Hydrologic Automation

PROCESSING OF PRECIPITATION DATA FROM A NETWORK OF AUTOGRAPHIC AND AUTOMATIC RAINGAUGES

by Nathaniel B. Guttman (1)
National Climatic Data Center, NOAA
Asheville, NC 28801

ABSTRACT

Precipitation data are collected from many independent networks of rain gauges for application in numerous fields. This paper describes the procedures used in the processing of data collected from the hourly precipitation network. Climatological questions about data homogeneity are also addressed. The treatment of instrument problems and the meteorological screening techniques could be used in part if not in totality by other network data processors.

1. INTRODUCTION

Precipitation data are collected from many independent networks of rain gauges. The data have numerous applications in hydrology, architecture, agriculture, forensics, engineering, meteorological research and other fields. Very little documentation exists in the open literature regarding the conversion of autographic or semi-automated records to digital records, the validation of the data, and the consistency of observations. This paper describes the procedures used by the National Climatic Data Center to process hourly precipitation data. The treatment of instrument problems and the meteorological screening could be used in part if not in totality by other network data processors.

In January 1984 the National Climatic data Center instituted a new system to process hourly precipitation data. The old procedures were more labor intensive and costly. They evolved over many years into a patchwork system that did not provide optimum standardization or objectivity. The new system was designed to employ a mix of man and computer that would automatically identify and correct the more common instrument problems as well as bring to the attention of meteorological technicians inconsistencies in punched paper tape or weighing rain gauge chart annotation made by the observers. The new system also screens the data for meteorological inconsistencies.

Hourly precipitation data enter the system in one of four forms -- punched paper tape, autographic charts, manuscripts and digital records. Fischer-Porter rain gauges provide monthly cumulative rainfall data in the form of punched paper tape for about 2,000 stations. Universal gauges, which are gradually being replaced by the Fischer-Porter gauge, produce continuous ink traces of cumulative rainfall on a chart for about 600 stations. Manuscript forms are used by only a handful of stations. Digital data are created as a by-product from other processing systems for approximately 250 National Weather Service first order stations.

Weather observers are responsible for removing tapes and charts as prescribed by the National Weather Service Observing Handbook No. 2 in the 1972 edition (2). They are also responsible for annotating information such as station identification, time-on and time-off of the tapes and charts. An internal National Climatic Data Center study of one month's data showed that 40 percent of all tapes and charts have some problem that affects the data processing. Seventy percent of these problems are due to observer error. Gauge problems also exist. For the Fischer-Porter gauges typical problems are oscillations, bad punches, timer malfunctions, tape misalignment, multiple punching and dead batteries. Universal gauge problems include timer malfunctions and blurred traces.

The new processing system automatically screens the data according to preset rules. This technique eliminates much of the subjectivity that previously existed. It also allows more time to be spent on problem data instead of correct data.

2. NON-METEOROLOGICAL PHASES OF DATA VALIDATION

The first stage is an inventory and check-in function that keeps track of what data have been processed and of any previous problems. The information is annotated via interactive software for use by both the analyst and the automated processing system. The analyst indicates that a station's data has entered the processing system, verifies station identification information, and looks for observer error or mechanical problems. If necessary, the analyst will reject data that meet specified criteria. The criteria depend upon the type of data and are summarized in Table 1. The critical errors are those affecting the determination of the station, period of record, or date. They are critical because the data cannot be processed. Non-critical errors are noted interactively for follow-on analysis and correction; the data,

however, can be processed even though some of it may be deleted or modified during later stages.

The next stage is the data entry function in which Fischer-Porter tapes, Universal charts and manuscript forms are converted into digital form. During this stage, National Weather Service first order station hourly precipitation data are also entered into the system.

The processing system for the punched paper tapes automatically checks for consistency between the end time of the previous month's data and the beginning time of the current month's data. It is also designed to find timing problems such as slow clocks and erroneous observer notations. Adjustments are made so that timing errors will not propagate into subsequent months. Timing conditions that are automatically checked are shown in Figure 1. Panel a shows the condition when timing is correct. Panels b and c illustrate problems with the beginning time of a month while d and e show problems with the ending time of a month. Panels f through i indicate problems with both the beginning and ending time of a month. The checks assume that daylight savings time and recalibration/reset times have been taken into account. The adjustments to times are made in a systematic and objective manner and are retained in the data files created during the inventory and check-in stage.

