Creating a Serially Complete, National Daily Time Series of Temperature and Precipitation for the Western United States

JON K. EISCHEID

Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado

PHIL A. PASTERIS

Natural Resources Conservation Service, Portland, Oregon

HENRY F. DIAZ

NOAA Climate Diagnostics Center, Boulder, Colorado

MARC S. PLANTICO AND NEAL J. LOTT

NOAA National Climatic Data Center, Asheville, North Carolina

(Manuscript received 10 November 1998, in final form 16 October 1999)

ABSTRACT

The development of serially complete (no missing values) daily maximum–minimum temperatures and total precipitation time series over the western United States is documented. Several estimation techniques based on spatial objective analysis schemes are used to estimate daily values, with the ''best'' estimate chosen as a missing value replacement. The development of a continuous and complete daily dataset will be useful in a variety of meteorological and hydrological research applications.

The spatial interpolation schemes are evaluated separately by interpolation method and calendar month. Cross validation of the results indicates a distinct seasonality to the efficiency (error) of the estimates, although no systematic bias in the estimation procedures was found. The resulting number of serially complete daily time series for the western United States (all states west of the Mississippi River) includes 2034 maximum–minimum temperature stations and 2962 total daily precipitation locations.

1. Introduction

There is great demand in the climate community and many federal agencies for quality-controlled, serially complete climate datasets for natural resource modeling. With many agencies increasingly relying on models to determine the appropriate management decisions, the need for accurate climate datasets for model validation and verification has never been greater.

Traditionally, model users must first identify, then remove or correct extreme errors and/or missing data. These developers may or may not be knowledgeable about the intricacies of the data being processed and often develop algorithms to overcome data problems that may introduce additional uncertainties into the data. Uncoordinated and independent data correction and es-

timation for specific regions or states may be redundant, expensive, and sometimes erroneous. The effect of missing data, or data gaps, in the calculation of monthly mean temperature can result in errors that exhibit temporal and spatial patterns (Stooksbury et al. 1999). This project has the objective to create serially complete daily datasets in a systematic, well-documented fashion that can be utilized for many hydrologic and other natural resource conservation models.

2. Project objective

The objective of this project is to create a serially complete (no missing data values) daily temperature and precipitation dataset (initially for the period 1951–91) for the United States in support of a wide variety of ecosystem resource models. The completed serially complete daily dataset will be archived at the National Climatic Data Center (NCDC) in Asheville, North Carolina, and will be made available to the Natural Resources Conservation Service (NRCS), U.S. Depart-

Corresponding author address: Dr. Jon K. Eischeid, CIRES, University of Colorado, Campus Box 449, Boulder, CO 80309-0449.
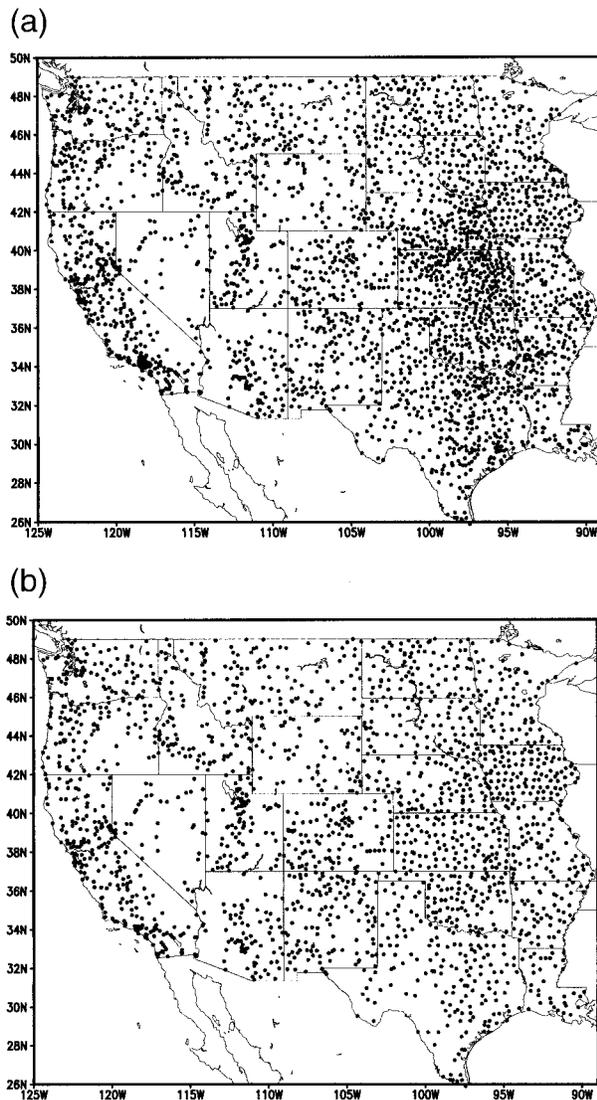E-mail: jon@cdc.noaa.gov

(a)

(b)

FIG. 1. (a) Geographical distribution of daily precipitation measurements designated as target stations (2962 stations). (b) Geographical distribution of mean daily maximum–minimum temperature measurements designated as target stations (2034 stations).

TABLE 1. An example of adjusting estimated data to accumulated totals (−9999 denotes missing 24-h values).

	Station identification number: 140471		
	Adjustment ratio is 1.35 (1.80/1.33)		
Date	Original estimate (in.)	Adjusted estimate (in.)	Reported
25 Aug 1979	0.03	0.04	−9999
26 Aug 1979	0.14	0.19	−9999
27 Aug 1979	0.24	0.32	−9999
28 Aug 1979	0.00	0.00	−9999
29 Aug 1979	0.92	1.25	−9999
Sum	1.33	1.80	1.80 in.

TABLE 2. Frequency (%) of the number of times each of the interpolation methods was chosen as "best" for the two time categories. See section 3 for definitions of the method codes.

	Method					
	NR	IDW	OI	MLAD	SBE	MED
	Category 1, A.M. reader					
Jan precipitation	8.2	4.0	6.8	61.2	2.3	17.5
Jul precipitation	15.8	5.2	10.8	44.1	1.6	22.5
Jan max temp	6.7	3.1	4.0	83.7	0.3	2.1
Jul max temp	7.6	3.6	0.9	83.3	0.6	4.0
Jan min temp	9.5	6.2	8.6	62.5	1.2	12.0
Jul min temp	7.6	1.8	1.8	84.4	0.3	4.0
	Category 2, P.M. reader					
Jan precipitation	5.0	2.5	5.3	67.4	1.1	18.8
Jul precipitation	20.1	9.7	5.1	46.6	2.5	16.0
Jan max temp	2.6	1.6	6.1	84.6	0.2	5.0
Jul max temp	6.9	2.6	2.4	83.5	0.6	5.8
Jan min temp	7.5	4.0	11.3	59.3	0.4	17.5
Jul min temp	4.6	1.2	2.0	89.1	0.4	2.8

ment of Agriculture, and the climate community through the Unified Climate Data Access Network. The goal is to create serially complete datasets based on approximately 4775 stations available from the NCDC Climatography of the United States No. 81—Monthly Station Normals, 1961–1990 (NCDC CLIM 81) (Owenby and Ezell 1992). The purpose of this paper is to present the basic details/results of the creation of serially complete data for states west of the Mississippi River.

