INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **21**: 1305–1319 (2001)
DOI: 10.1002/joc.679

ESTIMATING MISSING DAILY TEMPERATURE EXTREMES USING AN OPTIMIZED REGRESSION APPROACH

ROBERT J. ALLEN and ARTHUR T. DeGAETANO*
Northeast Regional Climate Center, Cornell University, Ithaca, NY, USA

Received 11 March 2000
Revised 2 April 2001
Accepted 7 April 2001

ABSTRACT

A variation of a least squares regression approach to estimate missing daily maximum and minimum temperatures is developed and evaluated, specifically for temperature extremes. The method focuses on obtaining accurate estimates of annual exceedence counts (e.g. the number of days greater than or equal to the 90th percentile of daily maximum temperatures), as well as counts of consecutive exceedences, while limiting the estimation error associated with each individual value.

The performance of this method is compared with that of two existing methods developed for the entire temperature distribution. In these existing methods, temperature estimates are based on data from neighbouring stations using either regression or temperature departure-based approaches.

Evaluation of our approach using cold minimum and warm maximum temperatures shows that the median percentage of correctly identified exceedence counts is 97% and the median percentage of correctly identified consecutive exceedence counts is 98%. The other existing methods tend to underestimate both single and consecutive exceedence counts. Using these procedures, the estimated exceedence counts are generally less than 80% of those that actually occurred.

Despite the fact that our method is tuned to estimate exceedence counts, the estimation accuracy of individual daily maximum or minimum temperatures is similar to that of the other estimation procedures. The median absolute error (MAE) using all temperatures greater than or equal to the 90th percentile (T_{90}) $-$ 1.1°C for ten climatically diverse stations is 1.28°C for our method, while the other methods give MAEs of 1.27 and 1.17°C. In terms of median error, however, the tendency for underprediction by the existing methods is pronounced with -0.77 and -0.61°C biases. Our optimized method is relatively unbiased as the resulting mean error is -0.12°C. Copyright © 2001 Royal Meteorological Society.

KEY WORDS: median absolute error; optimized regression analysis; temperature extreme

1. INTRODUCTION

The ubiquitous nature of missing daily temperature observations is a problem in many climatological applications. Researchers have relied on a variety of techniques to estimate missing daily temperature data depending on the region, time of year, or spatial distribution of neighbouring stations. Collectively, these methods can be organized into three basic categories: (i) within-station; (ii) between-station; and (iii) regression-based.

Within-station methods make use of temperatures recorded on the previous and following days to estimate a missing observation. For example, to estimate a missing maximum temperature on 1 July, the maximum temperatures from 30 June to 2 July could be averaged (Oliver, 1973). Averages can also be computed using more than 1 day on either side of the missing day.

* Correspondence to: Northeast Regional Climate Center, Cornell University, 1115 Bradfield Hall, Ithaca, NY 14853, USA; e-mail: atd2@cornell.edu

Between-station techniques use temperatures at neighbouring stations to estimate missing data values. Kemp *et al.* (1983) describe such a method that is based on the assumption that the difference between daily temperatures at adjoining stations is equal to the difference between the monthly average temperatures at those stations. DeGaetano *et al.* (1995) describe another between-station approach that is based on the assumption that temperatures at nearby stations will depart from their respective daily normals by similar amounts.

With the regression-based approach, data from one or more neighbouring stations is used to develop a regression equation, which is subsequently used to estimate missing temperatures. Using several different regression techniques, Kemp *et al.* (1983) have shown that this type of approach tends to produce more accurate estimates than both within-station and between-station methods. Similarly, Eischeid *et al.* (1995) show that multiple regression, using the least absolute deviation (LAD) criteria, is superior to four between-station methods for estimating missing monthly temperature averages. Although based on monthly data, Eischeid (unpublished document) suggests that the same approach can be applied to daily values.

In their comparisons, Eischeid *et al.* (1995) include the single-best estimator method which is analogous to substituting the temperature at a nearby station for the missing value. The surrogate temperature is obtained from the neighbouring station that exhibits the highest correlation with the missing value station. The normal ratio method is similar, but uses a weighted average of surrounding stations. Weights are calculated according to the correlation coefficient (r) between the daily time series from the target (missing value) station and each neighbouring station, as well as the number of temperatures used to compute r. The inverse distance method is a distance weighted, area average estimate. The assumption of this method is that temperatures at the target station (Y) and neighbouring stations (X_i, where i is the ith neighbouring station) are related according to the geographic distance between the stations. Optimal interpolation is a spatial interpolation technique that assigns weights to the observed difference values at the selected nearby stations. The weights depend on the spatial autocorrelations between the x_i (temperatures at station X_i) and are mathematically modelled as a function of the distance separating the surrounding stations and station Y.

Each of these previous methods has been developed and refined to minimize the bias and mean absolute errors associated with the overall distribution of daily maximum or minimum temperatures. Little attention has been given to the relationship between estimation error and temperature magnitude. In fact, the achievement of good, overall error statistics is generally at the expense of extremes. Specifically, these methods have a strong proclivity to underestimate temperatures in the right tail of the overall distribution and to overestimate temperatures in the left tail. With the analysis of extreme temperature series becoming an important issue in terms of global warming detection and impact assessment (IPCC, 1996), accurate estimates of missing extreme temperatures have correspondingly become more important.

Regression-based methods tend to underestimate warm extreme temperatures and to overestimate cold extremes (i.e. estimates are warmer than observed) because of the disproportion of moderate temperatures compared with more extreme observations. Many more moderate temperatures will go into the development of the regression equation, in effect maximizing its ability to predict these temperatures, while sacrificing its ability to estimate the more extreme values.

Similarly, between-station methods are also apt to underestimate warm extreme temperatures because of this disproportionality. Given the distribution of daily temperatures, there is a higher probability that x_i will be less extreme than the corresponding temperature at station Y. For example, assume that y is very extreme (e.g. $\geq$ 99th percentile). Although x_i will also be extreme, it is more likely that x_i will be lower (rather than higher) than y.

Of course, within-station methods underpredict warm extreme temperatures as well. Clearly averaging temperatures from successive days precludes an estimate that is greater than the temperatures of the non-missing days.

This paper describes a variation of a least squares regression approach to estimate extreme daily maximum and minimum temperature observations. The method focuses on obtaining accurate estimates of annual exceedence counts (e.g. the number of days $\geq T_{90}$ of daily maximum temperatures), while

sacrificing only a small degree of accuracy in the estimation of the individual daily temperatures. The method also focuses on maximizing the accuracy of the estimated number of consecutive exceedence counts, such as the number of 2-, 3- or 4-day runs $\geq T_{90}$ of maximum temperatures per year.

