Climatic Data Bases

Jeffrey A. Andresen and Robert F. Dale
Department of Agronomy
Purdue University, West Lafayette, Indiana 47907

Introduction

Many climatological and agronomic models require daily values of meteorological variables. There are at least two general concepts for the use of weather data in modeling responses to weather and climate in an area, whether it be evaluating crop growth response to the weather or determining the effect of the weather on some dependent operation. In one concept, the model is run with the average weather data for the area, and in the other the model outputs, run with the weather data from each of several stations in and surrounding the area, are averaged. It is the first approach we are treating in this paper, that is, to create a daily series of average weather variables for an area which series can then be used for model estimates for that area. Data bases for meteorological variables that are representative of a mean for a county or climatological division (CD) containing several to many counties are usually difficult to obtain or develop. How do you calculate a "true" area mean from the weather stations in the area? At best, most areas have densities of about one station per county. Also, topographical differences may introduce greater variability within a division, making any areal averaging attempt even more difficult.

Several techniques have been developed and used to estimate county or CD means, including straight arithmetic or geometric averages and various objective analyses. Outputs from the schemes range from a single areal mean to a uniform grid network of the meteorological variable over the area of interest, making field analysis of the variable possible. In most all of these methods, the terrain is assumed to be uniform, which is probably valid for large areas of the Central U.S.

Besides the spatial variability of the variables involved, a source of error frequently ignored is that which arises from the use of data originating from different networks. In Indiana, data originate from several sources: the National Weather Service hourly reporting stations from which weather summaries are obtained for the calendar day ending at midnight, the climatological substation network in which once daily observations generally are taken either at 1700-2000 local time (PM) or at 0700-0900 local time (AM). Because of these observational differences, weather data taken for a specific day may actually be more representative of the previous day, especially for the maximum temperature recorded at AM stations. Aside from publishing the maximum temperature on the day following that of its occurrence, Schaal and Dale (1977) found that mean daily temperatures calculated for AM stations averaged 1.3°F lower than those for PM stations.

The objectives of this research were to test for differences between divisional daily averages for several weather variables when calculated with three commonly-used areal averaging techniques, as well as for the effect of correcting for the heterogeneites arising from different observation times.

Data

Data from Indiana climatological data (USDC, NCDC, 1982) were obtained from the Midwest Agricultural Weather Service Center (MAWSC) at Purdue University. Variables used in the study were daily precipitation, maximum and minimum temperatures, and Class A pan evaporation. A sampling period of 1 January, 1981 through 31 December, 1981 was chosen because the year was near normal in most

respects. The sample area chosen for the study was the Indiana Central Climatological Division, CD5, since data outside the area borders were readily available, a condition necessary to prevent border bias effects. The stations in the Indiana data base for the different variables are shown in Figures 1a, b, and c. The stations were chosen

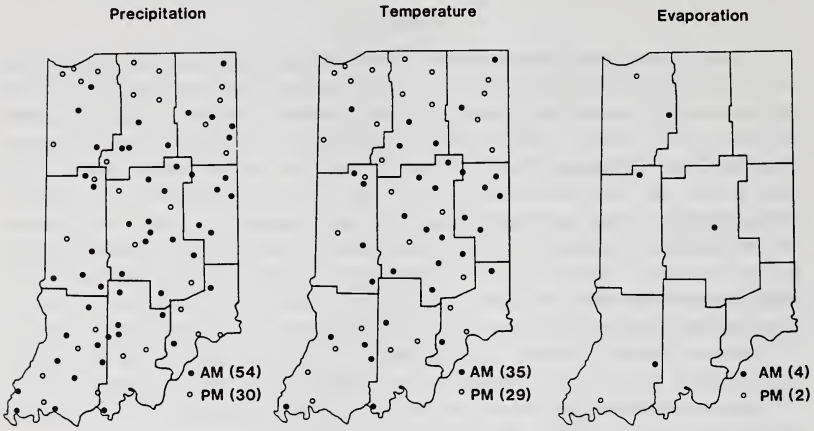


Figure 1. Indiana climatological stations used for a) precipitation, b) temperature, and c) pan evaporation daily divisional averages. Closed and open circles denote AM and PM stations, respectively.

on the basis of best possible areal uniformity and for completeness of station record. The percentage of stations taking AM observations ranged from 67% for pan evaporation to 55% for temperature.