This project is funded by the NRCS National Water and Climate Center located in Portland, Oregon, and cosponsored by the NCDC and the Climate Diagnostics Center located in Boulder, Colorado.

3. Project approach

The project uses a multistep approach to process observed daily precipitation totals and mean daily maximum and minimum temperatures to create serially complete daily datasets. The creation of a serially complete dataset includes the replacement of missing daily values through the use of simultaneous values at nearby stations to calculate an estimated value for that particular day (all days in which an estimate is derived for a missing value are flagged as such). Station histories are reviewed and appropriate stations are selected and designated as "target stations" for estimation. Data observation times are reconciled and categorized to allow accurate spatial interpolation from neighboring stations determined to have sufficient record (greater than 10 yr) to provide stable estimation statistics with the target stations.

Six different methods of spatial interpolation are used to create the serially complete dataset. The methods are defined as 1) the normal ratio method (NR), 2) simple inverse distance weighting (IDW), 3) optimal interpolation (OI), 4) multiple regression using the least absolute deviation criterion (MLAD), 5) the single best

TABLE 3. Summary statistics for the monthly distribution of observed vs estimated correlation coefficients for maximum temperature. In general, the poorest estimates are found in the topographically diverse terrain of mountain regions as well as coastal areas of California, Oregon, and Washington.

	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	0.975	0.976	0.975	0.972	0.967	0.961	0.946	0.950	0.966	0.970	0.974	0.973
Median	0.957	0.961	0.960	0.957	0.954	0.944	0.926	0.929	0.953	0.958	0.958	0.955
Lower hinge	0.924	0.932	0.936	0.936	0.934	0.928	0.892	0.898	0.930	0.940	0.932	0.921
Number of stations	2034											

estimator, and 6) the median (MED) of the previous five methods (Eischeid et al. 1995).

Statistical summaries are generated using cross correlations between observed daily values and those estimated for each of the six different methods described. The six techniques respond to variations in season and geography, and the best estimation method is selected based on the efficiency of the estimate over time. The cross correlations are used to measure the efficiency of each method, and the method that exhibits the highest correlation relative to the other methods is utilized to replace missing values. Performance in terms of the root-mean-square (rms) statistic is also provided.

4. Dataset description and processing

The primary dataset used in this project was the NCDC Summary of Day (TD3200). Quality control performed by NCDC on this dataset included a procedure (Reek et al. 1992) that identified and flagged nearly 400 000 data discrepancies (the number of discrepancies is less than 1% of the total observations). To be included in the serialization procedure, a station could not have more than 48 missing months of data for the entire period of record (1951–91). A month was marked as missing if it contained more than 14 consecutive days of missing temperature or precipitation.

The 22 states west of the Mississippi River were examined and 2962 precipitation, 2034 maximum, and 2035 minimum temperature reporting stations were selected for serialization. The 2962 target precipitation stations are shown in Fig. 1a and the 2034 maximum–minimum target temperature stations are shown in Fig. 1b. Although only target stations are shown here, many more stations were used with less stringent serial requirements (i.e., 10 yr of record) in the estimation procedures as a means of enhancing the pool of potential predictors. Also, stations from bordering states were extracted to improve the spatial distribution of sites surrounding target stations located near state borders. The resulting total number of stations, including target stations, utilized for each of the estimation methods is 6353
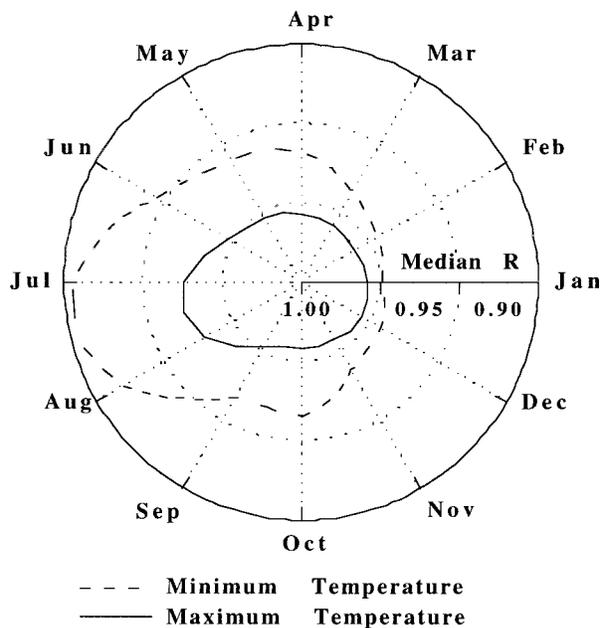

FIG. 2. Monthly distribution of the observed vs estimated correlation coefficients (R) for maximum and minimum temperatures.
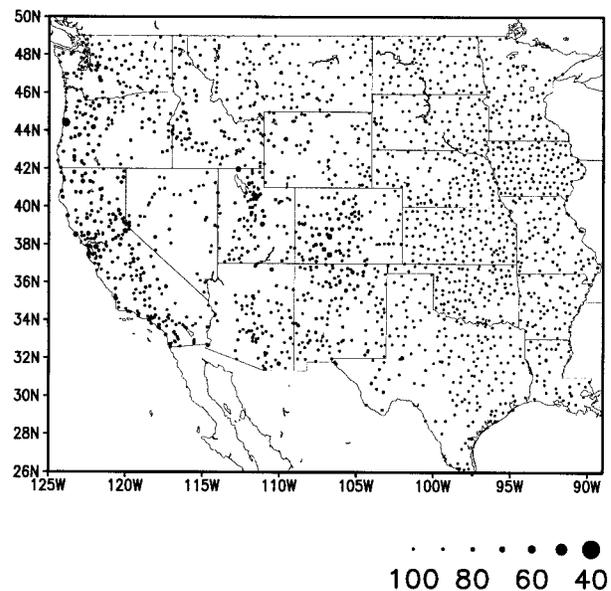

FIG. 3. Geographical distribution of the observed vs estimated correlation coefficients for maximum temperatures for Jan. The scale of correlation coefficients is inverted (i.e., low R to high R is plotted as large to small) to highlight the poorest correlations.
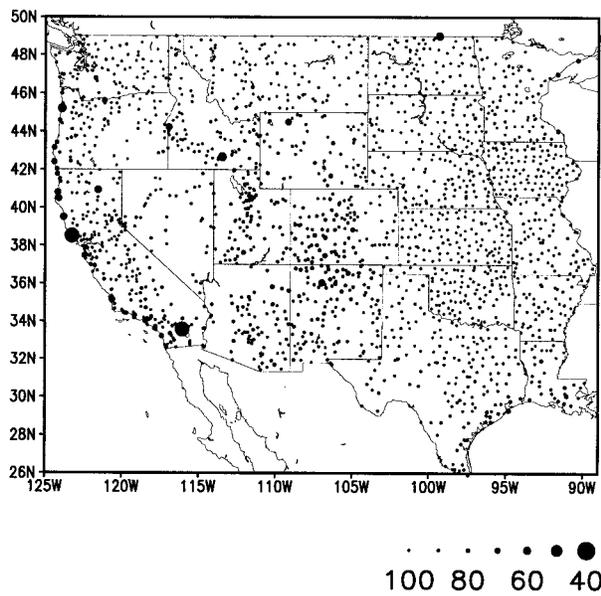
FIG. 4. As in Fig. 3 but for Jul.