2. EXISTING ESTIMATION METHODS

Estimation procedures representing the regression-based and between-station approaches are used as a basis for comparison in this paper. Eischeid *et al.* (1995) describe a multiple regression approach using LAD. DeGaetano *et al.* (1995) use a modified version of Steurer's (1985) standard departure method. A brief description of these methods is given in the following subsections.

2.1. Multiple regression, least absolute deviations criteria

Eischeid *et al.* (1995) use LAD to estimate monthly mean temperatures. In addition, Eischeid (unpublished document) suggests that the same approach can be applied to daily values. Thus, we have adopted the LAD technique to estimate daily maximum or minimum extreme temperatures. Temperatures from one or more neighbouring stations are used as predictors in a regression equation that is used to predict missing temperatures for station Y. Neighbouring stations are selected according to their geographic distance from the target station (station Y) and ranked by their correlation coefficient, r. The neighbouring stations with the largest $r \geq 0.35$ are used in the estimation procedure. Between one and four stations are included in the regression. Separate correlation coefficients are calculated for individual months, as are the regression coefficients. This takes into account the fact that one station may be better suited to estimate a missing day in July, whereas a different station may be more appropriate in December.

Multiple regression using the least absolute deviations criteria is a robust version of the general linear least squares estimation. The principal advantage of LAD is its resistance to outliers. Regression coefficients are estimated so as to minimize the sum of the absolute deviations of y from the values predicted by the model. Mathematically, this is represented by:

$$\sum \left| \sum x_{ij} b_j - y_j \right|, \tag{1}$$

where x, $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ denote a set of n temperatures at m surrounding stations. y, $j = 1, 2, \ldots, n$ denote the corresponding observations at station Y and b is the regression coefficient.

Eischeid *et al.* (1995) estimate monthly mean temperatures using 3467 stations from the Global Historical Climate Network (GHCN) (Vose *et al.*, 1992) and use these estimates as a means of quality control. The procedure uses the five aforementioned interpolation methods to estimate each monthly time series. Comparisons among the techniques are based on the correlation coefficient between the true monthly time series at station Y and each of the estimated monthly time series. The method with the largest r is chosen as the best estimator, with LAD the best estimator for the majority of the records. Estimates were found to be relatively unbiased, with mean errors of 0.002 and 0.000°C for January and July, respectively. The standard deviation of mean error for these months was 0.869 and 0.697°C, respectively.

Conventional regression-based procedures, like LAD, have a propensity to underestimate warm extreme temperatures because the abundance of less extreme temperatures biases the regression toward these values. Cold extreme temperatures are overestimated (i.e. estimates are warmer than observed) for an analogous reason. The computation of separate regressions for each month in Eischeid *et al.*'s (1995) procedure fails to compensate for this shortcoming. Furthermore, compared with least squares regression, LAD regression is less sensitive to outliers. Often, these outliers are similar to the extreme temperatures that require estimation.

2.2. DeGaetano's temperature departure procedure

DeGaetano *et al.* (1995) use a between-station method that assumes that on any given day, nearby stations with similar observation times will experience temperatures that deviate from their daily normal by a similar amount.

Briefly, standard departures, Z_i are calculated for each daily temperature according to:

$$Z_i = \frac{x_i - \bar{x}_i}{S_i}, \tag{2}$$

where i is any nearby station and x is this station's corresponding daily temperature. $\bar{x}_i$ is estimated as the smoothed daily temperature normal following a procedure developed by Epstein (1991). In Epstein's procedure, the daily climatology of maximum or minimum temperatures is estimated by the sum of harmonic components based on the 12 monthly means. S_i is estimated as the monthly standard deviation. A daily average standard departure Z_{avg} is computed using

$$Z_{\text{avg}} = (1/N) * \sum Z_i, \tag{3}$$

where N is the number of stations that have valid daily temperatures. Estimates of missing values are calculated using the equation:

$$y_j = Z_{\text{avg}} S_j + \bar{y}_j, \tag{4}$$

where j is the station to be estimated.

To test the validity of their approach, DeGaetano *et al.* (1995) estimated non-missing daily temperatures at 12 sites in the northeastern US. In general, 75% of the estimates for daily maximum and minimum temperature were within 1.7°C of the actual value. Median absolute errors (MAEs) for minimum temperatures tended to be greater than those associated with maximum temperatures. MAEs for minimum temperatures were approximately 1.0°C, whereas those for maximum temperatures were near 0.5°C.

Methods using temperature departures tend to underestimate warm extreme temperatures for reasons similar to those of regression-based approaches. Examining Equation (4) for maximum temperatures, $\bar{y}_j$ will be less than the actual temperature (y_j) because an extreme is being estimated. This is not a problem if the Z_{avg} on a particular day is similar to the corresponding standard departure (Z) at station Y. However, x_i will most likely fall below their corresponding threshold temperature, resulting in a relatively low Z_{avg} compared with Z.

The reason for the relatively low Z_{avg} is twofold, one an artifact of the method and the other a product of the nature of extreme temperatures. When estimating extreme temperatures at station Y, only $y \geq T_{90} - 1.1$°C are estimated. This allows temperatures that are actually $< T_{90}$ to be overestimated and counted as exceedences, balancing the potential underestimation of temperatures that are actually $> T_{90}$. Thus, y is restricted to be above a threshold. Neighbouring stations, however, do not have this same restriction and they could very well be cooler on a day when y is extreme. As DeGaetano *et al.* (1995) point out, this could occur when a cold front passes through station Y's neighbouring stations, but has not yet passed through station Y. On such a day, y could be quite extreme, whereas x_i will be considerably cooler. The reverse, however, will never occur because y is restricted to be above a threshold. Thus, the method screens y such that it is extreme, without a corresponding screening for x_i.

Since there are many more moderately extreme temperatures than very extreme temperatures, the probability is greater that an extreme y will be paired with a less extreme x_i. This is illustrated in Figure 1, which contains histograms of differences between the standard departures for Decatur, IL (Z) and the average neighbouring station standard departure (Z_{avg}). For all days ≥ 30.6°C, Figure 1(a) shows that approximately 70% of the time, Z is larger than Z_{avg}. For the most extreme days, Figure 1(d) shows that this is still true as Z is larger than Z_{avg} almost 80% of the time. Thus, because there is a greater probability for x_i to be less extreme than y, Z_{avg} will often times be less than Z, resulting in the underprediction of extreme y. This occurs regardless of the station density used to compute Z_{avg}.

