 1980). The 10 stations and their Köppen climate classifications are listed in Table 1.

At all 10 locations, the weakest influence of data gaps was observed during July and/or August. This was true for both the maximum and minimum temperatures (Figs. 6–8). The period of maximum influence of data gaps generally occurs during the winter with a few exceptions noted below.

1) WESTERN UNITED STATES

Of the nine selected stations, Olga 2SE, Washington, had the least amount of variation through time (Fig. 6a). Minimum temperatures at Olga are influenced most by data gaps during the winter and the least during the summer (Fig. 6b). Maximum temperatures at Ukiah, California (Fig. 6b), and Yuma, Arizona (Fig. 6c), were influenced most during April and June with a secondary peak during the fall. The minimum temperature at Ukiah was influenced most during the winter (Fig. 6e). At Yuma the minimum temperature was most influenced in the early fall (Fig. 6f).

2) CENTRAL UNITED STATES

In the interior plains the largest effect of gaps in maximum temperature data is seen in January and in the transitional seasons. Cloquet, Minnesota, has a maximum during January, April, and November (Fig. 7a). This relationship for maximum temperature is also evident in Fairbury, Nebraska (Fig. 7c), and Bridgeport, Nebraska (Fig. 7e).

For minimum temperature the least effect is in July and August (Figs. 7b,d,f). The effect is largest in January.

3) SOUTHERN UNITED STATES

The influence of data gaps on maximum temperature was greatest at all southern stations during the winter (Figs. 8a,c,e). The influence was least during July and August. At Albany, Texas, minimum temperatures were greatly affected by data gaps during the months of April and October (Fig. 8b). At Talladega, Alabama, the greatest influence occurred during February; however, there were secondary peaks during the transitional months of April and October (Fig. 8d). The influence of gaps in minimum temperature data at Belle Glade, Florida, was greatest during the winter and least during the summer (Fig. 8f).

5. Discussion

For regularly scheduled observations, datasets of substantial duration will most likely have gaps in data. The results of this study indicate that gaps in the data can have a large effect on the mean of maximum or minimum temperature. In general, the impact is most pronounced in continental locations during the winter. We found that there is a short-term persistence in daily temperature. Guttman and Plantico (1987) have also found a 1-day persistence in daily temperature. Regions that have daily persistence in temperature are least impacted by data gaps.

The northern plains of the United States can experience large temperature changes over a short time. During January, downslope winds from the Rocky Mountains can lead to rapid temperature rises followed by the invasion of extremely cold air from Canada. Farther south, the frequency of invasion of Canadian air masses rapidly decreases and, thus, winter temperatures have a greater persistence.