FIG. 5. Monthly distribution of the observed vs estimated rmse for maximum and minimum temperatures (°F).

precipitation and 4476–4463 maximum–minimum temperature stations.

The 22 states reflect a wide variety of terrain and a diversity of climatic regimes, which allows a means for testing the efficacy of daily estimates for regional and seasonal differences. In addition, with few exceptions, the geographic distribution across the western states is relatively uniform, which provides a stable estimation environment and a substantial number of serially complete stations for natural resource modeling.

Categorization of observation times

All estimation methods are dependent on a clear identification of each station's time of observation. Because observation times within the Cooperative Network can vary from station to station (and sometimes within station over the period of record), it is important to determine and group stations correctly by their observation time. The "true" observation time for a station can be inferred from the TD3200 dataset, although observation time irregularities do exist and must be accounted for. The true observation times for the period 1967–81 were recovered from the NCDC CLIM-81 dataset (Owenby and Ezell 1992). Once corrected observation times were
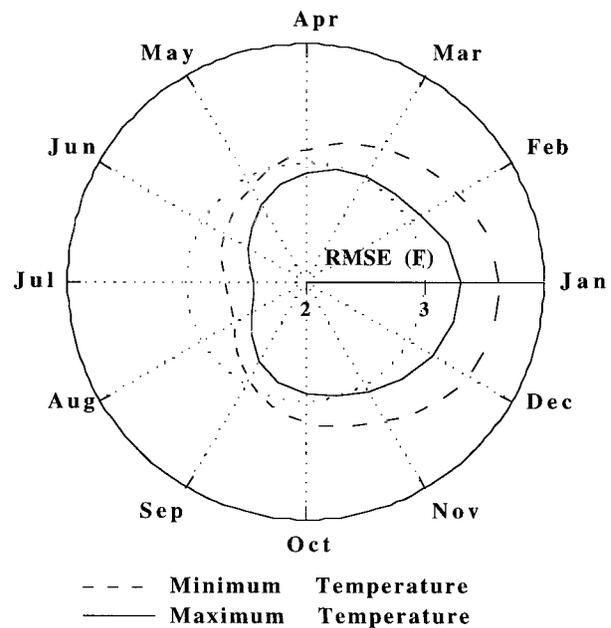
determined, the following observation time categorizations were made:

category 1: 0500–1100 LST—A.M. reader;
category 2: 1200–2000 LST—P.M. reader;
category 3: 2100–0400 LST—midnight reader.

Target station estimations were made using the "corresponding" observation time category from neighbor stations. After several target station estimation runs, it was determined that category-3 stations provided an insufficient sample size to provide accurate estimates. Category-3 stations were then combined with category-2 stations for the final runs. Implicit then in the following estimation analyses, is that estimates are computed twice: the "morning" and "evening" stations are done separately with stations in each category varying by parameter and month. Summary statistics characterizing the quality of derived daily values are presented collectively for the target stations.

5. Estimation methodology

The replacement of missing daily values for maximum–minimum temperatures and total precipitation in-

TABLE 4. Summary statistics for the monthly distribution of the rmse between observed and estimated maximum temperatures (°F).

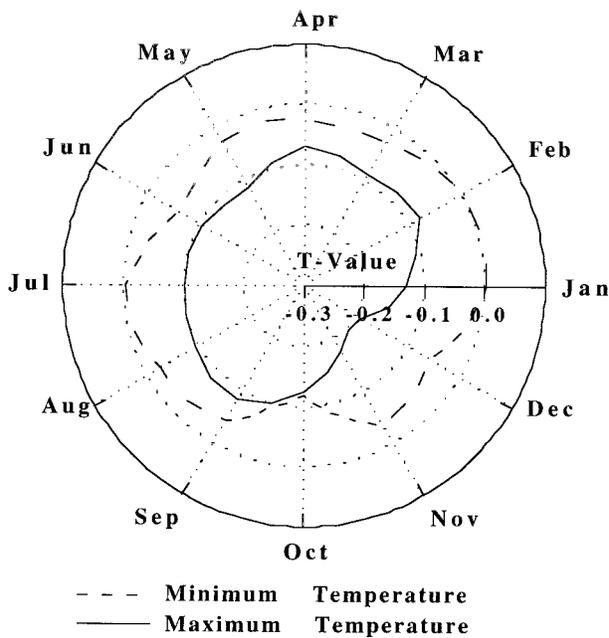|	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	3.96	3.70	3.58	3.52	3.28	3.04	2.83	2.94	3.30	3.46	3.68	3.88
Median	3.32	3.14	3.03	2.92	2.75	2.56	2.44	2.52	2.79	2.94	3.08	3.25
Lower hinge	2.73	2.61	2.54	2.50	2.35	2.18	2.12	2.19	2.40	2.49	2.58	2.71
Number of stations						2034						

FIG. 6. Monthly distribution of the *t* statistic for maximum and minimum temperatures.

cludes the use of nearby simultaneous values to calculate an estimated value at the target station over the period of time for which adequate data are available. The efficiency, or accuracy, of the estimates over a long period of time provides the information used to assess the quality of estimated daily values. Estimated daily values are used in lieu of missing values as a means of making a particular station serially complete.

There are numerous spatial interpolation methods available for point estimation with irregularly spaced data. Typically, the choice of methodology is dependent on several factors: the meteorological variable under consideration, the geographical area, the spatial distribution of surrounding observations, and the day–month–season for which the target station is to be estimated (Schlatter 1975; Bennet et al. 1984; Thiebaux and Pedder 1987).

Using regression-based methods, Kemp et al. (1983) found that the mean absolute error associated with minimum temperatures was reduced by 50% when compared with within-station methods. Additional investigations performed by the Northeast Regional Climate Center (DeGaetano et al. 1993) have shown that re-

gression-based methods of data estimation tend to be more accurate than within-station methods. Additional work (Huth and Nemesova 1995) has shown that other weather elements, such as relative humidity, wind speed, and cloudiness, contribute very little to regression-based methods and that temperature at neighboring stations has by far the highest spatial correlations.

