A Note on the Precision of Estimation of Missing Precipitation Data

James E. McDonald

Abstract—Climatological studies in which early precipitation records are used frequently lead to the necessity of estimating missing data. The present note summarizes a small series of tests of one objective estimation scheme identical in form to the normal-ratio method of Paulhus and Kohler but applied here to missing seasonal totals rather than to missing storm totals. The results suggest that estimation errors in the neighborhood of 25 pct must be expected in estimating seasonal totals using two index stations under conditions typical of those forced upon the investigator dealing with older records. Increasing the number of index stations beyond two does not seem warranted. Errors of 25 pct in seasonal totals, though not negligible, are small enough compared to coefficients of variation typical of seasonal totals of precipitation in the drier portions of the United States, that the use of such a method in climatic studies is desirable.

Introduction—In many meteorological and hydrologic investigations, it becomes necessary to use precipitation records which contain intermittent breaks due to change of observing personnel, absence of observer from his station, etc. Such breaks are, unfortunately, fairly frequent in older (say pre-1920) records from Weather Bureau cooperative observers whose unpaid services could not always be obtained entirely continuously, and the result is that the user of such data must either exclude from analysis those periods of record containing gaps, or else find an objective estimation scheme for filling the gaps. Although it would be difficult to defend with quantitative arguments the view that any sort of reasonable objective estimate is preferable to an isolated gap, this is in fact the writer's view. It becomes especially desirable to fill such gaps with systematic estimates in investigations where automatic analysis equipment is being used to process punchcard data, for programming such equipment to enable it to accept gaps is usually quite inconvenient.

In the cases where the investigator has available the original daily records, the problem of missing monthly or seasonal totals can be solved by estimating those particular daily precipitation amounts whose absence from the original records creates the missing monthly total. A rather extensive study of this problem has been conducted by *Paulhus* and *Kohler* [1952] in the context in which this problem arises in current processing of climatological data by the U. S. Weather Bureau. However, if one does not have original daily records available, as is most commonly the case, then a distinct question, not answered by the results of those investigators, arises. One may prefer to fill gaps in the monthly, seasonal, or annual records for stations by some objective scheme that entails use of no more detailed data than the widely available monthly totals.

In the U. S. Weather Bureau–University of Arizona cooperative punchcard climatology program just this problem arose with regard to seasonal precipitation totals missing from non-Arizona records whose originals were unavailable. To shed some light on the order of magnitude of errors incurred by one specific estimation scheme, a small synthetic experiment was conducted. The results are summarized here since, though limited in scope, they give a useful indication of the error level for an estimation problem not treated elsewhere in the literature, to the knowledge of the writer.

Estimation equation—The estimation scheme that has been tested and used in the University of Arizona precipitation studies is identical in form with that termed the 'normal-ratio method' by Paulhus and Kohler. It is best described in symbolic form as follows: Denote by P_{ik} the precipitation reported for the ith time-period (for example, the ith day of a month, ith month of a year, etc.) at the kth station of a group of climatological stations in a given area, and let $\overline{P}_{ik}$ be the mean value for that kth station and ith time-period over a specified long period of years. (Thus, if one were dealing with monthly totals, $i = 3$ might specify March, and $\overline{P}_{3k}$ would be the long-term mean March precipitation at the kth station.) If the jth station has no report for the ith period, then an estimate of this missing datum, denoted by (P_{ij}), is given by

$$(P_{ij}) = (1/n) \sum_{k=1}^{n} \frac{\overline{P}_{ij}}{\overline{P}_{ik}} \cdot P_{ik} \qquad (1)$$

where the sum over k from 1 to n and division of the right member by n implies that one uses as the final estimate an average of n individual estimates derived from actually reported values at n stations located near the jth station. Paulhus and Kohler have referred to the station whose missing datum is to be estimated as the 'interpolation station,' while the several stations whose data are employed to carry out the estimate are termed by them 'index stations.' It seems desirable to adopt their terminology here.

Discussion of estimation scheme—The following general comments about the scheme of (1) may be made:

(a) This scheme would seem, on general principles, to be superior to any that gave (P_{ij}) as some function of only the concurrently reported values, P_{ik}, for (1) weights the estimate in proportion to the ratio of the long-term means at the jth and kth stations, and in actual practice one usually has to deal with station-pairs involving unequal means. The scheme of (1) is, at the same time, clearly superior to mere insertion of $\overline{P}_{ij}$ as a crude estimate of (P_{ij}), for (1) allows any anomalous characteristics of the precipitation peculiar to the particular ith time-period in question to enter through the P_{ik}. Briefly, (1) gives weight both to long-term climatic relationships between the interpolation station (jth station) and the index stations and also to the departures from normal characterizing the ith period (as revealed by the departure from unity in the n ratios, $P_{ik}/\overline{P}_{ik}$), that appear as factors in the n terms of the sum on the right of (1). *Paulhus* and *Kohler* [1952] have shown, in fact, that this normal-ratio method gives substantially more reliable estimates of missing short-term precipitation amounts (missing days or storm-totals) than a simple scheme based just on current totals at nearby stations, and reference should be made to their paper for additional arguments in favor of a normal-ratio approach.