Averaging Techniques

Three averaging methods were chosen for the study: A simple arithmetic average of the stations within CD5, a reciprocal distance weighting method (REC), and a Purdue Regional Objective Analysis of the Mesoscale (PRO). Also for the maximum temperature and daily total precipitation, the simple arithmetic average was prepared both from the uncorrected UCR calendar day data (as published in Climatological Data) and after setting the variables back to the previous day for the AM stations (COR). In the straight averaging method, data were used from 16 precipitation and 15 temperature stations within CD 5. Because the data base is much more sparse for pan evaporation, only the REC and PRO methods were used in calculation of this variable. For both the REC and PRO methods a 17-row by 11-column grid network was superimposed over the entire state. These dimensions represent the approximate length (north-south) and width (east-west) of Indiana. This gave a grid resolution of approximately 28km × 28km, or roughly county size. The CD5 average in both cases was then calculated as the mean of the grid points within the division borders (15 points in all). The REC method for each grid point is given by:

$$REC_{pt} = \frac{\sum_{i=1}^{n} \frac{X_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}}$$

where REC_{pt} is the average of the variable for a grid point, X_i are the values of the variable at the n closest stations to the grid point, and d_i are the respective distances to the n closest stations. For the study, n was set equal to 3, which in most cases represented stations within 30km, except for pan evaporation.

The PRO objective analysis was developed at Purdue to aid in the study of mesoscale events (Smith and Snow, 1983). The scheme involves two separate passes, an initial pass to fit the station data to a grid as an exponential function of the distance of the station to the grid point, and secondary pass(es) to further improve the fit of the grid points relative to their immediate surroundings. Two corrective passes were used in all runs of the study. The number of stations involved in the calculation of a grid point value is determined by a preset radius, within which all stations are included down to a minimum of 2. A more detailed description of PRO is given in Smith and Leslie (1982).

Data Analysis

The data were split into four seasonal groups. It should be noted that the winter data set is not continuous, consisting of January, February, and December data from the same year, 1981. The primary method of comparison between the techniques was simple linear correlation for paired estimates from the methods. For daily maximum temperature and precipitation this was done with all four methods, including the straight average of the uncorrected and corrected data as two methods, for minimum temperature the three methods and for evaporation, because of the sparse network only the REC and PRO methods. Each method pair was also tested for significant differences between means and variances, through use of paired observation t-tests and variance ratio F-tests, respectively. For precipitation, bounded at one end of the distribution by zero, the data were log transformed to approximate a normal distribution as:

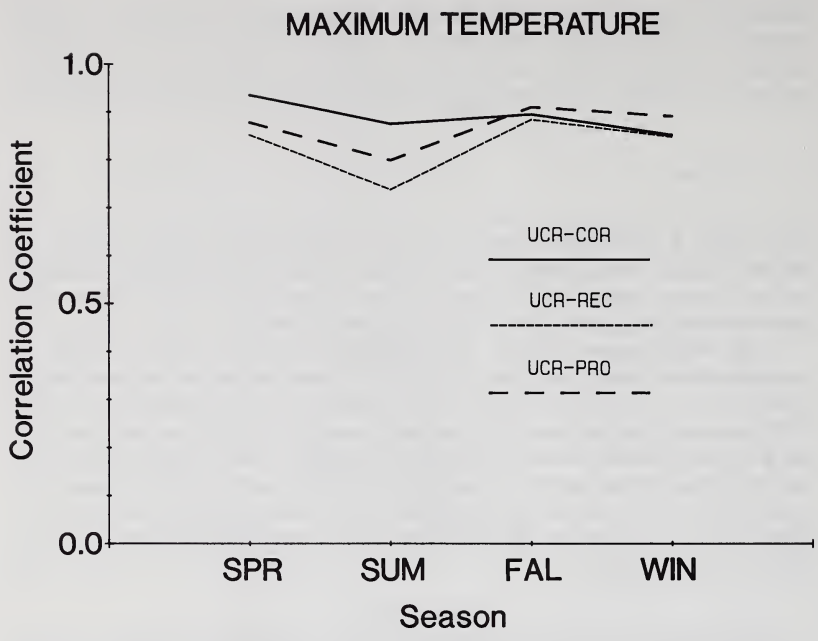
$$PPT_t = \ln (1 + PPT)$$

where PPT is the input daily precipitation, and PPT_t is the transformed value. All significance tests were carried out as the $\alpha = 0.10$ level.

To measure the relative efficiency of the methods, computing time samples were taken during the actual computer runs to determine the computational time required for a one day iteration (all 9 Indiana CD means calculated) for each of the methods.

Results and Discussion

To show the relation of the daily Central Division averages of the indicated variables calculated with the different smoothing methods, correlation coefficients were plotted on season for three variables in Figures 2-4. In Figure 2a, the solid line shows the correlations by seasons between the average divisional daily maximum temperatures computed from the uncorrected and corrected series of divisional daily mean maximum temperatures. Again, the uncorrected (UCR) daily divisional average maximum temperatures are computed from all 15 stations (Figure 1b) as published for the same calendar day in Climatological Data (USDC, 1982). The corrected (COR) daily divisional average maximum temperatures are calculated after setting the maximum



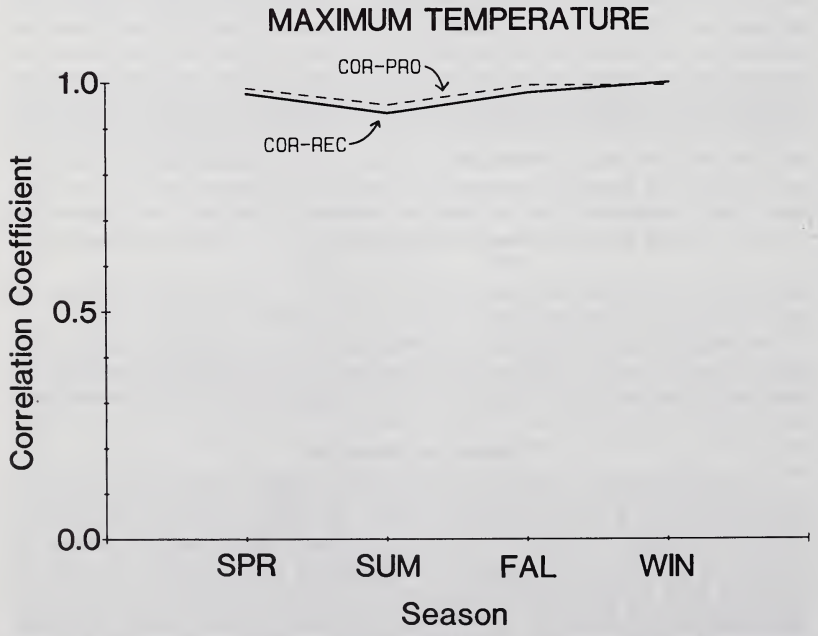


Figure 2. Seasonal correlation coefficients for daily divisional average maximum temperatures estimated with indicated methods with a) data uncorrected for observation time, and b) corrected data only.

temperatures for the 11 stations taking AM observations back to the previous calendar day. The correlations of UCR with COR divisional mean daily maxima decrease from 0.95 in the spring to 0.85 in the winter. The correlations between the UCR daily series and reciprocal distance averages with corrected data (REC) range between 0.75 in the summer to 0.85 in the fall, similar to those shown for the UCR and PRO correlation pattern.