DeGaetano goes on to mention that "while such methods are useful over limited areas, they are computationally intensive and therefore not feasible when data estimates are needed for a large number of stations over a long period of time" (DeGaetano et al. 1993). These limitations have been partially overcome with the use of new high-speed workstations and large mass storage capabilities that now provide the horsepower required to perform these intensive calculations in a reasonable time period.

Because estimates are required for each day separately over a variety of terrain with a differing number of available surrounding observations, we have chosen several different methods for testing. This section will focus on six methods of spatial interpolation. Each of the six methods are compared by month for each station, and the one with the highest correlation to the target station was chosen as the method used to replace any missing daily values at that location (Eischeid et al. 1995).

In any spatial interpolation scheme the selection and quantity of surrounding stations are critically important to the results of the interpolations. Problems arise when using climatological data because of missing values and the varying availability of stations through time. In order to determine which stations are to be used, surrounding stations are preselected based on their relationship with the target station. The 15 closest stations are identified for each target station and are ranked by the value of the correlation coefficient between the candidate station and its neighbors.

Correlation coefficients are computed for each month separately utilizing the daily data. The stations with the largest positive correlation coefficients, where the minimum criterion is an *r* of at least 0.35, are subsequently used in the estimation procedures. A minimum of one station is needed to compute the estimate at the target station, with a maximum of four. Tests have shown that inclusion of more than four stations does not significantly improve the interpolation and may in fact degrade the estimate.

TABLE 5. Summary statistics for the monthly distribution of the *t* statistic between observed and estimated maximum temperatures.

	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	0.34	0.34	0.33	0.36	0.38	0.45	0.53	0.55	0.41	0.31	0.25	0.27
Median	−0.13	−0.08	−0.09	−0.07	−0.12	−0.11	−0.10	−0.09	−0.08	−0.13	−0.18	−0.19
Lower hinge	−0.68	−0.52	−0.55	−0.50	−0.57	−0.62	−0.74	−0.70	−0.61	−0.60	−0.63	−0.74
Number of stations					2034							

TABLE 6. Summary statistics for the monthly distribution of observed vs estimated correlation coefficients for minimum temperature.

	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	0.967	0.965	0.961	0.949	0.942	0.926	0.902	0.911	0.948	0.948	0.959	0.963
Median	0.948	0.944	0.935	0.918	0.912	0.895	0.856	0.870	0.916	0.916	0.936	0.942
Lower hinge	0.911	0.906	0.884	0.863	0.859	0.846	0.785	0.810	0.867	0.877	0.896	0.902
Number of stations						2035						

The number (never greater than four) of neighboring stations meeting the criteria is not fixed in time. It varies depending on available station data for the year/month/day in question. As such, the interpolation models may also change in time. Moreover, the surrounding stations that may be optimal for a particular calendar month (e.g., January) may not be optimal for a different month (e.g., July). Thus, the station selection procedures are computed for each calendar month separately. The selected neighboring stations are identical for each of the six interpolation methods. A brief description of each method is given below.

a. Normal ratio method

The normal ratio (NR) method of spatial interpolation was first proposed by Paulhus and Kohler (1952). The current analysis uses a modified version, which is described by Young (1992). Weights for the surrounding stations used in the estimation algorithm are found according to

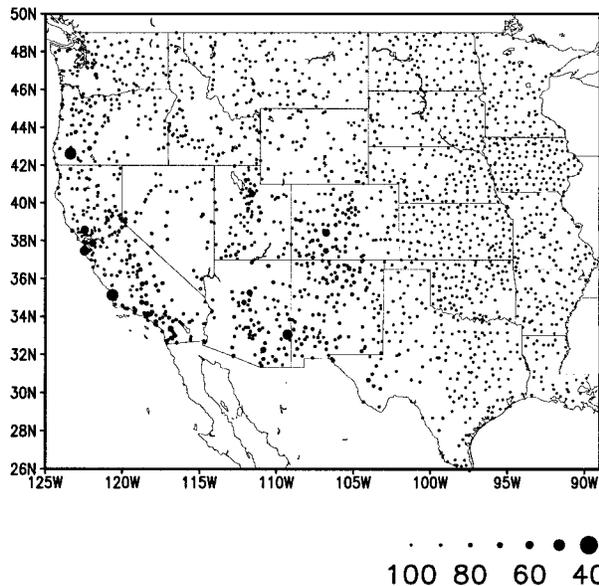$$W_i = \frac{r_i^2(n_i - 2)}{1 - r_i^2}, \qquad (1)$$

FIG. 7. Geographical distribution of the observed vs estimated correlation coefficients for minimum temperatures for Jan. The scale of correlation coefficients is inverted (i.e., low R to high R is plotted as large to small) to highlight the poorest correlations.

where r is the correlation coefficient for each daily time series between the target station and the ith surrounding station, n is the number of points used to derive the correlation coefficient, and W_i is the resultant weight.

b. Inverse distance method

The inverse distance method (IDW) is a simple distance-weighted ''area average'' estimate of the value at the target station. The assumption here is that surrounding stations are related to the target station by their proximity to the target station. This procedure is given by

$$\hat{Z} = \frac{\sum_{i=1}^{n} W_i Z_i}{\sum_{i=1}^{n} W_i}, \qquad (2)$$

where Z_i is the particular monthly anomaly at the ith surrounding station, and the weight function W_i is derived from the inverse of the distance from the target station to the ith surrounding station.

c. Optimal interpolation

Early uses of optimal interpolation (OI) in meteorology may be traced back to Gandin [see, e.g. Gandin and Kagan (1974)]. Since that time it has had wide usage in climatology and meteorology. In most applications OI is used to estimate values at a target site, for example, a grid point. Here we use a univariate OI to estimate values at a known station location. Optimal interpolation is a spatial interpolation technique that assigns weights to the observed difference values (observed minus first guess) at the selected neighboring station locations,

$$\hat{Z} = Z_f + \sum_{i=1}^{n} W_i(Z_{oi} - Z_{fi}), \qquad (3)$$

where Z_{oi} and Z_{fi} are the observed and first-guess values at the ith neighbor station, respectively, and Z_f is the first-guess value at the target station location being estimated. The first-guess values used in this particular application are concurrent values from the closest station with the highest correlation. The weighting coefficients W_i are determined in an objective manner such that the rms error (rmse) of the analyzed difference values at each target location is minimized over the spatial domain.

TABLE 7. Summary statistics for the monthly distribution of the rmse between observed and estimated minimum temperatures (°F).