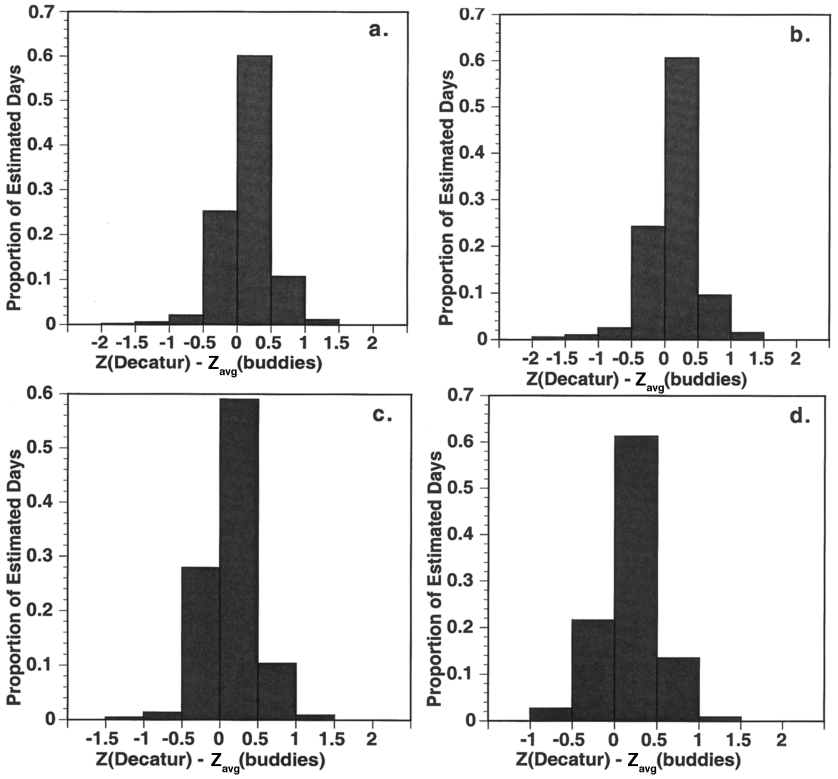

Figure 1. Histograms of $Z - Z_{\text{avg}}$ based on daily maximum temperature (T) at Decatur, IL, for cases where (a) $T \geq 30.6$°C, (b) 30.6°C $\geq T < 32.2$°C, (c) 32.2°C $\leq T \leq 34.4$°C and (d) $T > 34.4$°C

3. METHODOLOGY

Our methodology is similar to that of Eischeid *et al.* (1995) in that we use neighbouring stations to form a regression equation. Because our desire is to obtain accurate estimates of extreme temperatures and annual exceedence counts, we have incorporated numerous modifications to this base procedure. As will be discussed in more detail below, pairing each x_i with the corresponding daily maximum or minimum y results in a regression that is biased toward the less extreme temperatures. This occurs since many more observations are associated with less extreme temperatures, while only a few observations characterize the most extreme values. To alleviate this problem, a binning procedure is implemented that produces a relatively unbiased regression equation regardless of the magnitude of temperature, but one that is frequently still unable to accurately classify a day as exceeding (or not exceeding) a given threshold.

Cross-validation trials were used to select a set of neighbouring stations that minimize the mean absolute error associated with this binned regression. Each station-specific regression equation was then optimized to more accurately classify days as exceeding a given threshold, at some cost to accurately estimating the day's actual temperature. Using these optimized regressions, cross-validation trials were again conducted to select a subset of those stations that minimized the MAE. This final set of neighbouring stations is best able to accurately predict the number of annual extreme exceedences, while limiting the estimation error associated with each individual value.

3.1. Initial station selection

Some methods, such as Steurer's (1985), rely on station selection using broad and somewhat arbitrary climate or political boundaries. Others, such as Eischeid *et al.*'s (1995), use the nearest available station(s)

to reconstruct missing temperatures. We adopt this latter method to maximize the probability that all stations experience the same synoptic conditions on any given day. DeGaetano *et al.* (1995) have shown that using a distance criterion versus climate boundaries significantly reduces the overall range of errors.

Each station selected as a potential predictor must also have an observation time category identical to that of the station to be estimated. Using the daily US Historical Climatology Network (USHCN) (Easterling *et al.*, 1999) and its associated station history file, the daily temperature series at all stations were divided into periods that are free of observation category changes (as well as station relocations). Observation hours were grouped into categories based on Karl and Williams (1987), with 01:00–11:00 h local time (LT) defined as morning observations; 12:00–20:00 h LT as afternoon; and 21:00–24:00 h LT as midnight. Missing or unknown observation times were determined based on the inferred observation times given by DeGaetano (2000). Furthermore, actual observation times not supported by the inferred observation times were changed to reflect the inferred observation class when the disagreement persisted for more than 3 consecutive years and the average interdiurnal temperature difference exceeded 2.0°C.

Some final restrictions on the selection of neighbouring stations involve making sure there is adequate extreme data in common between a potential predictor station and station Y. A minimum of 100 temperature values $\geq T_{90} - 1.1$°C must be shared by the neighbour and station Y.

Temperatures $1.1\text{°C} < T_{90}$ are included to enable the possibility of overestimating a temperature slightly less than T_{90}, offsetting the underprediction of temperatures a few degrees greater than T_{90}. Furthermore, to guarantee that neighbouring stations truly reflect station Y's most extreme temperatures, each nearby station must have at least one non-missing temperature passing quality control screening on a day when y is 'extreme'. Extreme can be defined in a number of ways. The 95th percentile (T_{95}) is probably too lenient since $y \geq T_{95}$ comprise a good portion of the temperatures we wish to accurately estimate. Furthermore, use of the 99th percentile (T_{99}) is too stringent because of the possibility of eliminating nearby stations which otherwise are highly correlated with extreme y. Since cross-validation will be used as the primary measure of the ability of a neighbouring station to predict y, it would be premature to eliminate too many potential predictors. By defining 'extreme' as 75% of the range between T_{90} and the highest observed y, potential predictor stations represent the most extreme y without prematurely eliminating too many neighbouring sites. Thus, this definition was adopted.

3.2. Single least squared regression development

To estimate missing daily extreme temperatures, data from the closest 15 neighbouring stations that satisfy the above criteria serve as an initial pool of predictors. Given the station density of the USHCN, selecting > 15 stations represents an unrealistically large distance between the predictor and station Y.

The disparity between the number of temperatures near T_{90} for warm temperatures (or 10th percentile for cold temperatures) and occurrences of extreme temperatures is illustrated in Figure 2, which shows a scatter plot of warm extreme temperatures (≥ 31.7°C) for Fairhope, AL (Y) and the corresponding temperatures at a neighbouring station, Covington, LA (X_i). The frequency of occurrence for each temperature pair is labelled. Clearly, there is an abundance of temperatures near T_{90} and a paucity of values at higher thresholds. The influence of this discrepancy on the regression is also shown in Figure 2, as the black regression line passes well below the cloud of points representing the most extreme temperatures.