When the time of observation source of error is removed (COR), and only the smoothing methods compared, as shown in Figure 2b, the correlations range from 0.92 to 0.99. Considering the close agreement between the methods using corrected data, it is apparent that the differences caused by smoothing methods are insignificant compared to those created by the use of uncorrected heterogeneous data.

Since daily minimum temperatures usually occur on the day of observation for both AM and PM stations, there is only one, or correct, set of divisional mean daily minimum temperatures. The results for minimum temperature (Fig. 3) are similar to those for COR maxima (Figure 2b).

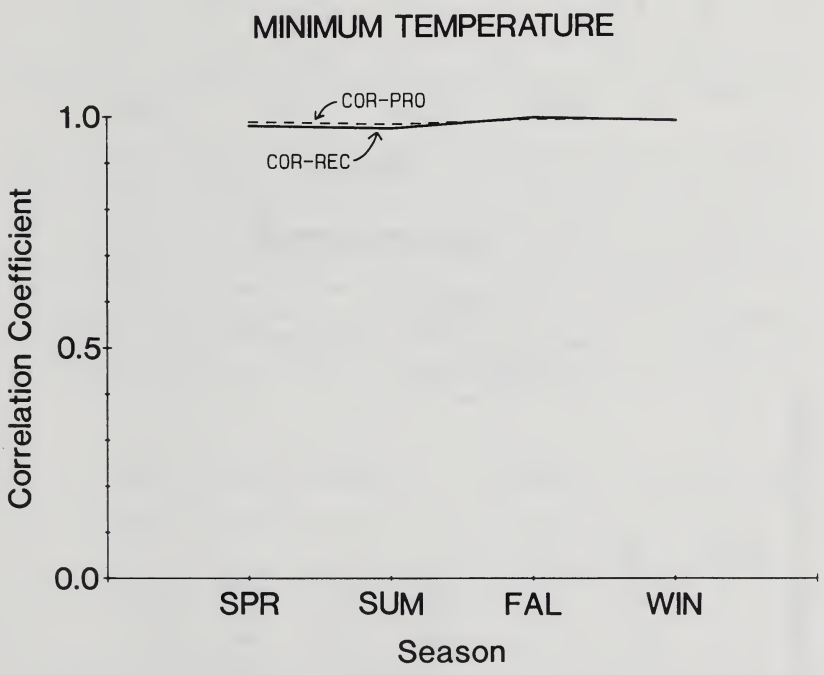
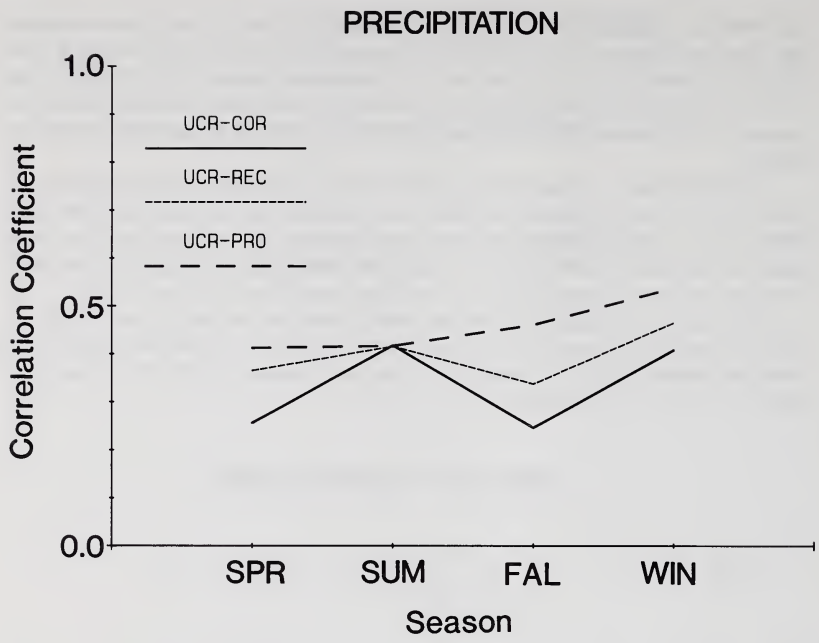


Figure 3. Seasonal correlation coefficients for daily divisional average minimum temperature estimated with indicated method.