	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	4.58	4.24	3.94	3.66	3.43	3.34	3.19	3.18	3.47	3.70	4.04	4.49
Median	3.62	3.47	3.29	3.11	2.91	2.78	2.68	2.69	2.96	3.18	3.35	3.59
Lower hinge	3.00	2.89	2.78	2.71	2.54	2.36	2.22	2.26	2.56	2.78	2.85	2.98
Number of stations						2035						

The weights are dependent upon the spatial autocorrelations among the surrounding observation values and are typically modeled mathematically as a function of distance separating the neighbors and the target location. Rather than model each spatial domain (the area surrounding each target station) individually, and to compensate for possible anisotropic fields, we use the actual relationships among all observations by directly using the calculated correlation coefficients.

Once the correlation coefficients are known, the weights needed to solve Eq. (4) are given by the solution of a system of linear algebraic equations. In matrix form this can be written as

$$W_i C_{ir} = G_r, \qquad (4)$$

where $i = 1, 2, \ldots, N$ (number of surrounding observations) whose coefficients are given by the correlation with selected neighboring station observations ($C, n \times n$) and correlation coefficients from the target station to each of the surrounding stations ($G, 1 \times n$ vector). Because the empirical correlation coefficients are used, the solution matrix could fail to be positive definite, but, in practice, this was a rare occurrence.
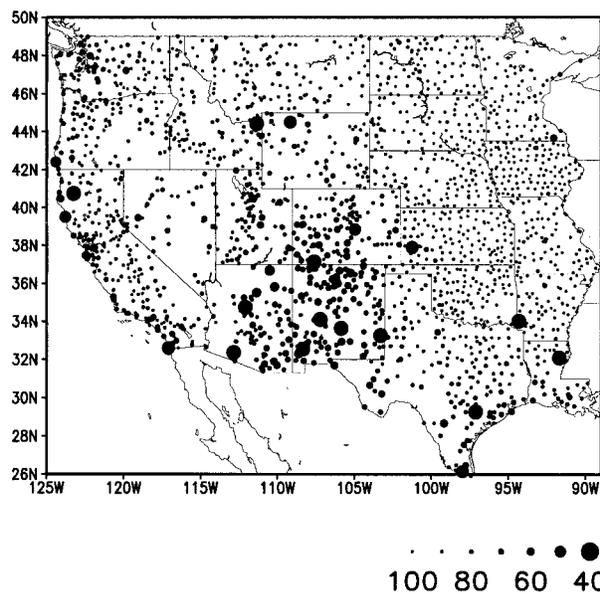

FIG. 8. As in Fig. 6 but for Jul.

d. Multiple regression, least absolute deviations criteria

The method of multiple regression using the least absolute deviations criteria (MLAD) is a robust version of the general linear least squares estimation. The method of least squares is an effective method when the errors are normally distributed and independent. However, for precipitation data especially, the assumption of normality over the wide range of situations can lead to poor estimations. The principal advantage of least absolute deviations is its resistance to outliers and to overemphasis of large-tailed distributions (Barrodale and Roberts 1973).

MLAD estimates the unknown parameters in a stochastic model so as to minimize the sum of absolute deviations of the neighboring station observations from the values predicted by the model. Regression coefficients b are calculated so as to minimize

$$\sum_i \left| \sum_j x_{ij} b_j - Y_i \right|, \qquad (5)$$

where $x, i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ denote a set of n measurements on m surrounding stations (independent variables), and $y, i = 1, 2, \ldots, k$ denote the associated measurement on the dependent (target station) value. The linear programming techniques of Barrodale and Roberts (1973) are used to accomplish this task.

e. Single best estimator

The single best estimator (SBE) is simple and analogous to using the closest neighboring station as an estimate for the target station. The target station is estimated using the actual observed value from the neighboring station that has the highest positive correlation with the target station.

f. Median

The median method (MED) is not a true interpolation model but is simply the median value obtained from the above five estimation methods. By using the median we implicitly account for the estimation formula to change over time, which may yield a better long-term estimate.

TABLE 8. Summary statistics for the monthly distribution of the *t* statistic between observed and estimated minimum temperatures. As a result, a greater number of stations are found to be statistically significant at the 0.5% level. In the aggregate, the estimation of minimum daily temperatures is less effective than that shown for daily maximum temperatures.

	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	0.89	0.70	0.61	0.50	0.55	0.55	0.74	0.73	0.54	0.50	0.62	0.81
Median	0.00	0.00	−0.03	−0.03	−0.03	−0.06	0.00	−0.03	−0.04	−0.12	−0.04	−0.06
Lower hinge	−0.91	−0.66	−0.76	−0.70	−0.68	−0.66	−0.85	−0.83	−0.76	−0.82	−0.72	−0.97
Number of stations						2035						

6. Basic continuity checks

After estimating daily maximum and minimum temperatures, a series of internal consistency checks were performed to ensure that estimates did not violate obvious constraints associated with recording maximum and minimum temperatures. Typical tests include identifying estimated maximum temperatures lower than a previous day's minimum and an estimated maximum lower than a minimum for the same day.

These inconsistencies were corrected by assigning corrected maximums or minimums where appropriate or averaging the maximum and minimum temperatures for the previous and subsequent days.

Distribution of accumulated precipitation

One of the more vexing problems associated with creating a serially complete daily precipitation time series is the distribution of multiday (accumulated) precipitation totals. These values are generally flagged with an ''A'' and do provide a valuable target value for the estimation procedure, because a primary goal of the estimation procedure is to not bias (neither increase or decrease) the observed monthly precipitation total.

FIG. 9. Monthly distribution of the observed vs estimated correlation coefficients (*R*) for precipitation.

Several adjustment methods exist; however, the simplest and most computationally efficient was to estimate independently the individual missing days and also the day shown with an accumulated amount, sum the estimates for the same period, and take the ratio of the observed accumulated to the sum of the estimated amounts.

The ratio would be applied to the daily estimates for the corresponding period to match the observed accumulated total flagged as A. The ratio method can increase or decrease the daily estimates and is constrained by the observed accumulated precipitation total. Table 1 illustrates the procedure for a station in Kansas for which the measured precipitation was reported as 1.80 in. on 29 August as an accumulation of the previous 5 days. The adjustment ratio is derived from the ratio of the measured total to the estimated total for the same 5 days (in the example the estimated total is 1.33 in.).

7. Description of files created

The estimation procedure produces three files: 1) the original data values, 2) the estimated values, and 3) the merged original values and estimates along with a flag that indicates the method used to estimate the value. These files were analyzed to produce the statistics discussed in the next section.