The regression's error based on all of the data points is satisfactory at 0.67 and 0.01°C for the mean absolute error and mean error, respectively. Examining the error for different extreme thresholds, however, reveals that the performance of the regression is clearly biased. Days when $y \leq T_{90}$ have a positive mean error of 0.44°C, while $y \geq T_{95}$ have a bias of -0.46°C. Every $y \geq T_{99}$ is underpredicted (Figure 2).

The deficiency of this approach is further illustrated by examining the ratio of the number of estimated days greater than or equal to (less than or equal to for cold temperatures) the extreme threshold (E_x) to the actual number of days greater than or equal to this extreme threshold (A_x):

$$\text{pct}_x = [(E_x/A_x) * 100], \tag{5}$$

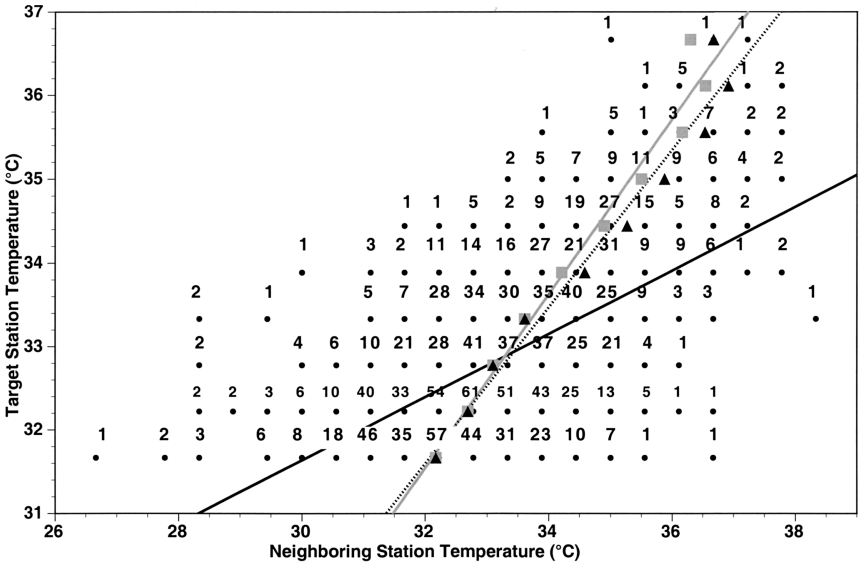

Figure 2. Scatter plot and regression of extreme warm maximum temperatures for Fairhope, AL (Y) versus Covington, LA (X_i) for all points (black circles and solid line); the median temperature at Covington (grey squares and line); and the median temperature at Covington plus an optimization factor of 0.37°C (black triangles and dotted line)

where x is the specified threshold. For days $\geq T_{90}$, pct_{90} equals 98%. However, at the more extreme percentiles, the regression performs poorly. pct_{95} is only 26.6% and pct_{99} equals 0%. This means that none of the actual days $\geq T_{99}$ at Fairhope are estimated as surpassing this 35°C threshold.

To avoid this unequal weighting, a binning procedure is employed. For each predictand temperature (y), the neighbouring station's median temperature (x_i) is computed. These temperature pairs (grey squares in Figure 2) then form the basis of a regression that gives equal weight to all observed temperatures greater than or equal to the threshold.

In Figure 2, the modified regression (solid grey) for Fairhope, AL, is shown. As can be seen from the error statistics (as well as visually), the regression is no longer biased toward the relatively cooler temperatures. For $y \leq T_{90}$, the mean error is 0.01°C and pct_{90} remains unchanged at 98%. There is a considerable improvement, however, in the regression estimates for days when $y \geq T_{95}$ and T_{99}. The mean error has been reduced to -0.08 and 0.16°C, respectively. Unfortunately, pct_{95} and pct_{99} are now over predicted, increasing to 128 and 165%, respectively. This is a problem that will be addressed by using an optimization procedure. Nonetheless, compared with the regression based on all of the extreme data, using the median neighbouring temperature produces a more unbiased regression.

3.3. Cross-validation of initial stations

After calculating the binned regression for a neighbouring station, the ability of the regression to estimate extreme y is evaluated using cross-validation.

The cross-validation errors were divided into three groups depending on the value of the nth withheld y. When estimating a warm temperature extreme, errors corresponding to $y <$ 92nd percentile (T_{92}) form group 1; errors for $T_{92} \leq y \leq$ 97th percentile (T_{97}) define group 2; and errors for $y \geq T_{97}$ are included in group 3. The average errors of each of these three groups are then calculated and a single mean absolute error and mean error are calculated. By keeping track of errors in this manner, the abundance of temperatures near the threshold do not bias the error statistics towards the less extreme thresholds.

After using this procedure on all initial stations, the stations are ranked according to their mean absolute error. Stations with a mean error $\geq$ 0.28°C or $\leq$ 0.28°C ($\pm$ 0.5°F) are discarded because they tend to give biased estimates for station Y and accuracy to the 0.56°C (1°F) is desirable to assure

satisfactory exceedence counts. In cases where the 0.28°C restriction omits all neighbouring stations, the restriction is relaxed in multiples of 0.06°C until at least one neighbouring station satisfies the criterion.

The median estimate given by the two stations with the lowest mean absolute error is used to obtain a new estimate for all non-missing extreme days at station Y. A new mean absolute error is then computed based on these estimates and this new MAE is compared with the MAE of the single best station. If the error is lower, the third best neighbouring station is incorporated into the procedure and new error statistics are calculated. Station are added to the procedure, one at a time, until the addition of another station increases the mean absolute error of the smaller set of stations.

3.4. Optimization of final stations

Although the above set of stations minimizes the MAE, these stations may still not be optimal for classifying a missing day as exceeding (or not exceeding) an extreme temperature threshold. In particular, most stations tend to overestimate the percentage of days exceeding T_{99} (or T_{01}). One reason for this problem relates to the greater quantity of temperatures near the less extreme temperatures versus the relative paucity of the more extreme temperatures (see Figure 2). Thus, for warm temperatures, the probability that the unbiased (in terms of mean absolute error) regression equation overestimates days that are actually less than the threshold is higher than the chance of underestimating days that are greater than the temperature extreme. For example, consider the calculation of pct_{95} using an unbiased regression equation with an MAE of 1.1°C. Assuming T_{95} at this station is 35°C, on average temperatures in the range of 33.9–36.1°C will be estimated as 35°C, given the formulation of the binned regression. In reality, however, there are many more days with temperatures between 33.9 and 35°C than there are between 35 and 36.1°C, and thus the chance of overestimating the extreme is greater. For warm temperatures, this leads to large values of pct_{95} and especially pct_{99}; for cold temperatures, this leads to large values of pct_{05} and pct_{01}.