Punched paper tape data are then filtered. Based on a test involving nearly 350 Fischer-Porter tapes for winter and summer, over 90 percent of the variations in data values resulting from gauge malfunctions fall into one of three categories: spikes, oscillations or bad values. A spike is generally defined as a drop or jump in the gauge value followed by a return to the previous value. Oscillations are pairs of values repeated one or more times with no other intervening values. Bad values are defined as a decrease or exceptionally large increase from one value to the next. Figures 2, 3 and 4 illustrate gauge malfunctions and the corrections that are described below.

Spikes are identified in a three step sequential screen. This approach is necessary because the tens, units and tenths punches operate independently to produce a value. The first step looks for and corrects spike in the tens digit. Letting t_i, u_i, and d_i represent the tens, units and tenths digits of the i-th gauge value v_i, the system performs the following operation.

$$\text{If} \quad |V_i - V_{i-1}| \geq 9.9 \qquad (1a)$$

$$\text{and} \quad |(u_i d_i) - (u_{i-1} d_{i-1})| \leq .1 \qquad (1b)$$

$$\text{then} \quad V_i = t_{i-1} u_i d_i \qquad (1c)$$

This step eliminates the spike by making the incorrect value consistent with the preceding and subsequent values.

The second step concerns one point spikes in the units or tenths digits.

$$\text{If} \quad |V_i - V_{i-1}| > .1 \qquad (2a)$$

$$\text{and} \quad |V_i - V_{i+1}| > .1 \qquad (2b)$$

$$\text{and either} \quad V_{i+1} = V_{i-1} \qquad (2c)$$

$$\text{or} \quad V_{i+1} = V_{i-1} + 1 \qquad (2d)$$

then v_i is deleted. This procedure operates similarly to the elimination of the tens digit spike. The third step looks for two point spikes in the units and tenths digits.

$$\text{When} \quad |V_i - V_{i-1}| > .1 \qquad (3a)$$

$$\text{and} \quad |V_{i+2} - V_{i+1}| > .1 \qquad (3b)$$

$$\text{and either} \quad V_{i+2} - V_{i-1} < .2 \qquad (3c)$$

$$\text{and} \quad V_{i+3} - V_{i+2} < .2 \qquad (3d)$$

$$\text{or} \quad V_{i-1} - V_{i-2} < .2 \qquad (3e)$$

$$\text{and} \quad V_{i+2} - V_{i-1} < .2 \qquad (3f)$$

then v_i and v_{i+1} are eliminated.

An oscillation is detected if

$$V_{i-2} < V_{i-1}\ , \qquad (4a)$$

$$V_{i-2} = V_i \qquad (4b)$$

$$\text{and} \quad V_{i-1} = V_{i+1}\ . \qquad (4c)$$

The data values that comprise an oscillation are determined by looking for equality of the first, third, fifth, etc., values and of the second, fourth, sixth, etc., values. After detection, the oscillation is eliminated.

The filtering of bad values is a two step process. The first step flags a value as suspicious if it is less than or equal to the previous value or if it is at least 3.0 higher than the previous value. The second step adjusts or deletes the values. As each value is examined, a correction term is subtracted from the value. The correction term, initially zero, is the difference between the previous suspicious value and the value before it. If the current value were originally flagged as suspicious, and if it remains suspicious after this adjustment, then it is deleted. The correction term acts as an adjustment to the baseline or reference point value.

Data entry from Universal rain gauges includes analysis and digitizing of the ink traces on the autographic charts. The analysis is a manual identification of the beginning and

ending times of precipitation events, of problems with chart readability and of consistency between successive charts. Annotations are made concerning error codes and periods for which data cannot be used. The charts are then digitized and stored at 15 minute resolution. The quality of the digitizing is checked by comparing the sum of the 15 minute values with the difference between the gauge values at the end and beginning of the chart.