8. General results

In order to test the efficacy of the estimation techniques, each of the six interpolation methods is compared with respective nonmissing observations. For each station six estimates are computed for the entire period of record as if all observations were missing. Comparisons among the techniques are then based on the correlation coefficient *R* between the original observed daily values and the corresponding estimated values. At each station, for each month, six correlation coefficients are calculated and ranked. The method that exhibits the largest *R* is considered to be the most representative and is used to determine which estimates are used for further analysis in addition to replacement of missing values. In Table 2 we show these results, as a percentage of the total, for all stations in the study. In general, stations may, and quite often do, require the use of different models depending on the month in ques-
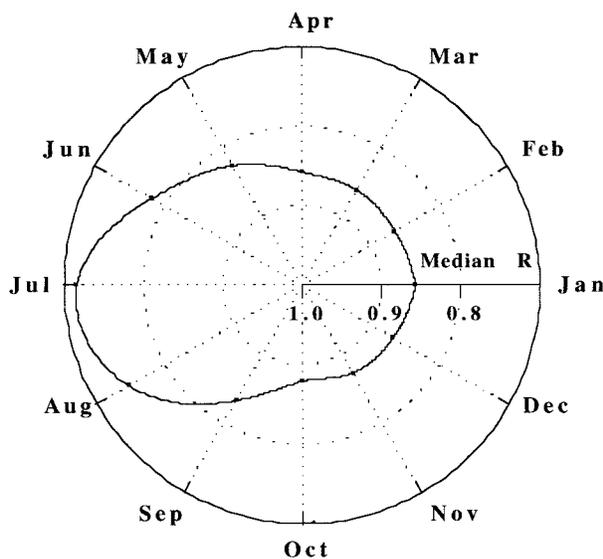
TABLE 9. Summary statistics for the monthly distribution of observed vs estimated correlation coefficients for precipitation.

	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	0.910	0.914	0.910	0.900	0.870	0.831	0.800	0.819	0.882	0.918	0.920	0.919
Median	0.857	0.864	0.861	0.857	0.826	0.782	0.716	0.747	0.833	0.879	0.873	0.867
Lower hinge	0.783	0.787	0.787	0.792	0.758	0.707	0.615	0.638	0.765	0.817	0.804	0.794
Number of stations						2962						

tion though it is clear that the MLAD interpolation method outperforms the other five estimation techniques. Additionally, there is no apparent time-of-observation bias with regard to method, that is, there was no significant difference in the observed versus estimated error for morning times versus afternoon/evening times. A flag indicating which method was used to estimate the value, and the corresponding correlation coefficient, are provided with the metadata. In general our concern was how well we were able to estimate daily temperature and precipitation values. A description of the accuracy/ efficiency of each of the methods is not presented here [see Eischeid et al. (1995) for an analysis of these methods for monthly time series of temperature and precipitation]. Rather, the following section outlines the efficacy of the ''best'' estimate regardless of which estimating method was chosen. It should be noted that, although the results are presented on a monthly basis, the summary statistics are computed from daily values.

a. Maximum temperature

The efficiency of the best estimation is summarized in Fig. 2 via the range of observed versus estimated correlations found for all maximum temperature reporting stations. The median correlation is above 0.90 for all months ranging from a low of 0.93 in July and August to a high of 0.96 during the spring and autumn transition seasons (see Table 3). The spatial distribution of the maximum temperature correlations for all stations is shown in Figs. 3 and 4 for January and July, respectively.

In general, the poorest estimates are found in the topographically diverse terrain of mountain regions as well as coastal areas of California, Oregon, and Washington. Conversely, the best estimates are obtained in areas of relatively even terrain where spatial relationships among surrounding stations (those used in the estimation procedure) are stronger and cover a greater areal extent than mountainous regions. In addition, the regions that exhibit the best correlations typically have a better, less scattered, spatial distribution of surrounding stations.

The distributions of the rmse between the observed and estimated series are presented in Fig. 5 and Table 4 for each month. For nearly all stations the rmse is less than 5°F in all months with the lowest values shown for the summer months (July median of 2.4°F) and the largest during winter (January median of 3.3°F). The geographic areas of small/large rmse are generally analo-
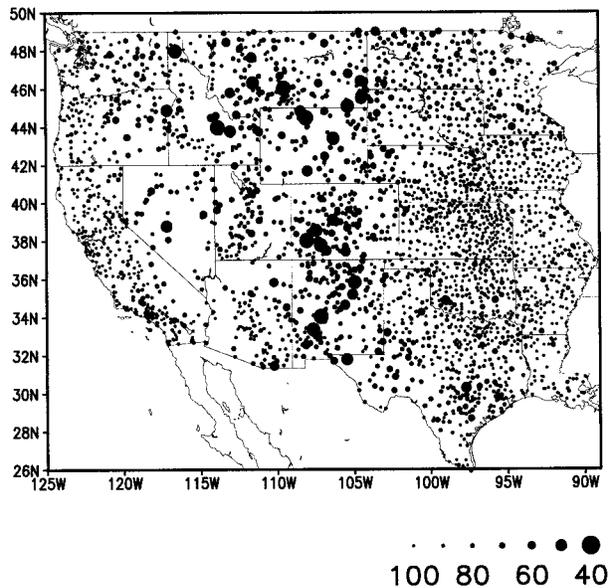

FIG. 10. Geographical distribution of the observed vs estimated correlation coefficients for daily precipitation for Jan. The scale of correlation coefficients is inverted (i.e., low R to high R is plotted as large to small) to highlight the poorest correlations.
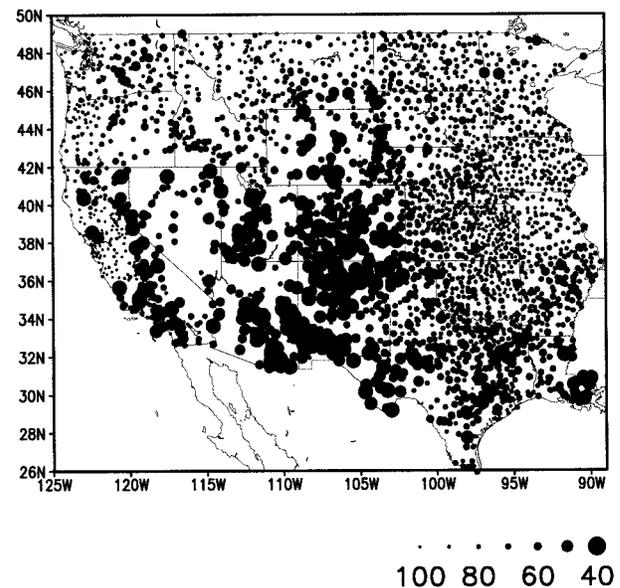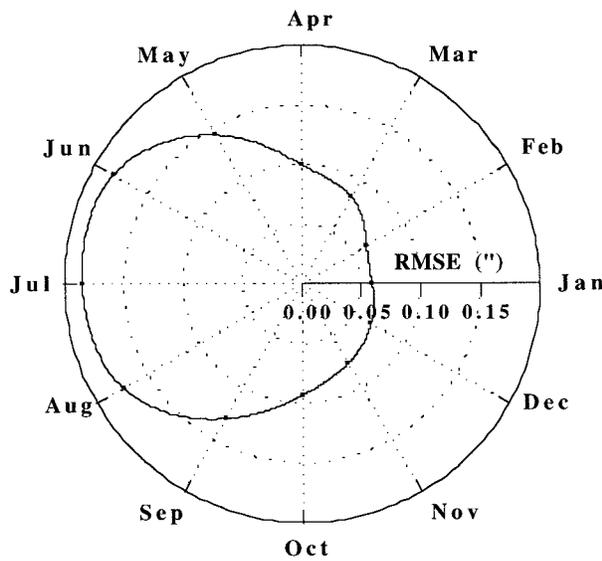

FIG. 11. As in Fig. 10 but for Jul.

