Nordic Hydrology 4, 1973, 17–27
Published by Munksgaard, Copenhagen, Denmark

FILLING IN GAPS IN RAINFALL RECORDS BY SIMULATED DATA

B. SAMUELSSON

UNDP – FAO, Iraklion, Greece

The concept of the rainfall coincidence rate is introduced and is used as a tool for correcting the errors in the rainfall records which are due to shifts. A method of simulating missing rainfall data without smoothing out the rainfall distribution pattern is presented. Dates are allocated to the generated raindays according to a procedure which takes into account the coincidence rate as well as the occurrence of rainfall over groups of consecutive days.

For any kind of water resources study, and particularly when applying a watershed model, the rainfall data should, as far as possible, be homogenous with regard to both time and space. This necessitates records over a common, reasonably long, period from stations which are well distributed over the study area and which are located so as to produce a representative picture of both the macro-climatological characteristics and those micro-climatological elements which influence the areal rainfall distribution.

To proceed from a collection of observers' notes, of varying degrees of accuracy, to computed rainfall depth area values is a long and cumbersome task, and there is need for a method of shortening and simplifying this task, while making sure, at the same time, that an objective procedure, which makes the best possible use of the original data, is not replaced by mere personal judgement, which is always somewhat fallible.

This paper describes a step-by-step procedure according to which the minimum of data handling is done manually, the bulk of this time-consuming and tedious work being executed by a computer.

CHECKING DATA FOR SHIFTS

Partly due to different observation times, and partly because it is known from experience that some observers more or less consistently refer their observations back to the previous day, it is necessary to try to trace and correct shifting errors. As a choice between the two extremes, i. e. checking individual storms or periods of convective activity, or not checking at all, a fair compromise could be to check on a year-by-year basis. Usually this could be regarded as sufficient because, in general, mistakes in reporting dates are consistent.

As a mathematical tool in this checking procedure the coincidence rate, c, can be used (Samuelsson 1971):

$$c = \frac{2C}{N_1 + N_2} \tag{1}$$

where C is the number of days of rainfall coinciding at the two stations, index 1 and 2, during a certain period of time; N_1 and N_2 are the total number of raindays at stations 1 and 2, respectively, during the same period.

The computation of coincidence rates is then carried out in two steps, in which all rainfall records are, in principle, geared to those of any not too distant first-class meteorological station which is climatologically related to the study. First, a set of control stations are selected (second-class long-term climatological stations) and the records from each of them compared with those from the first-class station. As a second step, all the remaining station records are compared with those of the nearest control station. For each pair of stations thus formed, the coincidence rate is computed with the records from the stations to be checked, at first unchanged and then shifted forward by one day.

In cases where the shifting results in an increase of the coincidence rate. the shifted parts of the records are regarded as correct and are repunched.

GENERATION OF MISSING RAINFALL DATA

Yearly and monthly totals

There are many ways by which missing yearly and monthly totals of rainfall can be estimated, the most common probably being to estimate missing yearly totals from annual isohyetal maps and to compute the missing monthly totals from weighted monthly totals from three or four neighbouring stations. This method shares the disadvantage of other similar methods as regards the smoothing out of the rainfall pattern. Sometimes it may be considered necessary to

preserve, as far as possible, the irregularities in the rainfall area distribution pattern.

For this purpose, a combined linear regression analysis and randomization method is introduced. The coefficients A_y and B_y, in the regression equation.

$$y = A_y + B_y\, x \tag{2}$$

where x and y represent total monthly rainfalls, is obtained for every month by selecting a station from 3–4 neighbouring long-term stations to act as control station. The selection is made so that, for each month, the station which gives the highest value of the correlation coefficient is chosen as control station.

Next, the y-values is recomputed from (2) for overlapping periods and the standard deviation, $S\Delta_y$, of the residuals, Δ_y is determined.

The total monthly rainfall to be generated, y, is then computed from the expression

$$y = A_y + B_y\, x + K \cdot S\Delta_y \tag{3}$$

where K is a frequency factor dependent only on probability and type of distribution of Δ_y.