Persistence is most striking in coastal regions where

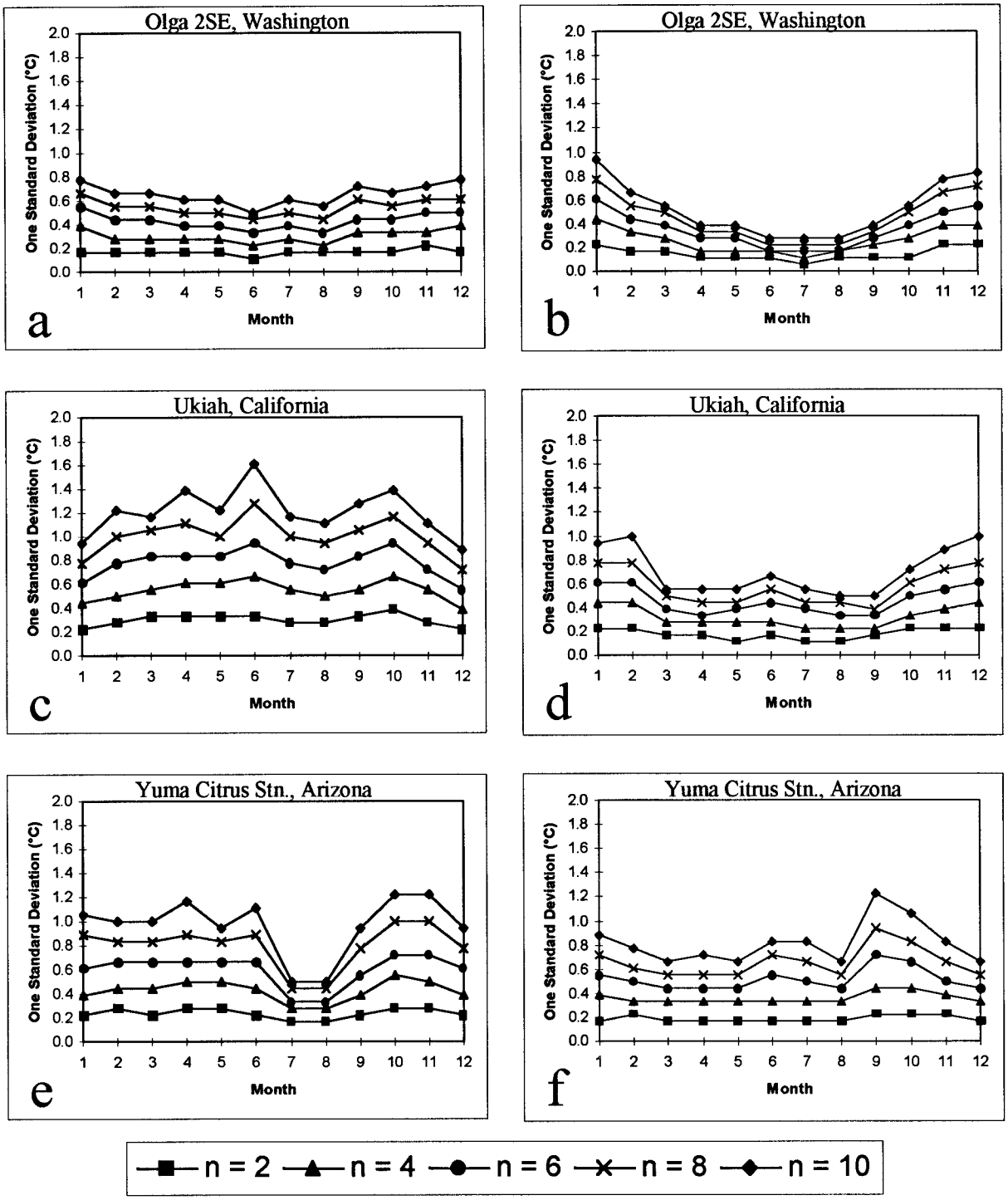

FIG. 6. Standard deviations by month for maximum [(a), (c), and (e)] and minimum [(b), (d), and (f)] temperatures for $n = 2$, 4, 6, 8, and 10 data gap days at Olga, WA; Ukiah, CA; and Yuma, AZ.

the water tends to modify extremes, especially in January. During January, the Pacific Ocean has a greater modifying influence than the Atlantic Ocean. This greater modifying impact of the Pacific Ocean is due to the strong westerly airflow of winter. In the Pacific coastal region the airflow is generally onshore while it is generally offshore in the Atlantic coastal region. Thus, the Pacific coastal region is modified by the onshore flow while the Atlantic coastal region is being modified by a more continental flow.