The pattern is even more enhanced for precipitation (Figures 4a, b), with correlation coefficients ranging between 0.25 and 0.55 for the UCR with COR, UCR with REC, and UCR with PRO comparisons (Figure 4a). When only the "corrected" divisional daily precipitation means are used, the correlations of COR with REC increase to 0.96 in the spring and winter, with a low of 0.72 in the fall. The correlations of COR with PRO are above 0.96 in all seasons, showing closest agreement between the cor-



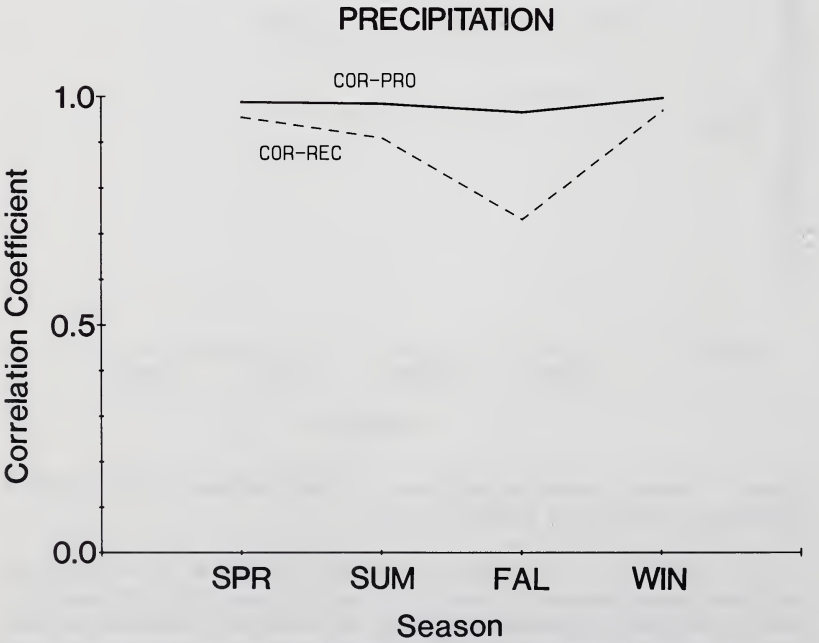


Figure 4. Seasonal correlation coefficients for daily divisional average precipitation estimated with indicated method with a) data uncorrected for observation time, and b) corrected data only.

rected daily divisional average precipitation and that obtained with the PRO objective method. We should point out that the "corrected" daily divisional mean precipitation series is not as correct as that for the maximum temperature, because we have about a ⅔ probability of aligning the daily precipitation on the proper calendar day.

The mean and mean absolute daily differences (errors) between divisional daily average maximum temperatures for the different smoothing methods are shown in Table 1 for the four 90- to 92-day seasons. Since there are only two days of the 90 to 92

TABLE 1. Seasonal mean and mean absolute daily differences between divisional daily average maximum temperatures for three pairs of averaging methods. * denotes significant difference at the $\alpha = 0.10$ level.

No. days	Mean Differences (°F)			Mean Absolute Differences (°F)		
in sample	COR-UCR	COR-REC	COR-PRO	COR-UCR	COR-REC	COR-PRO
SPR(92)	0.0	−1.4*	−1.4*	3.6	2.6	2.0
SUM(92)	0.2	−0.2	−0.3	1.8	1.2	0.9
FAL(91)	−0.3	−0.5*	−0.7*	4.1	0.9	1.2
WIN(90)	0.1	−0.6*	−1.0*	4.4	0.6	1.2
ANN(365)	−0.1	−0.6	−0.8	3.5	1.3	1.3

days in each season (the first and last) which are different between the COR and UCR series, the mean differences between the UCR and COR series are near 0. The differences between the mean absolute differences (signs ignored), however, range from 4.4 °F in the winter to 1.8 °F in the summer, obviously important should the divisional daily averages be used in any sensitive crop or industry response model. The difference between the daily divisional average maximum temperatures computed with the COR and REC methods (COR - REC) were slightly larger than 0 and consistently negative, ranging from −1.4° F in the spring to −0.2° F in the summer. Also, spring, fall, and winter differences were found to be significantly different from 0 at the $\alpha = 0.10$ level, partially because of the small standard error (Table 2). The range for the