By generating a random number between 0 and 1 with the same distribution as Δ_y, the corresponding random value of K can be determined from the cumulative probability expression for the identified distribution, and this value is then used for computing a certain total monthly rainfall value from (3).

The process of generating a random value of K, which has to be repeated for every y-value to be computed, is for most applicable distributions available as a standard sub-routine for at least large and medium-sized computers.

The simulated total yearly rainfalls are obtained simply by adding the computed monthly totals.

Daily rainfalls

The same combined linear regression analysis and randomization method described above is applied to generate, month by month and for each specified class (interval of daily amount of rainfall), the total number of raindays, as well as the number of days of rainfall coinciding with those at the control station. Using a random number generator with rectangular distribution, the amount of rainfall for each rainday is simulated and subsequently adjusted in order to keep the total monthly rainfall unchanged. The allocation of dates, finally, to the raindays is accomplished by a logical procedure, taking into account the occurrence of rain in groups of consecutive days (passing of cyclones or unstable air masses).

The Appendix gives a detailed description of the procedure step by step; its programming into any computer language would be a straightforward matter.

CONCLUSIONS

The philosophy behind the generation of rainfall data, as described above, is that historic rainfall records will never be repeated, but that their statistical characteristics are preserved.

As was pointed out by the writer (Samuelsson 1972), there are meteorological reasons to believe that extreme rainfall events, caused by abnormalities in the general atmospheric circulation pattern, have statistical properties different from those of less extreme rainfall events. The generation of rainfall data should be based, therefore, on historical records over a long enough period of time to include a sufficient number of extreme synoptic situations, in order to provide an acceptable overall statistical representation.

APPENDIX

GENERATION OF DAILY RAINFALLS

The computations should be executed month by month, and the same pair of stations should be used as in the generation of monthly rainfall figures. Rainfall figures should be rounded off to full mm. For stations the records of which are to be generated, index "x" refers to existing short-term records, index "y" refers to be records to be generated. Control stations: without x- or y-index

M = monthly rainfall
D = daily rainfall
N = number of raindays during the month
C = number of raindays common to the pair of stations during the month

$$c = \text{coincidence rate} = \frac{2C}{N + N_x} \quad \text{or} \quad \frac{2C}{N + N_y}$$

S = standard deviation
K = frequency factor.

The following rainfall classes should be used:

$D \geqq 1$ mm	Class 1
$D \geqq 5$ mm	Class 2
$D \geqq 10$ mm	Class 3
$D \geqq 20$ mm	Class 4
$D \geqq 30$ mm	Class 5
$D \geqq 40$ mm	Class 6
$D \geqq 50$ mm	Class 7
$D \geqq 75$ mm	Class 8
$D \geqq 100$ mm	Class 9
$D \geqq 150$ mm	Class 10
$D \geqq 200$ mm	Class 11

1. For the period of the short-term record find:
$N_1\ N_2 \ldots, N_{11}, N_{x1}, N_{x2}, \ldots N_{x11}, C_1, C_2, \ldots C_{11}, c_1, c_2 \ldots c_{11}, c_1, c_2 \ldots c_{11}, S_{c1}, S_{c2} \ldots S_{c11}$.

2. Determine coefficient A_j and B_j in regression equations

$$N_{xj} \equiv A_j + B_j N_j \quad (j = 1, 2, \ldots 11) \tag{1}$$

3. Evaluate ΔN_{xj} = the difference between observed and computed (1) number of raindays; compute standard deviation, $S(\Delta N_{xj})$.