To compensate for this problem, the regression equations are optimized by substituting percentiles other than the median in their development. This changes the slope and intercept of the regression such that estimates for the less extreme temperatures are kept fairly constant, while estimates of the more extreme temperatures are adjusted downward for warm temperature estimation, or upward for cold temperature estimation. This is illustrated by the dotted line in Figure 2, which shows that for each Fairhope temperature $(y) \geq T_{97}$ (33.9°C), the corresponding median temperature at Covington (x_i) has been increased (shifted to the right) by 0.37°C. All of the other temperature pairs remain unchanged. The resulting regression equation has a less steep slope and higher intercept than the initial (grey) regression. This new optimized regression is better able to classify a temperature as exceeding a given extreme threshold.

The 0.37°C increment was arrived at through the optimization of the percentage of days correctly identified as exceeding the 90th, 95th and 99th percentiles. Mathematically,

$$(\mathrm{pct}_{90} + \mathrm{pct}_{95} + \mathrm{pct}_{99})/3 \tag{6}$$

is optimized. Optimization begins with an unrealistically high adjustment of 3.3°C (Table I). To pivot the regression, the adjustment is added to each of the neighbouring station's median temperatures when $y \geq T_{97}$. Equation (6) is then recalculated and the adjustment re-evaluated (by halving or doubling), depending on whether Equation (6) yields an average greater or less than 100%. If Equation (6) is $> 100\%$, the current regression tends to overestimate the number of extreme days, so a larger increment is warranted. If, however, Equation (6) is $< 100\%$, the current regression tends to underestimate the number of extreme days, so a smaller increment is necessary. The value of the final increment is the one that produces a regression equation that minimizes the difference between Equation (6) and 100%. Table I illustrates this optimization procedure. The initial value (without optimization) of Equation (6) is quite large at 133%, most of this attributed to the very high values of pct_{95} (138%) and pct_{99} (164%). The regression is pivoted about T_{97} by adding 3.3°C to each of the neighbouring station's median temperature values when the corresponding temperature at Fairhope is $\geq T_{97}$, or 33.9°C. As expected, this is an

Table I. Illustration of the adjustment procedure used to pivot the regression based on the percent of correctly identified exceedences (pct_{90}, pct_{95}, pct_{99}) and the average percent (Equation (6)) of correctly identified exceedences

Iteration	Adjustment (°C)	pct_{90} (%)	pct_{95} (%)	pct_{99} (%)	Average (%)
1	0.00	98	138	164	133
2	3.33	76	26	10	37
3	1.67	85	120	130	112
4	2.50	85	115	85	95
5	2.08	95	115	120	110
6	2.29	95	115	105	105
7	2.24	95	112	100	102
8	2.18	95	112	85	97
9	2.23	95	112	105	104
10	2.22	95	105	98	99

over-adjustment with pct_{95} falling to 26% and pct_{99} to 10%. The initial adjustment is halved to 1.67°C and is added to each of neighbouring station's temperatures when $y \geq 33.9$°C. This overcompensates for the initial 3.3°C adjustment as pct_{95} and pct_{99} once again exceed 100%. The intermediate value of 2.5°C is used for the next iteration. Continuing in this manner, pct_{99} falls between 90 and 110% with an increment of 2.29°C. pct_{95}, however, does not fall between 90 and 110%, so we continue to optimize at finer increments until this occurs. After four additional iterations, the regression is finally optimized when Equation (6) assumes a value of 99.3%.

Our choice of T_{97} was based upon a set of trials using larger and smaller percentiles. Various percentiles between the 95th and 99th were examined. The percentage of correctly classified days exceeding the 90th, 95th and 99th percentiles for the different approaches are shown in Figure 3. Using the T_{99} as a pivot point to optimize pct_{99} within 10% of 100% (approach 1) results in relatively unbiased values of pct_{90}, pct_{95} and pct_{99}, but shows moderately large spreads. pct_{95} is skewed upwards with a median of 104% and upper decile near 120%. Similarly, pct_{99} has a relatively large spread that is skewed downwards. Here a median of 95% and a lower decile of 78% result. Unfortunately, continuing to pivot the regression about the 99th percentile in an attempt to bring down pct_{95} fails to produce the desired result. In fact, pct_{99} often becomes more biased before pct_{95} is improved.

Intuitively, choosing a lower pivot threshold, such as T_{95}, and optimizing pct_{95} (instead of pct_{99}) within 10% error limit (approach 2) seems feasible. Although this improves the distribution of pct_{95} somewhat, the improvement is at the expense of pct_{99} which now, although unbiased, exhibits a high spread (Figure 3). pct_{90} also becomes skewed toward underestimation.

Our final approach uses T_{97} as an intermediate pivot point and optimizes pct_{99} to within a 10% error limit and then continues optimizing pct_{95} at finer increments until it, too, falls within 10% or pct_{99} falls outside its 10% error limit (approach 3). Using T_{97}, the median of pct_{90} has decreased slightly from 98% using approach 1 to 97%. However, the spread of pct_{95} is now more symmetrical and the median has improved from 104% using approach 1 to 100%. Furthermore, pct_{99} has also improved. The median has increased to 96% from 95% (approach 1) and the spread has decreased as well.

This final, optimized regression is no longer the best (in terms of minimizing the squared errors) for all temperatures $\geq T_{90} - 1.1$°C. This is true because we have changed the slope and intercept of the original binned (and least squares) regression (which used all data $\geq T_{90} - 1.1$°C). In most cases, this will introduce a slight negative bias in the final error statistics because we are intentionally underestimating the more extreme temperatures. This offsets the fact that the abundance of temperatures less than the extreme threshold results in more temperatures being incorrectly classified as exceeding this threshold than vice versa. Thus, the final regression is able to more accurately classify a day as exceeding (or not exceeding) a given threshold, in particular the more extreme thresholds.