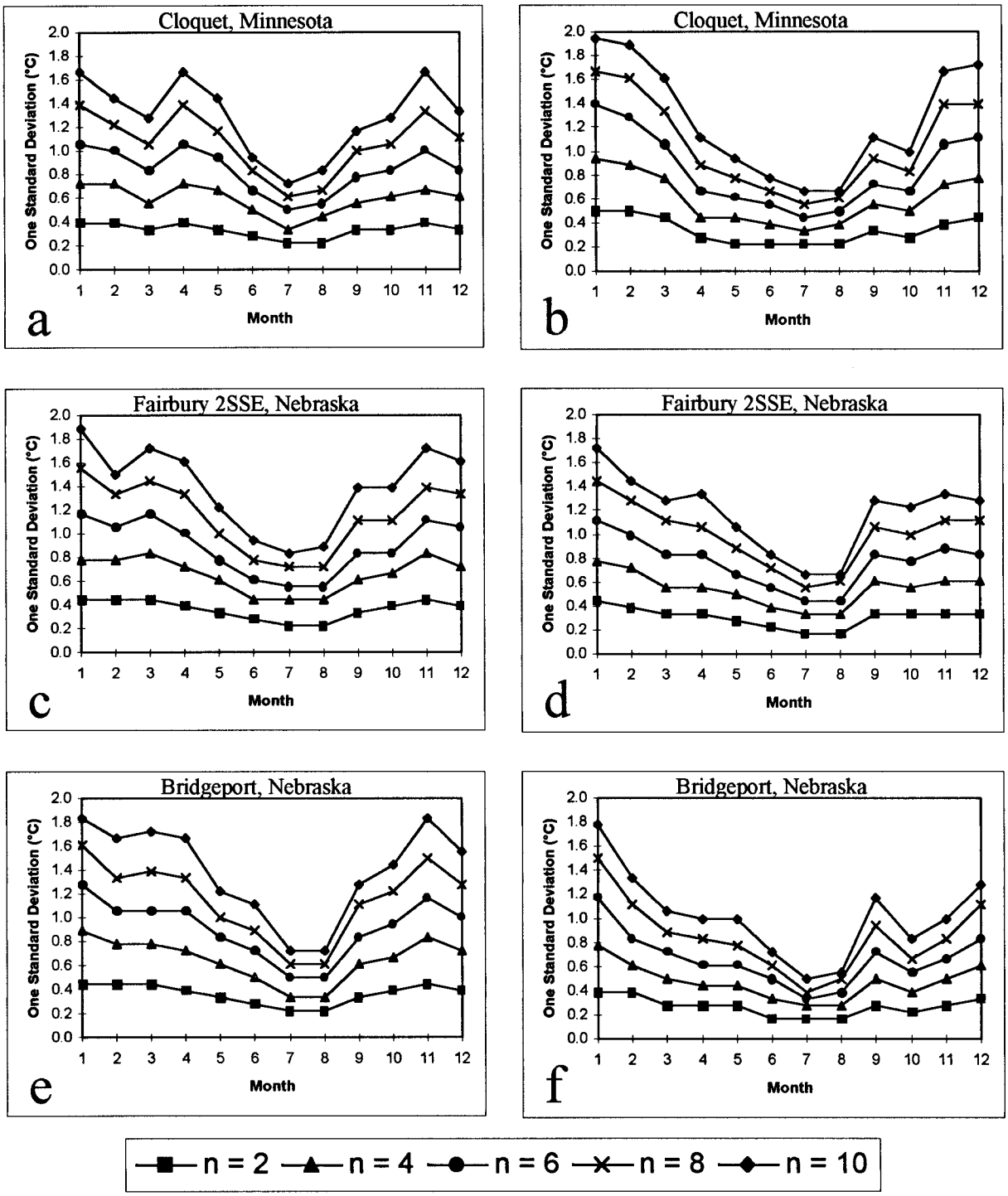

FIG. 7. Standard deviations by month for maximum [(a), (c), and (e)] and minimum [(b), (d), and (f)] temperatures for $n = 2$, 4, 6, 8, and 10 data gap days at Cloquet, MN; Fairbury, NE; and Bridgeport, NE.

The July situation is different and more complex. During July, the Atlantic Ocean will have a greater modifying impact than the Pacific Ocean. The strong westerly airflow of winter has relaxed and other flow patterns predominate. The Atlantic coastal region is generally under the influence of a subtropical high pressure system, the ''Bermuda high.'' The whole region will be under southerly flow. The warm Gulf