TABLE 2. Seasonal standard deviations (°F) and standard errors (in parentheses (°F)) of daily maximum temperature differences for three pairs of averaging methods.

	COR-UCR	COR-REC	COR-PRO	Effective Sample Size	Lag 1 Serial Corr. Coeff.
SPR	13.8(1.4)	8.5(0.8)	6.2(0.6)	10	0.80
SUM	4.1(0.1)	3.1(0.1)	2.6(0.1)	28	0.53
FAL	12.8(0.7)	2.3(0.1)	3.3(0.2)	19	0.65
WIN	14.1(0.9)	1.3(0.1)	3.1(0.2)	16	0.70

mean absolute differences between the COR and REC methods was from 2.6 °F in the spring to 0.6° F in the winter. Since the corrected maximum temperature series were used in both COR and REC, at least part of the differences is caused by the patterns of the areal weighting of the stations within the division. In the COR average, all 15 stations (Figure 1a) in the Central Division are weighted equally, and in the REC average, 15 uniform grid points are weighted equally, each of the grid points based on distance-weighted observations at the three closest stations. The COR and PRO difference pattern is about the same, showing a small but consistent negative

bias. The mean absolute differences for both COR with REC and COR with PRO averaged between 1 and 2°F.

Another measure of the error between the three methods of estimating the daily divisional average maximum temperature series is given in Table 2, the standard deviations of the differences, together with the standard error. While the computation of the squares of the daily differences is statistically straight forward, the effective sample number of days to be used in estimating the mean square error, the standard deviation, and the standard error is not. Since the daily temperature and precipitation from one day to the next are not independent, the autocorrelations and the effective sample size in the formula for determining "independent" days, calculated from Brooks and Carruthers (1953, p. 326, Eq. 280), are also included in Table 2. The standard deviations of the differences between the divisional average daily maximum temperatures estimated with UCR and COR ranged from 14.1 to 4.1°F. The standard errors (in parentheses in Table 2) were used in testing the significance of the mean differences in Table 1. The standard deviation of the daily differences between COR and REC and COR and PRO ranged from 8.5 to 1.3.

There were no significant differences for the divisional daily minimum temperature COR vs. REC and COR vs. PRO, or for pan evaporation REC vs. PRO. Those tables are not included.

The mean differences and the mean absolute differences between the average divisional daily precipitation estimated with the UCR, COR, REC, and PRO methods are shown in Table 3. None of the mean differences are significantly different from

TABLE 3. Seasonal mean and absolute daily differences between divisional daily average precipitation for three pairs of averaging methods.

	Mean Differences (in.)			Mean Absolute Differences (in.)		
	COR-UCR	COR-REC	COR-PRO	COR-UCR	COR-REC	COR-PRO
SPR	0.00	0.00	0.01	0.18	0.02	0.05
SUM	0.00	0.01	0.00	0.14	0.02	0.06
FAL	0.00	0.01	0.00	0.10	0.02	0.05
WIN	0.00	0.00	0.00	0.08	0.01	0.02
ANN	0.00	0.00	0.00	0.12	0.02	0.04

zero, as the mean differences for all methods are very near 0. Crosiar (1982) also found no significant differences between REC, a PRO-like method, and a Thiessen polygon method. The mean absolute differences, however, ranged from 0.18 to 0.08 for the UCR vs. COR, considerably larger than the 0.01 to 0.02 for COR with REC and 0.02 to 0.06 for COR with PRO.