4. Compute for period with records from control station only:

$$N_{yj} \equiv A_j + B_j N_j + KS\ \Delta_{Nxj} \tag{2}$$

$$c_{yj} \equiv \bar{c}_j + KS_{cj} \tag{3}$$

$$C_{yj} \equiv \frac{c_{yj}}{2}(N_{yj} + N_j) \tag{4}$$

5. Allocate daily rainfalls, D_{kj}, to the number of raindays $(N_{yj} - N_{y(j+1)})$ by random numbers with rectangular distribution. $(\Sigma k_j = N_{yj};\ N_{y(j+1)} = 0$ for $j = 11)$

6. Compute

$$\delta_j = M_y - \sum_{}^{j}\ \sum_{}^{k} D_{kj}$$

7. Examine δ_{11}

7.1 $\delta_{11} < 0$:

Assume that $\delta_n > 0$ and $\delta_{n+1} < 0$. Then distribute

$$\delta_{n+1} = M_y - \sum_{j=1}^{n+1}\ \sum_{}^{k} D_{kj}$$

proportionally to D_{kj} over the total number of raindays, N_{yl}.

7.2 $\delta_{11} > 0$:

7.2.1 $N_{y11} > 0$

Then distribute

$$\delta_{11} = M_y - \sum_{j}^{} \sum_{k}^{} D_{kj}$$

in proportion to D_{kj} over the total number of raindays, N_{yl}

7.2.2 $N_{y11} = 0$

Assume that $N_{y(n+1)} = N_{y(n+2)} = \ldots = N_{y11} = 0$.

Then generate a number, $N_{y(n+1)}$, so that $N_{y(n+1)} \leqq N_{yn}$ and allocate daily rainfalls, $D_{K(n+1)}$, to the $\sum K_{(n+1)} = N_{y(n+1)}$ number of raindays. If $\delta_{n+1} > 0$ then generate a new number $N_{y(n+2)} < N_{y(n+1)}$ and allocate daily rainfalls $D_{k(n+2)}$, etc.

8. Allocation of dates to the N_y raindays

8.1 The N_{y1} raindays

8.1.1 Half or less of the dates of the N_1 raindays are consecutive

8.1.1.1 $C_{y1} \geqq N_{y1}$

8.1.1.1.1 $N_1 \geqq N_{y1}$

Distribute the N_{y1} days randomly among the dates of N_1

8.1.1.1.2 $N_1 < N_{y1}$

Allocate the N_1 dates and distribute the remainder, $(N_{y1} - N_1)$ days, randomly among the remaining dates of the month.

8.1.1.2 $C_{y1} < N_{y1}$

8.1.1.2.1 $N_1 \geqq N_{y1}$

Select randomly C_{yl} days from the N_{yl} days. Distribute these C_{yl} days randomly among the dates of N_1.

Distribute the remainder, $(N_{yl} - C_{yl})$ days, randomly among the dates left over after the N_1 dates have been extracted.

8.1.1.2.2 $N_1 < N_{yl}$

8.1.1.2.2.1 $C_y1 \geqq N_1$

Allocate the N_1 dates and distribute the remainder, $(N_{yl} - N_1)$ days, randomly among the remaining dates of the month.

8.1.1.2.2.2 $C_{y1} < N_1$

Select randomly C_{y1} days from the N_{y1} days. Distribute these C_{y1} days randomly among the dates of N_1.

Distribute the remainder, $(N_{y1} - C_{y1})$ days, randomly among the remaining dates of the month.

8.1.2 More than half of the dates of N_1 are consecutive, the ratio being $= f_1$ $(1/2 < f_1 \leqq 1)$ and the number of groups of consecutive days being g1.

8.1.2.1 $C_{y1} \geqq N_{y1}$

8.1.2.1.1 $N_1 \geqq N_{y1}$

8.1.2.1.1.1 $2_{g1} \leqq (f_1 N_{y1})$

8.1.2.1.1.1.1 $g_1 = 1$

Allocate randomly one of the $(f_1 N_{y1})$ days to one of the dates of the group.

Select randomly a second of the $(f_1 N_{y1})$ days and allocate this day to a concutive date within the group.

Continue until all the $(f_1 N_{y1})$ days have been allocated to dates within the group.

Distribute the remainder, $(N_{y1} - f_1 N_{y1})$ days, randomly among the remaining N_1 dates.