Manuscript data are converted to digital records. These records are then checked by comparing daily totals on the manuscript against the computed daily sum of the digital hourly values. Digital data for National Weather Service first order stations are created and validated in other systems and are merged into the hourly precipitation data file without any additional non-meteorological data validation.

3. METEOROLOGICAL SCREENING OF DATA

The first step in the screening procedure is to identify suspicious hourly precipitation data by comparison with data from stations that are near the hourly data station being validated. These nearby stations are not, however, hourly stations themselves. They are part of the larger Climatological Data network of cooperative observers and National Weather Service observers. "Nearby" is defined as being located within a one degree latitude-longitude box centered on the hourly station and not differing from the elevation of the hourly station by more than 1,000 ft. The data from the closest nearby station in each quadrant centered on north, east, south and west is retained for comparison with the hourly station data. If a quadrant has no nearby station, then the next closest station's data in any quadrant, if available, is retained.

The retained data are daily precipitation amounts. Discontinuities exist, however, in the beginning times of a day. Hourly precipitation amounts are summed from midnight to midnight to create a daily total. The Climatological Data station amounts are often recorded for other 24-hour periods, such as from morning to morning or from evening to evening. Very often precipitation indicated in the nearby station record has fallen partially or entirely on the previous calendar day. Therefore, nearby station data must be compared to the hourly precipitation data on the same day as well as on the previous day.

In addition to nearby station data, daily precipitation totals are retained for any Climatological Data station that may be collocated with the hourly station. About 1,750 or the 3,000 hourly precipitation network stations are so collocated. For these stations more direct comparison between precipitation totals is possible than when only nearby data exist.

Comparison of hourly precipitation with collocated station data are designed to verify precipitation events. Validation of actual rainfall amounts is not possible because of differences of observation times between the two data sets. An inconsistency is said to exist, however, if on a given day the difference in precipitation amounts between the two sets is greater than .1 inch or if a two day difference is greater than .2 inch. Monthly total rainfall is also compared. If both the hourly data station and the collocated station reported precipitation during the month, and if the ratio of the hourly station total to the collocated station total is less than .8 or greater than 1.2, then the two data sets are said to be inconsistent. If either the hourly or the collocated station reported more than .3 inch of precipitation in the month, but the other station was dry, then the sets are flagged as being inconsistent.

When collocated stations do not exist or if the collocated station data are missing, then nearby stations are compared to the hourly stations. The occurrence of precipitation at the nearby stations on a given day coupled with no precipitation at the hourly station, or rainfall at the hourly station coupled with no rainfall at the nearby stations defines an inconsistency. The average of the monthly precipitation totals from the nearby stations is compared to the total monthly precipitation at the hourly station. If the two differ by more than 40 percent, then they are inconsistent. Also, if either the hourly station total or the nearby station average is more than .3 inch, but the other is zero, then an inconsistency is said to exist.

Another step in screening identifies suspect hourly precipitation data by value alone. The data will pass through the screen if any hourly value is not more than 1 inch, if any daily total is not more than 10 inches, or if any monthly total is not more than 15 inches.

The number of days with rain in a month and the daily and monthly precipitation amounts at about 2,000 hourly stations are compared to climatological values. The data are flagged if they fall outside the climatological values associated with the .05 and .95 probability levels.

A Poisson distribution, modified to account for persistence, is used to develop the number of rain days that can be expected at the specified probability levels. The density function, as described by Brooks and Carruthers (3), is

$$f(n) = \begin{cases} \dfrac{\prod_{k=1}^{n} \left[m + (n-k)f\right]}{n!\,(1+f)^{(m/f+n)}} & n \geq 1 \\ \dfrac{1}{(1+f)^{m/f}} & n = 0 \end{cases} \qquad (5)$$

where n is the number of rain days in a month, m is the mean number of rain days in the month, and f is the presistence factor minus one. The persistence factor is defined as the ratio of the variance to the mean. Sample mean number and variance of days of precipitation per month were computed from the historical records. If the number of days of precipitation in a given month is less than or greater than the climatological values corresponding to the .05 and .95 probability levels, respectively, then the data are flagged.