The standard deviations of the differences, with their respective standard errors, between methods for estimating divisional mean daily precipitation are shown in Table 4. Again the standard deviation of the differences, decoded from the log transform, was greatest for the UCR vs. COR comparison, ranging from 0.25 in the summer to 0.15 inches in the fall. Even though the mean difference is near 0, the standard deviations show these are large errors, even for nonsensitive models! The standard deviation of the differences between estimates of the divisional mean daily precipitation for COR vs. REC ranged between 0.04 in the spring and summer to 0.02 in the winter, and for COR vs. PRO from 0.10 in the summer to 0.05 in the winter. As with the maximum temperature, the smoothing method had much less effect on the error than the use of the proper data base.

Table 4. Seasonal standard deviations (in.) and standard errors (in parentheses (in.)) of divisional daily precipitation differences for three pairs of averaging methods.

	COR-UCR	COR-REC	COR-PRO	Effective Sample Size	Lag 1 Serial Corr. Coeff.
SPR	0.20(0.02)	0.04(0.00)	0.07(0.01)	78	0.08
SUM	0.25(0.03)	0.04(0.00)	0.10(0.01)	54	0.26
FAL	0.15(0.02)	0.03(0.01)	0.09(0.01)	91	0.00
WIN	0.18(0.04)	0.02(0.00)	0.05(0.01)	59	0.21

The time (seconds) taken by the computer per iteration (daily) for all nine divisions for the different methods and variables are shown in Table 5. A great time difference can be seen between the straight average and the objective analyses, an obvious result of the greater complexity of the latter. A pattern is also detected between REC and PRO. For both, the greater the number of stations included in the analysis, the longer the computation time required, but the REC method was affected to a greater extent. The required time per iteration increased 122s from a 6-station network (evaporation) to an 84-station network (precipitation) while the PRO method changed 111s for the same increase.

Table 5. Mean computational times (seconds per daily iteration) for all nine climatological divisions in Indiana with indicated methods for precipitation, temperature, and pan evaporation.

Method	Precipitation	Temperature	Evaporation
COR	0.81	0.78	--
REC	129	127	7
PRO	131	119	20

Conclusions

Three different methods of areal averaging were compared. When corrected for network observational time differences, no significant differences existed between method means and variances for daily divisional average precipitation. There were some significant differences between methods for daily divisional average maximum temperatures, but throughout the study the greatest dissimilarity between the techniques was caused by a heterogenous data base, including both PM and uncorrected AM stations. Considering this, the simplicity of the data network correction, and the large computational requirement differences, it appears that for climatological division daily averages, the straight arithmetic average would be preferable to an objective analysis under most circumstances, with no appreciable loss in accuracy. For cases in which a straight average is not possible (pan evaporation), either method discussed previously would be suitable, although the REC method was quicker computationally for small data networks.

Acknowledgments

The authors wish to express their gratitude to the entire staff of the MAWSC facility at Purdue for their cooperation, help, and use of facilities for this study. We also wish to thank Mrs. Lois Edwards for her help in preparing this manuscript.

Literature Cited

Brooks, C. E. P. and N. Carruthers. 1953. Handbook of Statistical Methods in Meteorology. Her Majesty's Stationery Office, London. 412 p.

Crosiar, C. L. 1982. Characterization of Regional Climate. M. S. Thesis, Univ. of Missouri-Columbia.

Schaal, L. A. and R. F. Dale. 1977. Time of Observation Temperature Bias and "Climatic Change." J. Appl. Met. 16:215-222.

Smith, D. R. and F. W. Leslie. 1982. Evaluation of a Barnes-Type Objective Analysis Scheme for Surface Meteorological Data. NASA tech. memo-82509, Marshall Space Flight Center, 25 p.

Smith, D. R. and J. T. Snow. 1983. Objective Analysis of Mesoscale Disturbances Using Surface Meteorological Observations. Ind. Acad. Sci. 92:432.

U. S. Department of Commerce, NOAA, National Climatic Data Center. 1982: Climatological Data, Indiana, 1981, Vol. 86, Monthly and Annual.