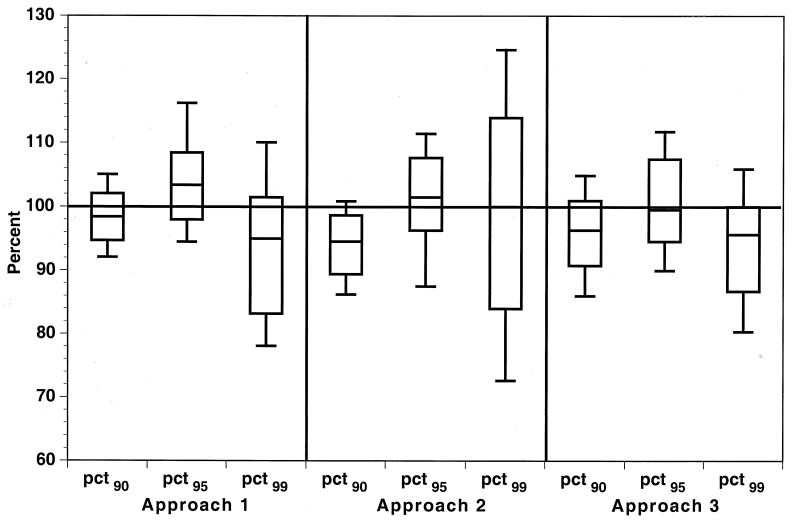

Figure 3. Box plots of pct_{90}, pct_{95} and pct_{99} using warm maximum temperatures at 20 climatically diverse stations and pivoting the regression at T_{99} (approach 1), T_{95} (approach 2) and T_{97} (approach 3)

After optimizing each of the final neighbouring stations separately, the two that produced the lowest mean absolute error prior to optimization are combined as they were before, but the optimized regression equations are used and the result of Equation (6) is evaluated. This process is repeated for the remaining final stations, each time recalculating Equation (6). The final predictor stations used to estimate a given day are those that bring Equation (6) closest to 100%. When more than one predictor station is used, the median predicted temperature is used as the final estimate for a given day.

The optimization routine works in a similar manner for cold extreme temperatures. The only differences include substituting pct_{10}, pct_{05} and pct_{01} into Equation (6) and subtracting the increment from station Y's median temperature when $y \leq 3$rd (instead of the 97th) percentile.

4. RESULTS AND COMPARISONS WITH OTHER METHODS

To assess the accuracy of our method, extreme temperatures were estimated for non-missing days at ten climatically diverse stations across the US. Comparisons are made with the methods of DeGaetano *et al.* (1995) and Eischeid *et al.* (1995). In all cases, these comparisons were based on total exceedences accumulated over homogenous subperiods at each station.

Figure 4(a) is a plot of A_{90} versus E_{90} for warm maximum temperatures. The results clearly show that both of the previous methods tend to underestimate extreme temperatures, as most points fall below the line $x = y$. Our optimized regression method, on the other hand, produces unbiased results with most points falling on, or near, the 1:1 line. The median of pct_{90} for our method is 97%, compared with 79 and 74% for the other methods.

Similarly, Figure 4(b) and (c) shows further deterioration of both the non-optimized methods for temperatures $\geq T_{95}$ and T_{99}, respectively. These previous methods severely underpredict these temperatures at nearly every station, with the median of pct_{99} equal to 54% for both. The optimized method, however, is nearly unbiased with the median of pct_{99} equal to 96%.

Figure 5 illustrates the performance of the three methods using runs of extreme temperatures; namely the number of consecutive days $\geq T_{90}$. Figure 5(a) shows that the prior methods tend to underestimate the number of two-day runs $\geq T_{90}$, with median pct_{90} values of 81 and 76%, respectively. The optimized method produces unbiased results, with most of the values falling on or near the line $x = y$ and giving a median pct_{90} of 98%. Figure 5(b) and (c) shows that this underestimation continues for 3 and 4

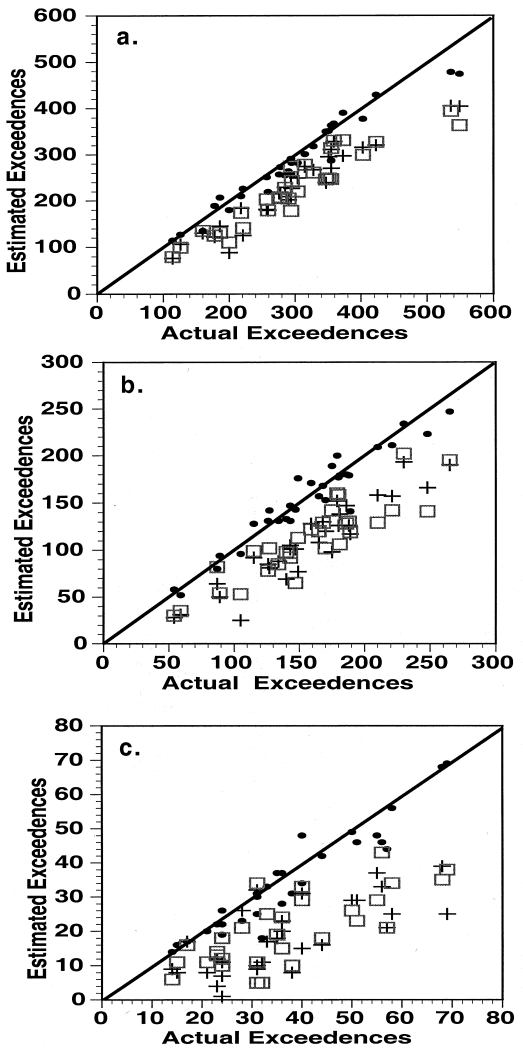

Figure 4. Scatter plots of actual versus estimated (a) T_{90}, (b) T_{95} and (c) T_{99} exceedences for maximum temperatures using the methods of Eischeid *et al.* (boxes); DeGaetano *et al.* (black crosses) and our optimized method (black circles)

consecutive days $\geq T_{90}$, with median values ranging from 74 to 83%, while the optimized median values are both 97%.

The estimation of cold minimum temperatures (Figures 6 and 7) produces similar results. The existing procedures underestimate both single and consecutive exceedence counts of the 10th, 5th and 1st percentiles. The optimized method, on the other hand, produces much better results. The median value for pct_{10} is 97%; 105% for pct_{05}; and 95% for pct_{01}. The corresponding values for Eischeid *et al.*'s (1995) method are 85, 83 and 79%. For the method of DeGaetano *et al.* (1995), the percentages are 82, 85 and 76%.

There are a few points in Figure 6(b) that indicate overestimation ($> 10\%$ error) of pct_{05}. For cold minimum temperatures, this means that the regression had a tendency to underestimate a given day, resulting in the prediction of more cold extreme temperatures than were actually observed. These points correspond to data from Presque Isle, ME, and values near the beginning of the 20th century for Greensboro, AL. In these cases, the neighbouring stations were limited and unable to be satisfactorily optimized, which resulted in the relatively high errors.