8.1.2.1.1.1.2 $g_1 > 1$

Allocate randomly one of the $(f_1 N_{y1})$ days to one of the dates of the largest of the g_1 groups.

Select randomly a second of the $(f_1 N_{y1})$ days and allocate this day to a consecutive date within that group.

Select randomly a third of the $(f_1 N_{y1})$ days and allocate this day to one of the dates of the second largest of the g_1 groups.

Select randomly a fourth of the $(f_1 N_{y1})$ days and allocate this day to a consecutive date within that group.

Continue until two of the $(f_1 N_1)$ days have been allocated to consecutive dates within each of the g_1 groups.

Repeat the process starting with the largest of the g_1 groups and allocate consecutive dates within each group.

Continue until all the $(f_1 N_{y1})$ days have been allocated to dates.

Distribute the remainder, $(N_{y1} - f_1 N_{y1})$ days, randomly among the remaining N_1 dates.

8.1.2.1.1.2 $2g_1 > (f_1 N_{y1})$ (in this case $g_1 > 1$, because if $g_1 = 1$ then $(f_1 N_{y1}) \leqq 1$)

8.1.2.1.1.2.1 (f_1N_{y1}) is an even number

Proceed according to 8.1.2.1.1.1.2

8.1.2.1.1.2.2 (f_1N_{y1}) is an odd number

Proceed according to 8.1.2.1.1.1.2 until one of the (f_1N_{y1}) days is left

Allocate this day to a consecutive date within the largest of the g_1 groups.

Continue according to 8.1.2.1.1.1.2

8.1.2.1.2 $N_1 < N_{y1}$

Allocate the N_1 dates.

Select randomly one of the remaining ($N_{y1} - N_1$) days and allocate this day to a date next to the largest of the g_1 groups.

Select randomly a second of the remaining days and allocate this day to a date next to the second largest of the g_1 groups.

Continue and, if necessary, repeat the process until $f_1(N_{y1} - N_1)$ days have been allocated to dates.

Distribute randomly the remaining $N_{y1} - f_1(N_{y1} - N_1)$ days among the remaining dates of the month.

8.1.2.2 $C_{y1} < N_{y1}$

8.1.2.2.1 $N_1 \geqq N_{y1}$

Select randomly C_{y1} days from the N_{y1} days.

Proceed according to 8.1.2.1.1 ((f_1C_{y1}) instead of (f_1N_{y1}) days to be allocated to dates).

Distribute the remaining days, ($C_1 - f_1C_1$), randomly among the remaining N_1 dates.

Select randomly of the remaining ($N_{y1} - C_{y1}$) days and allocate the day to a date among those remaining after the dates of the N_1 days have been extracted.

Selected randomly a second of the remaining days and allocate this day to a consecutive date, not included in the dates of the N_1 days.

Continue until $f_1(N_{y1} - C_{y1})$ days have been allocated to dates.

Distribute randomly the remaining $N_{y1} - f_1(N_{y1} - C_{y1})$ days among the remaining dates of the month.

8.1.2.2.2 $N_1 < N_{y1}$

8.1.2.2.2.1 $C_{y1} \geqq N_1$

Proceed according to 8.1.2.1.2

8.1.2.2.2.2 $C_{y1} < N_1$

Proceed according to 8.1.2.2.1

8.2 The N_{yn} raindays (n = 2, 3, . . . 11)

Let $N_{c(n-1)}$ be the days representing coinciding dates of the $N_{y(n-1)}$ and the N_n days.

8.2.1 Half or less of the dates of the N_n raindays are consecutive

8.2.1.1 $C_{yn} \geqq N_{yn}$

8.2.1.1.1 $N_n \geqq N_{yn}$

8.2.1.1.1.1 $C_{y(n-1)} \geqq C_{yn}$

Distribute the N_{yn} days randomly among the coinciding dates of $N_{y(n-1)}$ and N_n

8.2.1.1.1.2 $C_{y(n-1)} < yn$

Allocate the coinciding dates of $N_{y(n-1)}$ and N_n, and distribute the remainder randomly among the remaining $N_{y(n-1)}$ dates.