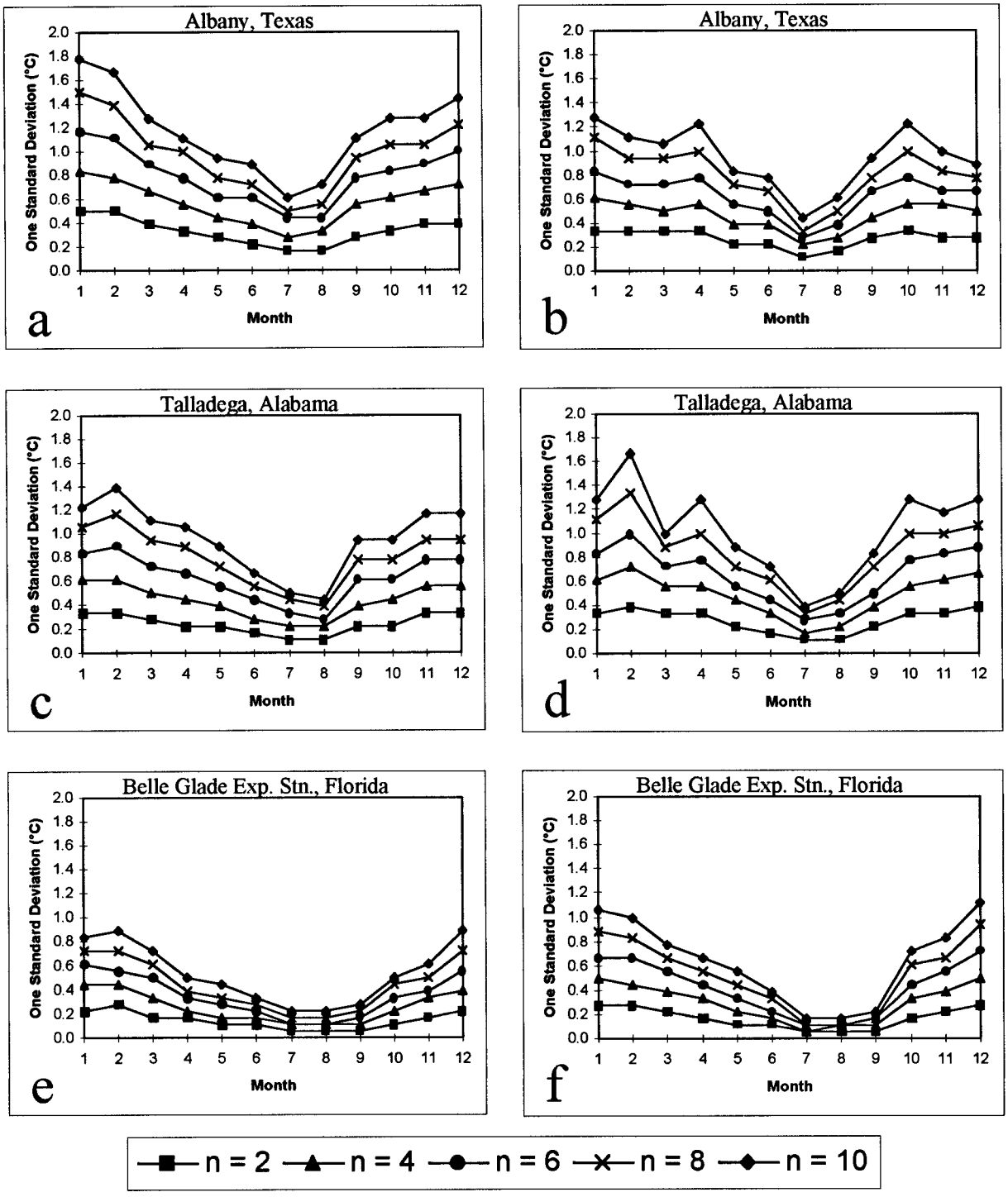

FIG. 8. Standard deviations by month for maximum [(a), (c), and (e)] and minimum [(b), (d) and (f)] temperatures for $n = 2$, 4, 6, 8, and 10 data gap days at Albany, TX; Talladega, AL; and Belle Glade, FL.

Stream will be advecting very warm (>25°C) water northward just off the coast. Thus, southerly airflow or onshore flow from the Atlantic Ocean will give the coastal region hot and humid weather with little day to day change. In the Pacific coastal region, onshore flow is cool and damp while offshore flow is warm and dry. Thus, the Pacific coastal region will be impacted by air masses that have a greater variability than the

Atlantic coastal region. This leads to a greater day to day temperature variability.

If the data gaps can affect the measure of central tendency, then it is logical to suspect that gaps can also have a significant effect on other analyses. For example, based on our results we would expect the effect of data gaps to be greater on heating degree day analyses than on cooling degree analyses. The effect of data gaps on growing degree day analyses will be variable with generally more effects due to data gaps early and late in the growing season.

6. Conclusions

In this paper we have quantified the effects of gaps in data on the calculations of monthly mean maximum and minimum temperatures in the continental United States. Our results show that it is critical that the possible effects of data gaps on particular analysis are evaluated and *reported.*

The effect of data gaps is most severe during the winter. In January, one standard deviation from the true mean for maximum and minimum temperatures ranges from less the 0.3°C for 2-day gaps to greater than 2.2°C for 10-day gaps. In July, the range is from less than 0.1°C for 2-day gaps to greater than 0.7°C for 10-day gaps.

The effects of data gaps also have a spatial/climatic region component. Marine west coast climates (Olga 2SE, WA) were affected less severely than continental (cold) climates (Cloquet, MN).

These results will assist researchers in climate variability and climate trend studies to quantify the confidence intervals when there are gaps in the data record. With only a 2-day data gap in continental regions, the standard deviation is on the order of 0.5°C. Thus, caution is recommended when monthly means are calculated with data gaps. With climate variability and climate trend analysis, data gaps are especially a concern when they occur early or late in the record. These results show the need for researchers to have good metadata so that they can quantify the impacts of data gaps in their climate studies.

REFERENCES

Guttman, N. B, 1991: January singularities in the Northeast from a statistical view. *J. Appl. Meteor.,* **30,** 358–367.

——, and M. S. Plantico, 1987: Climate temperature normals. *J. Climate Appl. Meteor.,* **26,** 1428–1435.

Hughes, P. Y., E. H. Mason, T. R. Karl, and W. A. Brower, 1992: United States Historical Climatological Network daily temperature and precipitation data. Carbon Dioxide Information Analysis Center, Environmental Sciences Division, Publ. 3778, 129 pp. [Available from Carbon Dioxide Information Analysis Center, P. O. Box 2008, MS 6335, Oak Ridge, TN 37831.]

NOAA, 1996: Climatological Data Nebraska. Vol. 101, No. 1, 49 pp. [Available from National Climatic Data Center, Room 120, 151 Patton Ave., Asheville, NC 28801-5001.]

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in FORTRAN: The Art of Scientific Computing.* 2d ed. Cambridge University Press, 963 pp.

Trewartha, G. T., and L. H. Horn, 1980: *An Introduction to Climate.* 5th ed. McGraw-Hill, 416 pp.