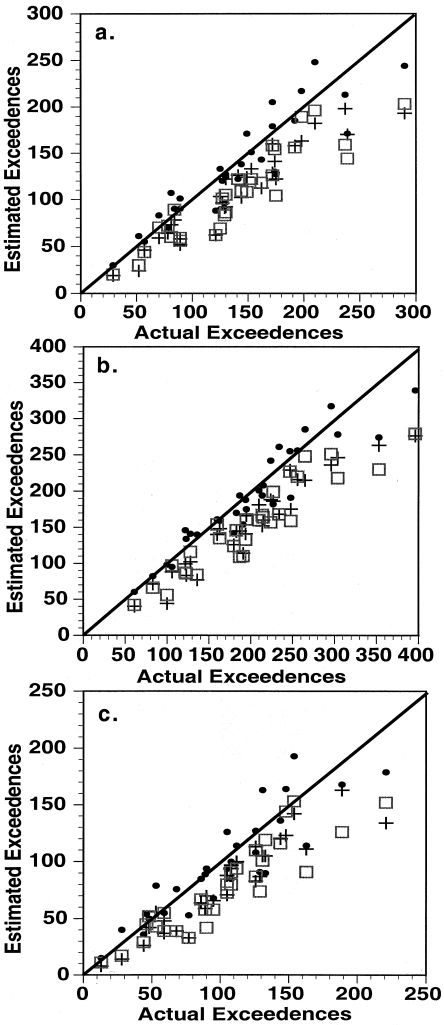

Figure 5. Scatter plots of actual versus estimated occurrences of (a) 2-, (b) 3- and (c) 4-day runs of maximum temperatures exceeding T_{90} using the methods of Eischeid *et al.* (boxes), DeGaetano *et al.* (black crosses) and our optimized method (black circles)

In general, cold minimum temperatures are the most difficult to estimate because of the strong effect of microclimatic variations between neighbouring stations. For example, if nearby stations are located in a small valley while station Y is situated outside the depression, radiational cooling and cold air drainage will significantly effect cold minimum x_i. Furthermore, because the range of temperatures between the 5th and 1st percentiles (T_{05} and T_{01}, respectively) is greater than the corresponding $T_{95}-T_{99}$ range for warm maximum temperatures, there tends to be more days with temperatures a few degrees greater than T_{05} and less days with temperatures a few degrees less than T_{05}. This results in an increased probability (compared with warm extremes) to overestimate the number of days $\leq T_{05}$.

Despite the fact that the binned least squares regression methodology has been tuned to estimate exceedence counts, individual temperature extremes are estimated with accuracy very similar to that of the other estimation procedures. This is illustrated in Figure 8, which contains box plots of mean errors for the three methods using ten climatologically diverse stations. The MAE using all temperatures $\geq T_{90} - 1.1$°C (Figure 8(b)) is 1.28°C for our method, 1.27°C for DeGaetano *et al.* (1995) and 1.17°C for Eischeid *et al.* (1995). These previous methods, however, are biased toward underestimation as the mean error (Figure 8(a)) for the method of DeGaetano *et al.* (1995) is -0.77 and

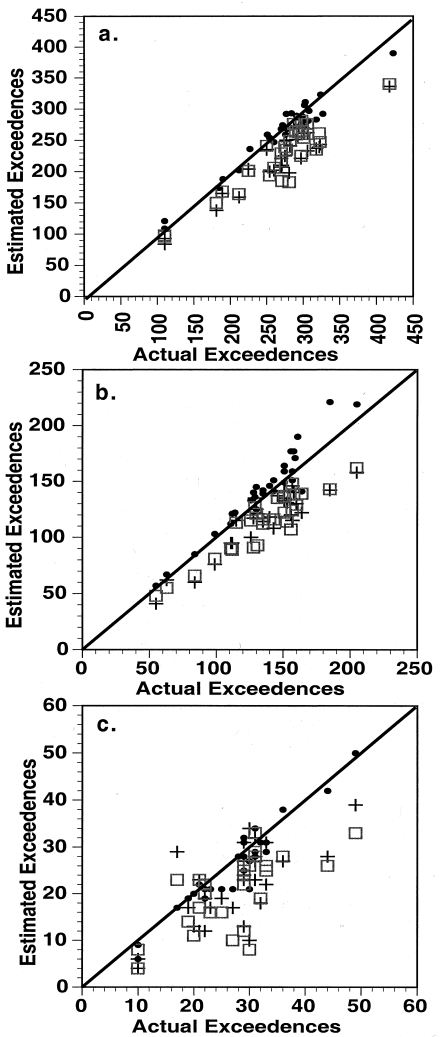

Figure 6. As in Figure 4 but for non-exceedences of (a) T_{10}, (b) T_{05} and (c) T_{01} for minimum temperatures

-0.61°C for Eischeid *et al.*'s (1995) method. The optimized method is less biased with a mean error of -0.12°C.

As the temperature to be estimated becomes more extreme (i.e. $> T_{99}$), the day-to-day accuracy of our method is higher than DeGaetano *et al.*'s (1995) method and approximately equal to the method of Eischeid *et al.* (1995). This is shown in Figure 8(d), where the MAE for the optimized method is 1.19°C, 1.33°C for DeGaetano *et al.* (1995) and 1.17°C for Eischeid *et al.* (1995). For the very extreme temperatures, the binned least squares methodology is once again not as biased as the other two methods. Figure 8(c) shows that the mean error for our method is -0.54°C, -1.17°C for DeGaetano *et al.* (1995) and -0.89°C for Eischeid *et al.* (1995).

5. SUMMARY

A binned least squares regression approach has been developed to estimate missing extreme maximum or minimum temperatures. The method focuses on obtaining accurate estimates of exceedence and consecutive exceedence counts by using a binning and optimization procedure. The method is equally valid for maximum or minimum temperatures and high (e.g. T_{90}) or low (e.g. T_{10}) extremes.

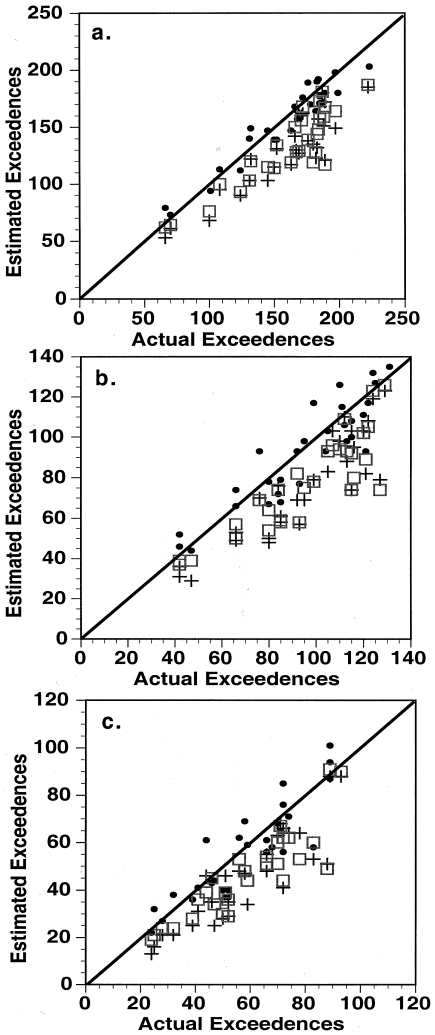

Figure 7. As in Figure 5 but for (a) 2-, (b) 3- and (c) 4-day runs of minimum temperatures not exceeding T_{10}

Compared with more conventional data estimation routines, our method significantly improves estimates of single and consecutive extreme exceedence counts. Evaluation of our approach using cold minimum and warm maximum temperatures showed that the median percentage of correctly identified exceedence counts was 97% and the median percentage of correctly identified consecutive exceedence counts was 98%. The other existing methods tended to underestimate both exceedence and consecutive exceedence counts. Using these procedures, estimated single and consecutive exceedence counts were generally less than 80% of those that actually occurred.