Daily and monthly precipitation amount climatologies were developed from the gamma distribution with probability density

$$f(r;\beta,\gamma) = \begin{cases} \frac{1}{\beta^{\gamma}\Gamma(\gamma)} r^{\gamma-1} \exp\left[-r/\beta\right] & r,\beta,\gamma > 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Where β is the scale parameter, γ is the shape parameter and Γ is the gamma function. From the cumulative form of the density function the amount of precipitation corresponding to a probability level can be determined. Maximum likelihood estimates of γ and β (4) were computed and debiased (5). Precipitation amounts corresponding to .95 probability level were then determined. Values that exceed this amount are deemed suspect. Monthly totals are also suspect if they are less than the amount corresponding to the .05 probability level.

Daily climatologies were created by applying the gamma distribution to the set of daily non-zero precipitation totals in a month over the historical period of record. Monthly climatologies were created in similar fashion.

Totals for a month over the historical period of record were modelled by the gamma distribution. Implicit assumptions in the comparisons of the hourly data with climatology are that the underlying meteorological conditions are stationary, for the month, over the historical period of record. Cycles and trends are assumed to be insignificant.

Computer identified inconsistencies or suspect data are reviewed by analysts. The analyst has at his disposal somewhat more information, such as the original chart or tape, than was available to the computer. He operates under the precept that flagged data are not necessarily invalid. Unless there is overwhelming evidence that invalidates a datum, he is encouraged to accept the value. The analyst recognizes that the screening criteria are arbitrary limits that are based more on statistics than atmospheric principles and therefore are to be used only as a guide to finding data problems or unusual events.

4. ESTIMATION OF TOTAL MONTHLY PRECIPITATION

In many cases missing data necessitates the estimation of monthly precipitation amounts. If an hourly station with less than 10 missing daily totals in a month is collocated with a Climatological Data station, then the total at the Climatological Data station is used to estimate the hourly precipitation data station monthly total, provided that the estimate is higher than the actual partial total at the hourly station. If the partial total is higher, then it becomes the estimate. The rational for this decision is that if the hourly precipitation data station with some missing data shows a larger total than a collocated station with complete data, then the missing hourly station records probably occurred when there was no precipitation. The monthly total is considered missing if the collocated station total falls outside the .95 probability climatological bound described in the previous section or if at least 10 hourly station daily totals are missing.

When a collocated station does not exist, but there are four nearby stations, the monthly total is estimated by a least squares regression. Four simultaneous equations of the form

$$Z_j = a_o + a_1 x_j + a_2 y_j \qquad j=1,4 \quad (7)$$

are solved to obtain the coefficients a_o, a_1, and a_2. The observed rainfall at station j is Z_j, and the distance and direction from the hourly station to the nearby station are represented in Cartesian coordinates by x_j and y_j. At the hourly station x and y are zero. The estimated precipitation is therefore the value of a_o. This procedure was proposed by Paulhus and Kohler (6).

If there are only two or three nearby stations, the estimated precipitation at the hourly station Z_e is a weighted average of the rainfall at the nearby stations j. Thus,

$$Z_e = \frac{\sum_j w_j Z_j}{\sum_j w_j} \qquad j = 2,3 \quad (8)$$

where w_j is the weight at station j. The weights are a function of the distance d from the nearby station to the hourly station such that

$$w_j = \begin{cases} \frac{30^2 - d_j^2}{30^2 + d_j^2} & 0 \le d_j \le 30 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

This function decreases from one at the hourly station to zero at a radius of 30 miles out from the hourly station.

When at least 10 days of data are missing at the hourly station, the modelled values are not computed and the monthly total is considered to be missing. The total is also missing if the modelled estimate is greater than the partial total at the hourly station and greater than the .95 climatological bound. If the modelled value is less than the partial total, then the partial total becomes the estimated total monthly precipitation.

If only one nearby and no collocated Climatological Data station data are available, then the estimated total monthly precipitation is the partial total at the hourly station. However, if at least 10 days are missing, the monthly total is considered to be missing.

The archived digital files and the publications of the hourly precipitation data distinguish between types of monthly totals. Missing values are coded as M, and estimated totals based on a full month of data are coded as E. Partial totals are denoted by I for incomplete.