(b) The scheme symbolized by (1) tacitly assumes that ratios of precipitations rather than sums or differences should be used in estimating missing data. This general point is discussed by *Conrad* and *Pollack* [1950, pp. 235–237], who conclude that in the case of precipitation data, it is the ratio and not the difference of two stations' receipts that are quasi-constant. Their conclusion rests, apparently, only on empirical evidence and such evidence is not elaborated by those authors. One might question the generality of the ratio-assumption on the grounds that inter-station precipitation regression equations frequently contain non-zero intercepts whose magnitude is sometimes so great as to seemingly vitiate the ratio method. However, this difficulty must, in almost all cases, become unimportant when the n index stations are all chosen from the immediate geographic (climatographic) vicinity of the interpolation station, for then the several regressions of the jth on the kth stations can reasonably be expected to exhibit zero or nearly zero intercepts.

(c) It is quite essential that all of the means, $\overline{P}_{ij}$ and $\overline{P}_{ik}$, be means computed for the same period of years, for otherwise secular trends in precipitation can very easily introduce systematic error into the estimation.

(d) If any information exists as to patterns of correlation between the interpolation station and surrounding index stations, it is obvious that one should select the index stations from regions of maximum correlation to optimize the estimate in (1), but lacking such information, proximity and topographic factors must serve as deciding considerations. In the program with which the writer is associated, only the latter have been used thus far, for reasons of present lack of detailed information concerning correlation patterns.

After having studied the efficacy of the estimation scheme of (1) it was found that essentially this same scheme has been discussed earlier by *Keeler* [1944], who tried several elaborations of this scheme, including one in which weighting coefficients dependent on the azimuthal distribution of the index stations were employed in making the estimate. Keeler gives no figures indicating comparative precisions of the several elaborations he tried, but states that he finally abandoned them in favor of the kind of scheme given here as (1). He tried values of n from one to five and states that three or more index stations are desirable, but gives no quantitative evidence of comparative performance of the scheme for various n-values. He suggests that the number of years of record entering into the precipitation means employed in the estimates should be "five years or more, if possible," which seems to the writer to be unwarrantedly optimistic. Keeler does not stress the importance of having the means appearing, respectively, in numerator and denominator of the right member of (1) pertain to the same time-periods, yet secular trend effects can, above all with such short averaging periods as accepted by Keeler, lead to quite serious errors of estimate. It bears emphasis that in the arid Southwest (and in similar regions of the world), variability of precipitation is so great that a mere five-year mean can be quite non-representative. Indeed, a study made by the writer of selected Arizona long-record stations revealed that the 95 pct confidence half-widths of seasonal means of precipitation were about ten per cent of the means, even for stations with records of up to 80 years, and rose to about 20 pct of the mean for records as short as 20 years. Consequently, unless the inherent relative errors of the estimation scheme itself should prove to be much larger than 20 pct, one ought not be satisfied with means based on much less than 20 years, and these should be the means for the same 20 years for the station whose missing datum is to be estimated and for the index stations.

Test of reliability of estimation scheme—As noted above, there are no data in the literature that provide a quantitative evaluation of the error level to be expected for missing seasonal values interpolated by the use of (1), so a simulated run was carried out, 'estimating' winter and summer seasonal totals for two Arizona stations, Tucson and Natural Bridge, for all years from 1900 to 1929, using a variety of predictor-station combinations.

For the purposes of the test of (1), four index stations were chosen for use in a series of synthetic estimates for Tucson, and four for the estimates for Natural Bridge.

These stations, and their airline distances from the two interpolation stations (Tucson and Natural Bridge) are as follows:

Tucson predictors: Oracle, 30 miles north-northeast; Benson, 45 mi east-southeast; Phoenix, 110 mi northwest; Bisbee, 85 mi southeast.

Natural Bridge predictors: Prescott, 60 mi west; Flagstaff, 60 mi north-northwest; Ft. Apache, 95 mi east-southeast (the 1931–52 mean used in the Ft. Apache estimations was that for Whiteriver, about five miles north of Ft. Apache); Phoenix, 70 mi south-southwest.

Previously, as part of another study, double-mass plots of the long-term precipitation records of Tucson and Natural Bridge had been made to determine whether any significant breaks occurred in the histories of these two stations. These double-mass plots are not reproduced here but are presented elsewhere [*McDonald*, 1956], and show that both records are free from discontinuities of magnitude that would vitiate their use in the present tests.