8.2.1.1.2 $N_n < N_{yn}$

Same as 8.2.1.1.1.2

8.2.1.2 $C_{yn} < N_{yn}$

8.2.1.2.1 $N_n \geqq N_{yn}$

8.2.1.2.1.1 $C_{y(n+1)} \geqq C_{yn}$

Select randomly C_{yn} days from the N_{yn} days.

Distribute these C_{yn} days randomly among the $N_{c(n-1)}$ dates.

Distribute the remainder, $(N_{yn} - C_{yn})$ days, randomly among the remaining dates of $N_{y(n-1)}$ not coinciding with the dates of N_n.

8.2.1.2.1.2 $C_{y(n-1)} < C_{yn}$

Proceed according to 8.2.1.1.1.2

8.2.1.2.2 $N_n < N_{yn}$

8.2.1.2.2.1 $C_{yn} \geqq N_{c(n-1)}$

Proceed according to 8.2.1.1.1.2

8.2.1.2.2.2 $C_{yn} < N_{c(n-1)}$

Proceed according to 8.2.1.2.1.1

8.2.2 More than half of the dates of the N_n days are consecutive, the ratio being $= f_n$ $(1/2 < f_n \leqq 1)$ and the number of groups of consecutive days being g_n.

8.2.2.1 $C_{yn} \geqq N_{yn}$

8.2.2.1.1 $N_n \geqq N_{yn}$

8.2.2.1.1.1 $C_{y(n-1)} \geqq C_{yn}$

Proceed according to 8.1.2.1.1 by substituting index 1 by index n except in the case of N_1 which should be substituted by $N_{c(n-1)}$.

8.2.2.1.1.2 $C_{y(n-1)} < C_{yn}$

Allocate the $N_{c(n-1)}$ dates.

Select randomly one of the remaining $(N_{yn} - N_{c(n-1)})$ days and allocate this day to one of the remaining $N_{y(n-1)}$ dates and so that it comes next to the largest of the g_n groups.

Select randomly a second of the remaining days and allocate this day to one of the remaining $N_{y(n-1)}$ dates and so that it comes next to the second largest of the g_n groups.

Continue and, if necessary, repeat the process until $f_n(N_{yn} - N_{c(n1-)})$ days have been allocated to dates.

Distribute randomly the remaining $N_{yn} - f_n$ $(N_{yn} - N_{c(n1-)})$ days among the remaining $N_{y(n-1)}$ dates.

8.2.2.1.2 $N_n < N_{yn}$

Proceed according to 8.2.2.1.1.2

8.2.2.2 $C_{yn} < N_{yn}$

8.2.2.2.1 $N_n \geqq N_{yn}$

8.2.2.2.1.1 $C_{y(n-1)} \geqq C_{yn}$

Proceed according to 8.1.2.2.1 by substituting index 1 by index n and so that the N_{yn} days are allocated to dates among those allocated for the $N_{y(n-1)}$ days.

8.2.2.2.1.2 $C_{y(n-1)} < C_{yn}$

Proceed according to 8.2.2.1.1.2

8.2.2.2.2 $N_n < N_{yn}$

8.2.2.2.2.1 $C_{yn} \geqq N_{c(n-1)}$

Proceed according to 8.2.2.1.1.2

8.2.2.2.2.2 $C_{yn} < N_{c(n-1)}$

Proceed according to 8.2.2.2.1.1

REFERENCES

Samuelsson, B. (1971) Surveys, demonstration and planning of water resources utilization. Technical Report 1, UNDP/FAO project CYP 6. FAO, Rome 1971.

Samuelsson, B. (1972) Statistical interpretation of hydrometeorological extreme values. *Nordic Hydrology 3* (4).

Received 21 June 1972.

Address:

Mr. B. Samuelsson,
c/o AGLO,
FAO,
I-00100 Rome, Italy.