5. CLIMATOLOGICAL IMPACT

The non-meteorological screening, which was designed to eliminate observer errors and mechanical problems, was tested on 104,226 hours of station data. The test compared the precipitation events determined under the new system with those data processed under the old system. The old system identified 2,588 hours with precipitation and 101,638 hours without precipitation. Only 89 percent of the hours with precipitation under the old system were said to have precipitation under the new system. Thus, 285 hours were considered dry by the screen but wet by the old procedures. Of the 101,638 hours determined to be dry by the old system, 99 percent were also considered dry by the screen. Therefore, 1,016 hours were changed from dry to wet by the screen. The net result is a gain of 731 precipitation hours under the new procedures.

Because the new procedures do not duplicate the old procedures, data sets derived from the two systems may be heterogeneous. Users of the data should be wary of making climatic inferences that result from mixed data sets rather than from physical causes. The heterogeneous data problem, however, could easily be tackled through appropriate statistical analysis and data adjustment. Once enough data become available from the new system, frequency distributions of the old and new data sets could be constructed and compared. Means, variances and other moments of the two distributions might also be compared. Time series analysis might also be appropriate.

Similar heterogeneities could exist in a data series of estimated monthly total precipitation. Modelled estimation techniques and the comparison with nearby station data were not previously used. Procedural changes may induce apparent but physically false climatic changes, but the problem with estimated monthly precipitation amounts is not nearly as severe as the non-meteorological procedural changes because estimated values are, by definition, suspect.

The meteorological screening should not create artificial climatic fluctuations. The procedures merely flag values that fall outside arbitrary, predetermined limits. These values are then manually analyzed for validity. Under the old system they would also have been manually analyzed.

6. CONCLUSIONS

It is recognized that changing a processing system may create heterogeneous data sets when users analyze data derived from differing systems. The advantages of the new methods of processing hourly precipitation data, however, far outweigh the heterogeneity problem. The primary benefit is that the new procedures introduce a level of consistency that did not previously exist. Automated steps are performed the same way each time according to be a prescribed set or rules. Subjective interpretation of the data validity, mechanical problems and observer practices is the undesirable alternative that previously existed. In the past this led to an inconsistency among analysts that was difficult to measure. Whether or not variances in judgment were random or systematic is a question that has not been addressed here. The mix of man and computer in the new system still allows judgmental decisions, but these decisions are now based on well defined guidelines.

A second benefit is that analysts now have more time to examine suspect data. Previously all data were manually screened; only suspect data are currently investigated. The automated processing of the bulk of the data is more economical in terms of time and money than the manual system.

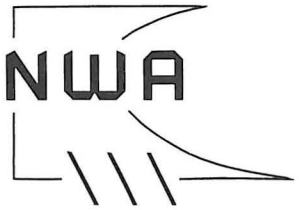
A third benefit is that data validation is based on a spatial comparison among nearby stations. Previously, the data were examined with little regard to meteorological conditions at surrounding stations. The new system not only looks at areal mesoscale weather events but also considers climatological expectations. The new system is therefore more palatable from a scientific viewpoint.

The end result of the new procedures is that the hourly precipitation data should be of higher quality than ever before. The users of the data, such as hydrologists, engineers, and planners, should be more confident that the

input to their analytical studies is a reasonably adequate representation of the precipitation events that occurred.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the developmental and programming work of Michael Mignono, Charles Phillips and Danny Fulbright. The suggestions of several staff members at the National Climatic Data Center are also acknowledged.