It may be questioned why the test was applied to index stations as far away as 110 miles when there are somewhat closer predictor stations that would surely be preferred in the actual applications of (1). The answer is twofold: First, it was desired to approximate the least favorable rather than the most favorable situations actually encountered, and second, it was desired to apply the test using recent mean ratios in combination with old reports, since this is the most frequent case encountered in practice. Specifically, in the current studies being conducted by the Institute of Atmospheric Physics, we have had to do the largest amount of estimation of missing data for periods before about 1920. But since the two means appearing in the right member of (1) must be means for exactly the same period (or at least for periods which are so nearly identical as to rule out distorting effects of secular trends), it becomes almost indispensable to use in the right side of (1) means over some recent period, common to all record-spans.

Availability of the recently published *Supplements* by the *U.S. Weather Bureau* [1955] has made it an obvious choice to use, throughout all our estimation work, the 22-year means for 1931–52. These 1931–52 means were therefore employed in the synthetic test, but were used to 'estimate' Tucson and Natural Bridge data for the 30-year period 1900–29 as a fairly realistic test of the practical case where gaps in old records are to be filled using means for a more recent period.

Estimates were made for values of n equal to 1, 2, and 4 in (1) in order to gain some notion of how much increase in accuracy might be expected to result from variations in n. Because of gaps in the 1900–1929 records for the four predictor stations, it was not possible to obtain 30 estimates in each case, so for uniformity, the first twenty estimates available, beginning with 1900 were used to evaluate the average error for the present purposes.

The precision of the estimation scheme might be measured in any of a number of ways, of which that chosen for use here was the computation of the average of the absolute values of the errors of 20 pct in each predictor season category. Use of the errors rather than their squares seems desirable since an occasional highly erratic prediction is weighted by a squaring process in an undue proportion for its seriousness. Use of averages of percentages is generally undesirable, and employment of logarithms of relative errors would have been somewhat better here; but since the individual percentage errors did not vary over an extremely large range, a simple average of percentages was felt to provide a fairly adequate

TABLE 1 – *Mean errors of estimation of missing precipitation data*

Predictor stations	Mean error of estimate for Tucson[a]		Predictor stations	Mean error of estimate for Natural Bridge	
	W[b]	S[b]		W[b]	S[b]
	pct	pct		pct	pct
Oracle	23	45	Prescott	22	23
Benson	29	28	Flagstaff	43	35
Bisbee	21	16	Ft. Apache	25	27
Phoenix	26	36	Phoenix	18	34
Oracle, Benson	18	25	Prescott, Flagstaff	31	27
Phoenix, Bisbee	19	23	Ft. Apache, Phoenix	17	27
Oracle, Benson, Phoenix, Bisbee	18	20	Prescott, Flagstaff, Ft. Apache, Phoenix	20	21

[a] Tabular values are means of absolute values of twenty percentage-errors of estimate.

[b] Winter (W) is here the six-month period from Nov. 1 to April 30; summer (S) is the remainder of the year.

TABLE 2 – *Distribution of signs of errors of estimation*

Predictor stations	Per cent of Tucson errors with positive signs		Predictor stations	Per cent of Natural Bridge errors with positive signs	
	W[a]	S[a]		W[a]	S[a]
Oracle	50	44	Prescott	63	59
Benson	32	36	Flagstaff	67	60
Bisbee	59	52	Ft. Apache	41	29
Phoenix	30	37	Phoenix	30	30
Oracle, Benson	32	35	Prescott, Flagstaff	66	62
Phoenix, Bisbee	37	59	Ft. Apache, Phoenix	30	29
Oracle, Benson, Phoenix, Bisbee	35	45	Prescott, Flagstaff, Ft. Apache, Phoenix	54	41

[a] See footnote b of Table 1.

measure of overall error, at least for the present purposes.

In Table 1 the results of the synthetic test are summarized. For each of the two interpolation stations (Tucson, Natural Bridge) and for each of two six-month seasons, a total of seven different series of estimates were made. The first four of each set of seven are single-index-station estimates, the next two are two-index-station estimates, and the last is a four-index-station estimate (that is the summation over k in (1) is taken in this last instance from 1 to 4). The percentage errors, of which 20 were averaged to give each of the entries of Table 1, were obtained by dividing the absolute value of the difference between the individual estimate and the actually reported value by the actually reported value, or in the symbolism of (1), $|P_{ij}) - P_{ij}|$ divided by P_{ij}, where P_{ij} is the true Tucson or Natural Bridge seasonal total and (P_{ij}) the corresponding estimate.