Despite the fact that the binned least squared method has been tuned to estimate exceedence counts, the estimation accuracy of an individual daily maximum or minimum temperature is similar to that of the other estimation procedures. The MAE using all temperatures $\geq T_{90} - 1.1$°C for 10 climatically diverse stations was 1.28°C for our method, while the other methods gave MAEs of 1.27 and 1.17°C. In terms of median error, however, the tendency for underprediction by the existing methods was pronounced with -0.77 and -0.61°C biases. Our optimization method was relatively unbiased as the resulting mean error was -0.12°C.

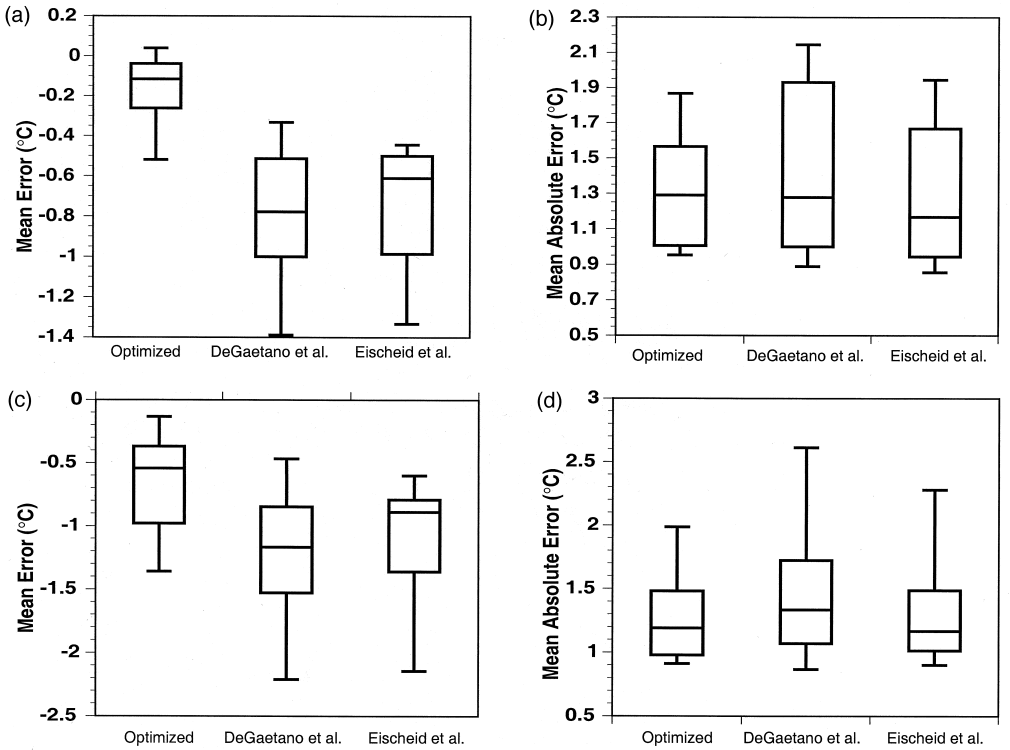

Figure 8. Box plots of (a) mean error and (b) mean absolute error for all days $\geq T_{90} - 1.1$°C and (c) mean error and (d) mean absolute error for all days $> T_{99}$ at 10 climatologically diverse stations using our optimized method, DeGaetano *et al.* and Eischeid *et al.*

ACKNOWLEDGEMENTS

This work was supported by the NOAA/NASA Grant NA76GPO351 and NOAA Cooperative Agreement NA67RJ0146. A pre-release version of the daily HCN data was obtained through the generosity of Dale Kaiser of the Carbon Dioxide Information Analysis Center at Oak Ridge National Laboratory.

REFERENCES

DeGaetano AT, Eggleston KL, Knapp WW. 1995. A method to estimate daily maximum and minimum temperature observations. *Journal of Applied Meteorology* **34**: 371–380.

DeGaetano AT. 2000. A serially complete observation time metadata file for US. *Daily Historical Climatology Network Stations Bulletin of the American Meteorological Society* **81**: 49–67.

Easterling DR, Karl TR, Lawrimore JH, Del Greco SA. 1999. *United States Historical Climatology Network Daily Temperature and Precipitation Data (1871–1997).* Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory: Oak Ridge, TN. ORNL/CDIAC-118, NDP-042/R1.

Eischeid JK, Baker CB, Karl TR, Diaz HF. 1995. The quality control of long-term climatological data using objective data analysis. *Journal of Applied Meteorology* **34**: 2787–2795.

Epstein ES. 1991. On obtaining daily climatological values from monthly means. *Journal of Climate* **4**: 365–368.

IPCC. 1996. *Climate Change 1995: The Science of Climate Change*. Cambridge University Press: Cambridge.

Karl TR, Williams CN Jr. 1987. An approach to adjusting climatological time series for discontinuous inhomogeneities. *Journal of Climate and Applied Meteorology* **26**: 1744–1763.

Kemp WP, Burnell DG, Everson DO, Thomson AJ. 1983. Estimating missing daily maximum and minimum temperatures. *Journal of Climate and Applied Meteorology* **22**: 1587–1593.

Oliver JE. 1973. *Climate and Man's Environment*. Wiley: Chichester.

Steurer P. 1985. Creation of a serially complete data base of high quality daily maximum and minimum temperatures. National Climatic Data Center, NOAA.

Vose RS, Schmoyer RL, Steurer PM, Peterson TC, Heim R, Karl TR, Eischeid JK. 1992. *The Global Historical Climatology Network: Long-term Monthly Temperature, Precipitation, Sea Level Pressure, and Station Pressure Data*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory: Oak Ridge, TN. ORNL/CDIAC-53, NDP-041.