FIG. 12. Monthly distribution of the observed vs estimated rmse for precipitation (inches).
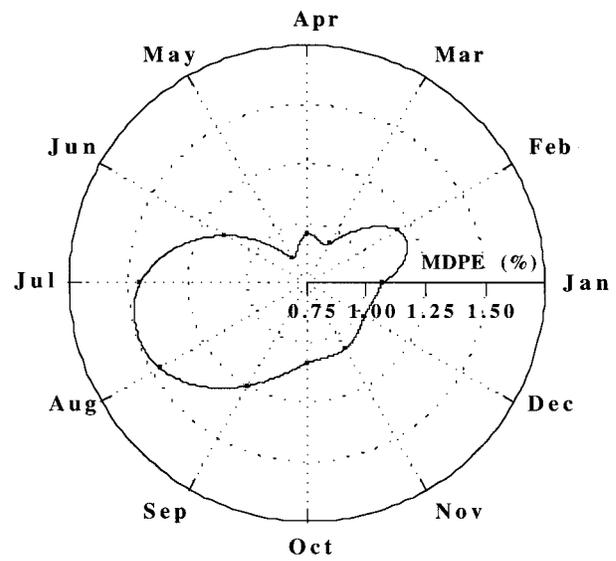

FIG. 13. Monthly distribution of the observed vs estimated mean daily percentage error (MDPE) for precipitation (%).

gous to regions where the correlations are high/low (see Figs. 3 and 4).

A simple *t* test was computed for observed and estimated maximum temperature for the 2034 stations and the range of monthly *t* values is summarized in Fig. 6 and Table 5. The range of the *t* statistic is typically larger, based on previous results, for the summer months and smallest during winter. Although a number of stations reveal significant differences, not unexpected given the large sample size, the preponderance of stations showed no statistically significant difference among the observed and estimated daily means. The spatial distribution (not shown) of the *t* statistic at the 0.5% level is spread evenly among the 22 states. No significant bias was detected in any of the months with respect to over- or underestimating observed values.

b. Minimum temperature

Analysis of minimum daily temperatures produced results broadly similar to the maximum daily temperatures. For the most part the distributions of observed versus estimated correlations (Fig. 2, Table 6) exhibit a greater range in values relative to results shown for maximum temperatures (Tables 5–7).

The medians are also lower but follow the same pat-

tern of poorer estimates during summer months (July median of 0.86) as compared with winter (January median of 0.95). The geographical pattern of minimum temperature correlations (Figs. 7 and 8) is analogous to that described for the maximum temperatures. Here, the poorer estimates for January are concentrated in a region encompassing Arizona–New Mexico–Colorado with the effect much more pronounced in July.

The monthly distributions of the rmse for the minimum daily temperatures shown in Fig. 5 (Table 7) parallel the monthly pattern shown for the maximum temperatures. Although the seasonal pattern of lower/higher rmse for the summer/winter months may be the same, the range of values is typically larger. The patterns of values for the rmse statistic for minimum temperatures are consistent with those for maximum temperatures. It should be noted that although the spatial pattern of stations is similar, the interquartile range (IR) of the rmse for all minimum temperature stations has a larger amplitude relative to identical values computed for maximum temperature.

Examination shows the monthly distribution of the *t* statistic for observed versus estimated minimum daily temperatures is broadly consistent with the pattern for maximum temperatures, with the results described above (Fig. 6, Table 8). The IR is larger than that shown

TABLE 10. Summary statistics for the monthly distribution of the rmse between observed and estimated precipitation (in.).

	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	0.105	0.112	0.123	0.147	0.211	0.239	0.235	0.224	0.201	0.151	0.125	0.116
Median	0.060	0.063	0.084	0.100	0.144	0.185	0.188	0.173	0.129	0.093	0.077	0.066
Lower hinge	0.039	0.043	0.060	0.072	0.084	0.085	0.089	0.090	0.077	0.063	0.054	0.043
Number of stations						2962						

TABLE 11. Summary statistics for the monthly distribution of the mean daily percentage error (MDPE) between observed and estimated precipitation (%).

	Month											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Upper hinge	1.50	1.70	1.40	1.43	1.35	1.71	2.08	2.05	1.70	1.49	1.51	1.52
Median	1.07	1.19	0.943	0.954	0.871	1.14	1.45	1.47	1.25	1.09	1.07	1.03
Lower hinge	0.738	0.805	0.644	0.681	0.667	0.854	1.09	1.07	0.932	0.799	0.759	0.699
Number of stations						2962						

for maximum temperatures although median values are near zero.

As a result, a greater number of stations are found to be statistically significant at the 0.5% level. In the aggregate, the estimation of minimum daily temperatures is less effective than that shown for daily maximum temperatures.

c. Total precipitation

As with temperature, the efficacy of the best estimate for daily precipitation is assessed via the relationships between the observed and estimated daily values. The distribution of correlation coefficients, as an indicator of the quality of the estimate for each station, is summarized in Fig. 9 and Table 9.

As expected, the range of values of R is much greater than that shown by maximum–minimum temperatures, and overall values are lower although the seasonal pattern remains the same: larger IR and comparatively lower values during summer (July median of 0.72) as opposed to the transition seasons and winter (January median of 0.86). In general, the worst estimates are found in the mountainous regions of the western United States for both January and July (Figs. 10 and 11). As with maximum–minimum temperature, the spatial pattern of high–low correlations is much less pronounced in January than for July. The topographical diversity and the lower density of stations in this region result in poorer estimations. Conversely, the best estimates are found in the lower elevations of the data rich regions, for example, the western coasts and eastern Kansas–Nebraska. The accuracy of the daily precipitation estimate is dependent on the quality and quantity of the surrounding stations utilized to estimate a value at a particular site. The same is true for maximum–minimum temperatures but the determination of daily precipitation totals is much more sensitive to these factors, particularly with regard to the elevation of the site to be estimated.