Table 1 reveals nothing about the sense of the departure of estimated value from reported value, since only absolute values enter into that tabulation. Since secular trend effects can be different at two stations for two different periods of time, it has to be admitted that making estimates in an early period such as 1900–1929 but employing means for a recent period such as 1931–1952, introduces possibility of systematic bias in the signs of the errors of estimate. To obtain a rough measure of this type of bias, each of the (fourteen) series of estimates was processed as follows: Using all available estimates for the entire period 1900–1929 for each predictor-season combination, the percentage of positive signs of departure of estimates from reported values were computed separately for each of the fourteen cases. The results are found in Table 2.

Discussion of results—The implications of Tables 1 and 2 may be summarized as follows:

(a) Very generally, all of the error-levels of Table 1 are in the neighborhood of 20–30 pct; that is, it appears that the method of (1) must be expected to yield in the neighborhood of a 25-pct error for estimated seasonal totals in Arizona. Somewhat curiously, this is of the same magnitude as the errors reported by *Paulhus* and *Kohler* [1952] for normal-ratio estimates of storm totals using three index stations. Apparently the smoothing effect of going to longer time periods, as in the present test, is roughly cancelled by lumping together the contribution of many individual storms that give widely variable amounts to the several stations under consideration, a compensation that would have been difficult to anticipate, in the writer's view.

(b) An improvement results, in general, as one increases the number of stations used, that is, as n is increased, but this trend is seen to be a bit erratic, and the gain is not extremely impressive as n increases from one to four.

(c) Distance between index station and interpolation station is not correlated in any very strong way with error of estimation. Thus, for the Tucson winter estimates, Bisbee is the best single-station estimator, though its distance from Tuscon is twice that for Benson and three times that for Oracle. Similarly Phoenix is the best estimator of Natural Bridge winter data though it is slightly farther from Natural Bridge than either Prescott or Flagstaff. The latter case is additionally surprising inasmuch as Phoenix is well out of the mountainous area within which Natural Bridge, Prescott, and Flagstaff all lie. The correlation coefficient for the relation of Natural Bridge to several other stations was computed for the available periods of record in related study [*McDonald*, 1956] and the results are of some interest here. For winter totals, the correlation with Phoenix was 0.85, which exceeded that for Flagstaff, 0.57, but was slightly smaller than was the value of 0.87 for Prescott. Distance and topographic similarity do not at all emerge as decisive clues to selection of optimal predictors in the present problem.

(d) In general, one sees that estimation is a bit more accurate in winter than in summer in Arizona. Thus, ten of the 14 cases of Table 1 show better estimates in winter than in summer. Again, however, the differences are not uniformly large.

(e) The distribution of algebraic signs of estimation errors, as revealed by Table 2, presents a somewhat confusing picture. No uniform trends appear evident, with the partial exception of the Tucson winter estimates, for which there is a noticeable tendency towards a deficiency of positive errors. It would seem that no very serious bias has been introduced into the present sample of estimates as a consequence of treating the practically important situation wherein one estimates old data using recent means, but beyond that observation, it seems difficult to generalize from Table 2.

Conclusions—On the basis of these results of the empirical test of the estimation scheme of (1), it is concluded that, at least for the Southwest, two index stations should be used as predictors, for the added labor of increasing n beyond two seems scarcely justified. It is further concluded that, if two index stations can be selected from within a radius of 100 miles centered on the station whose missing datum is to be estimated, the relative error of estimate may be expected to be in the neighborhood of 25 pct, on the average.

In most practical applications in studies of precipitation in the Southwest (except for the very earliest periods of record when stations were few and far between), the above requirements can be met. A 25-pct error in a single missing seasonal total, though hardly negligible, is sufficiently small (especially as compared with typical southwestern station coefficients of variation), that definite advantage is, in the writer's opinion, gained by carrying out this type of estimate rather than merely omitting missing periods from analysis.

Acknowledgments—The work reported here was done as one part of a study supported by the National Science Foundation under Contract NSF-G1101. The writer is indebted to Patricia Morgan for her assistance in the computational work.

References

Conrad, V., and L. W. Pollak, *Methods in climatology*, 2nd ed., Harvard Univ. Press, 459 pp., 1950.

Keeler, K. F., Some uses and deficiencies of climatological data, *Trans. Amer. Geophys. Union*, **25**, 420-432, 1944.

McDonald, J. E., Variability of precipitation in an arid region: A survey of characteristics for Arizona, *Technical Reports on the Meteorology and Climatology of Arid Regions, 1*, Institute of Atmospheric Physics, University of Arizona, Tucson, Ariz., 88 pp., 1956.

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

U. S. Weather Bureau, *Climatic summary of the United States—Supplement for 1931 through 1952, Arizona, Total precipitation*, U. S. Government Printing Office, Washington, D. C., 59 pp., 1955.

Institute of Atmospheric Physics, University of Arizona, Tucson, Arizona.

(Communicated manuscript received January 8, 1957, and, as revised, May 27, 1957; open for formal discussion until March 1, 1958.)