Error	Critical (C) or Non-Critical (NC)
1. Data not received in time for processing	C
2. Data cannot be identified	C
3. Data has invalid identification	C
4. Gauge type or status not consistent with station history	C
5. Data condition unacceptable (stained, shredded, illegible, etc.)	C
6. Beginning date/time cannot be determined or differs from previous month's ending time by more than 24 hours	C
7. Fischer-Porter tapes misaligned or guide holes not punched for entire tape	C or NC
8. Universal chart trace unreadable	C
9. Station name, number, date or time missing or incomplete	NC
10. Small time discrepancies between observer's notation, calculated time or previous charts and tapes	NC
11. Date or time missing on recalibration	NC
12. Illegible notations	NC
13. No line to indicate beginning of record	NC
14. Excessive Universal chart length (unable to distinguish individual precipitation events)	NC
15. Mechanical problems or gauge malfunctions	NC

Table 1. Inventory and check-in errors

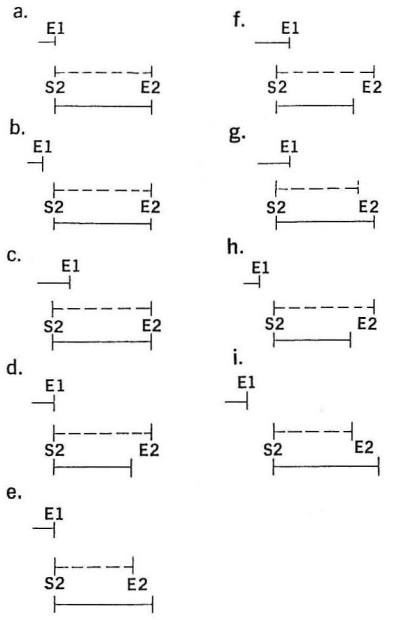


Figure 1. Computer checked timing conditions. E1 one is the ending time of the previous month, S2 is the starting time of the current month and E2 is the ending time of the current month. Dashed line indicates time deduced from observations, and solid line indicates time deduced from the edit procedures.

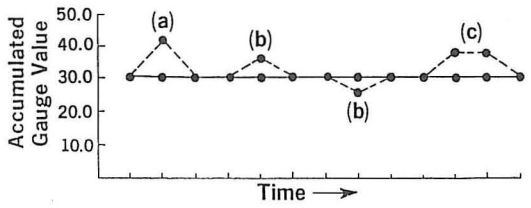


Figure 2. Spikes: a) one point spike in the tens digit, b) one point spike in the units or tenths digit, c) two point spike. Dashed line indicates original data series, and solid line indicates edited data series.

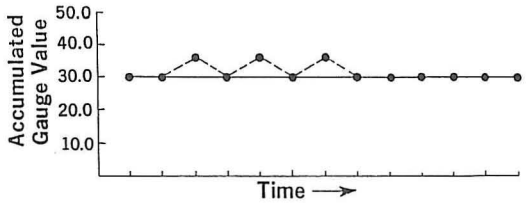


Figure 3. Oscillations. Dashed line indicates original data series, and solid line indicates edited data series.

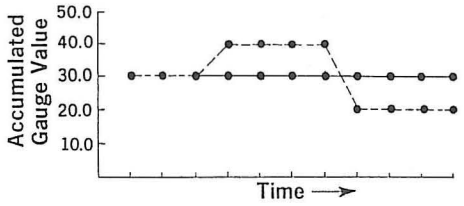


Figure 4. Bad values. Dashed line indicates original data series, and solid line indicates edited data series.

FOOTNOTES AND REFERENCES

1. Nathaniel B. Guttman received his B.A. in Meteorology and Oceanography from New York University and his M. Stat. in Statistics and Ph.D in Marine Sciences from North Carolina State University. He has worked as a Climatologist at the National Climatic Data Center and was a Visiting Research Professor in Meteorology at the U.S. Naval Academy. He has also taught at the University of North Carolina-Asheville, Mars Hill College and Western Carolina University. He serves as a member of the AMS Committee on Applied Climatology.

2. National Weather Service, 1972: Observing Handbook No. 2 (rev.), Natl. Ocean. Atmos. Admin., Washington, DC, 77 pp.

3. Brooks, C. E. P. and N. Carruthers, 1953: Handbook of Statistical Methods in Meteorology. Her Majesty's Stationery Office, London, pp. 315-318.

4. Thom, H. C. S. 1958: A note on the gamma distribution. Mon. Wea. Rev., 86, 117-122.

5. Crutcher, H. L. and R. L. Joiner, 1978: Gamma distribution bias and confidence limits. NOAA Tech. Rep. EDIS 30, Asheville, NC, 51 pp plus appendix.

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