Investigation of the rmse for observed versus estimated daily precipitation for all stations (Fig. 12, Table 10) yields results consistent with the seasonal pattern noted previously. The median rmse is less than 0.1 in. for all months except May–September, with peak values in June (0.18 in.) and July (0.15 in.). In addition, the IR is much larger for the months May–September. The location of stations that exhibit large/small rmse, relative to all other stations, is not a reliable indicator of

the absolute quality of a station's estimate, because of different climatic regimes. The pattern of large/small rmse is a reflection of the regional climate and is preserved in the observed/estimated differences. In other words, areas with climatically higher (lower) monthly rainfall would typically have higher (lower) rmse values. It was felt that a relative measure of precipitation estimation efficiency was needed. A statistic that would provide the kind of information that the t test did for the temperature field is described below.

The determination of whether the sample of estimated daily precipitation values is significantly different from observed totals is problematic, particularly if a simple t test is used. Precipitation distributions generally exhibit positive skewness (values tend to cluster about the median error rather than the mean) and thus do not lend themselves well to parametric tests designed to test mean differences. For this reason, a simple ratio test was employed to compare the observed versus estimated precipitation values and is described below.

At each station, for individual year/months the following is computed:

$$\left(\frac{\sum_{i=1}^{n} |E_i - O_i|}{N} \middle/ M\right) 100, \qquad (6)$$

where in a series of N days for a particular year/month E_i represents the ith estimate for the month and O_i the corresponding ith observation. The average absolute error for the month is then normalized by the total precipitation M for that month and expressed as a percentage. The result is a simple estimate of the mean daily absolute error standardized to reflect the secular variability of monthly precipitation totals for a wide variety of stations with differing seasonal cycles. The year/month ratios (expressed as percentage) are summarized for each station with the range of values for all precipitation reporting stations presented in Fig. 13 and Table 11. These values will be referred to as the mean daily percentage error (MDPE).

The monthly distribution of the range of MDPE among stations is consistent with that shown by the range of correlation coefficients (Fig. 9) and the seasonal distribution of rmse (Fig. 12). Although, as expected, median MDPE values are higher for summer months (July median of 1.5%) than winter (January median of 1.1%), the differences are small. Of note is the relative

similarity of the IR between months; this feature is not present for the seasonal range of correlation coefficients and the rmse. The lack of a pronounced seasonal cycle, that is, the parameters of the monthly frequency distributions are close to each other, suggests that the MDPE statistic is a more stable indicator of the efficiency of the estimate than that shown by the analysis of the correlation coefficient or the rmse.

9. Summary

This paper summarizes a set of procedures used to create serially complete daily temperature and precipitation datasets (1951–91) for the western United States. Determining target and estimator stations by scanning the quality of individual station records, reconciling metadata (including observation times and station locations), and categorizing observation times proved to be time consuming but necessary. Estimating the missing data values and cross validating the results proved to be relatively straightforward once preparatory work was completed.

Our results show that the efficacy of the estimation procedure and thus the reliability of the estimated missing values are dependent on a number of factors. For all three meteorological parameters the selection and quantity of surrounding stations are critically important to the results of the interpolations. We feel that the preselection of surrounding stations, based on their relationship with the station to be estimated, is an integral first step.

The quality of the estimates is strongly affected by seasonality—more so for daily precipitation as opposed to maximum–minimum daily temperatures. This effect is most severe during the summer months. Because of the changing spatial relationships among the temperature and precipitation fields, the derivation of the equations for estimation should be done separately for the season–month–day to be estimated. The quality of the daily estimates also has a spatial component. Stations at higher elevations are difficult to estimate accurately, in large part because of the topographical diversity of the surrounding stations leading to degradation of spatial coherence among stations.

In the aggregate, statistical tests of the resultant observed versus estimated time series show no systematic bias in the estimation procedures, particularly where the terrain and the density of surrounding stations is relatively uniform. In areas where the complexity of terrain (coastal, mountainous) dominates, the user should be aware of the limitations of the daily estimates. We feel that our methods produce the best possible estimates for all stations for a variety of conditions, but the error or bias at an individual site should be evaluated on a case-by-case basis. To assist the users in their own evaluation of the accuracy of individual estimates, the performance statistics for each individual station series are available as metadata with the serial dataset.

REFERENCES

Barrodale, I., and F. D. K. Roberts, 1973: An improved algorithm for discrete L1 approximation. *SIAM J. Numer. Anal.,* **10,** 839–848.

Bennet, R. J., R. P. Haining, and D. A. Griffith, 1984: The problem of missing data on spatial surfaces. *Ann. Assoc. Amer. Geogr.,* **74,** 138–156.

DeGaetano, A. T., K. L. Eggleston, and W. W. Knapp, 1993: A method to produce serially complete daily maximum and minimum temperature data for the Northeast. NRCC Research Publication RR 93-2, 9 pp. [Available from NRCC, Cornell University, Ithaca, NY 14853.]

Eischeid, J. K., C. B. Baker, T. R. Karl, and H. F. Diaz, 1995: The quality control of long-term climatological data using objective data analysis. *J. Appl. Meteor.,* **34,** 2787–2795.

Gandin, L. S., and R. L. Kagan, 1974: Construction of the system of heterogeneous data objective analysis based on the method of optimal interpolation and optimal agreement (in Russian). *Meteor. Gidrol.,* **5,** 3–10.

Huth, R., and I. Nemešová, 1995: Estimation of missing daily temperatures: Can a weather categorization improve its accuracy? *J. Climate,* **8,** 1901–1916.

Kemp, W. P., D. G. Burnell, D. O. Everson, and A. J. Thomson, 1983: Estimating missing daily maximum and minimum temperatures. *J. Climate Appl. Meteor.,* **22,** 1587–1593.

Owenby, J. R., and D. S. Ezell, 1992: Monthly station normals of temperature, precipitation, and heating and cooling degree days, 1961–90. *Climatography of the United States.* No. 81, National Climatic Data Center. [Available from NCDC, Asheville, NC 28801.]

Paulhus, J. L. H., and M. A. Kohler, 1952: Interpolation of missing precipitation records. *Mon. Wea. Rev.,* **80,** 129–133.

Reek, T. S., S. R. Doty, and T. W. Owen, 1992: A deterministic approach to validation of historical daily temperature and precipitation data from the cooperative network. *Bull. Amer. Meteor. Soc.,* **73,** 753–762.

Schlatter, T. W., 1975: Some experiments with a multivariate objective analysis scheme. *Mon. Wea. Rev.,* **103,** 246–257.

Stooksbury, D. E., C. D. Idso, and K. G. Hubbard, 1999: The effects of data gaps on the calculated monthly mean maximum and minimum temperatures in the continental United States: A spatial and temporal study. *J. Climate,* **12,** 1524–1533.

Thiebaux, H. J., and M. A. Pedder, 1987: *Spatial Objective Analysis with Applications in Atmospheric Science.* Academic Press, 299 pp.

Young, K. C., 1992: A three-way model for interpolating monthly precipitation values. *Mon. Wea. Rev.,* **120,** 2561–2569.