GENERATION OF MONTHLY STREAM-FLOW DATA

F. H. Pearson*

ABSTRACT

Usually the length of a stream-flow record is insufficient to contain extreme values and critical sequences required for the design of water-resource system components. Data are sometimes required for an ungauged stream.

The Thomas-Fiering model may be used for the generation of a synthetic record of mean monthly stream flow of any required length. Data may be obtained by analysis of a shorter observed record for the stream, by analysis of regional characteristics of stream flow, or by analysis of the rainfall excess and consideration of storage in the catchment.

Different considerations apply to the generation of synthetic flows for low-flow studies than for other types of problem. The extent to which the characteristics of an observed flow record will be reproduced in a synthetic record generated from the observed record is discussed.

INTRODUCTION

In any area of planned human activity, the acquisition of data is a prerequisite to planning. Most types of data for planning, such as population data, consist of current and historical values of an item — each value being tagged with a date or some other measure of an independent variable. These values are used as a basis for predicting a value of the item at some future date.

Hydrological data are, however, essentially historical; that is, observed in the past. The value of the record is directly related to its length. The aim in the collection of data is generally to determine the form of the relationship of a dependent variable with an independent variable, for instance the dependence of stream flow on time. Often the relationship is not fixed but contains an element of variation which can be described by a probability distribution.

Having obtained the probability distribution of the observed events, the hydrologist is able to make an estimate of future events and from this attempt to solve his particular problem.

* Civil Division, Ministry of Works, Wellington.

One type of problem is that of estimating a future value of some stream-flow characteristic such as peak flood or low flow. The difficulty common to all problems of this type is that the period of record is usually short in comparison with the period required for a reliable direct estimate. For instance, Wisler and Brater (1959) suggested that 10 independent samples should provide a satisfactory determination of the size of flood that may be expected to occur with any given frequency, although a greater number would increase the accuracy. Usually the period of flow gauging is much less than the design period.

However, many natural phenomena including stream flow appear to conform to a continuous statistical distribution. The assumption that such a distribution exists enables the phenomena to be predicted on the basis of limited historical data.

The mathematical model used in this paper to generate synthetic stream-flow values will consist of a fixed effect and a random component, the latter being obtained from the assumed distribution. This distribution will contain parameters that can be estimated by stream-flow characteristics, such as the average flow, the variability of the flow, and the persistence of the flow. The three statistics thus obtained will determine the fixed effect of the model and the behaviour of the random component, or random noise.

The synthetic flow record will conform to the mathematical model that was selected to represent the observed flow record. Insofar as the statistics truly represent the observed flow record, the synthetic flow record also represents both the statistics and the observed flow record. Statistically, the synthetic flow record is indistinguishable from the historical flow record.

A synthetic record many times longer than the observed record may be generated and will provide a simple picture of possible stream-flow patterns which are not apparent in the historical record. In this sense the value of the synthetic record will be directly related to its length.

GENERATION OF A SYNTHETIC FLOW RECORD

A computer programme called GENSYN has been written for the generation of a synthetic record of mean monthly flow by the method of Thomas and Fiering (1962). The following statistics of flow for each month are required:

(1) mean,
(2) standard deviation,
(3) skew,
(4) regression coefficient,
(5) correlation coefficient.

Regression and correlation coefficients are for the regression of one month's flow on the next month's.

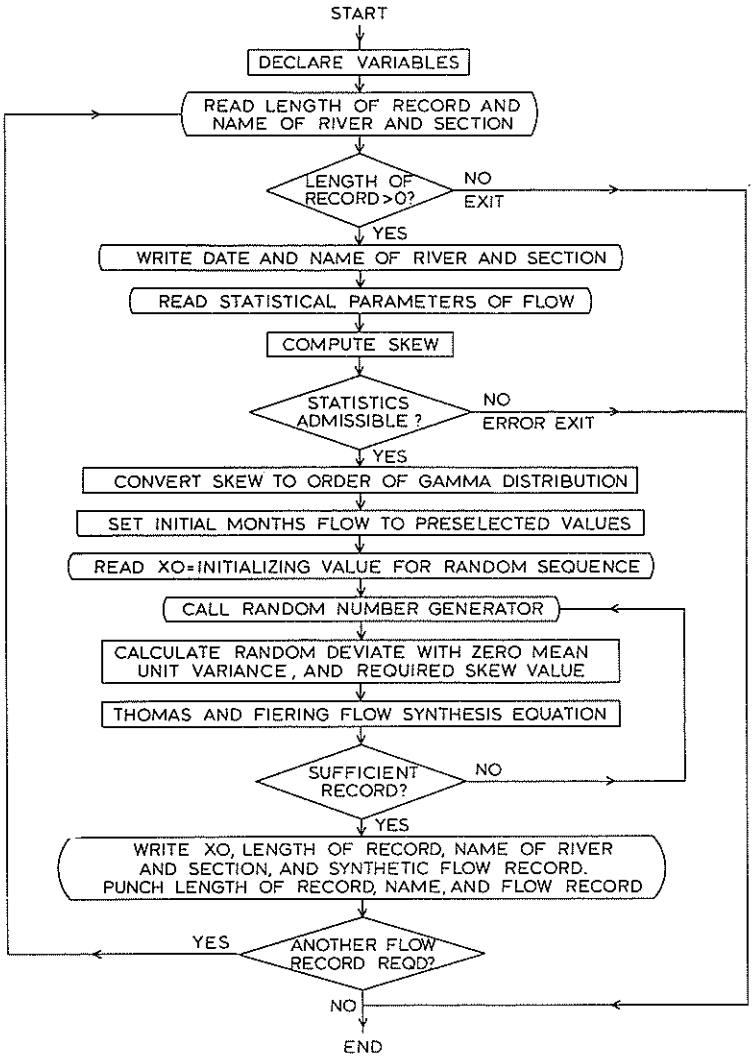

FIG. 1 — Simplified flow diagram for programme GENSYN.

The Thomas-Fiering equation is:

$$q_{t+1}=\bar{q}_{t+1}+b_t(q_t-\bar{q}_t)+s_{t+1}I_{t+1}(1-r_t^2)^{\frac{1}{2}}$$

where q_{t+1} and q_t are synthetic mean flows for months $t+1$ and t,
$\bar{q}_{t+1}$ and $\bar{q}_t$ are observed mean monthly flows for months $t+1$ and t,
b_t is the observed regression coefficient of q_t on q_{t+1},

s_{t+1} is the observed standard deviation of flow for month $t+1$,

I_{t+1} is the value of a random deviate at $t+1$ (I_{t+1} has zero mean and unit variance),

r_t is the observed correlation coefficient between q_t and q_{t+1}.

Statistics of monthly flow of the river for which a record is to be generated may be obtained by one of three methods, given in decreasing order of probable reliability:

(1) by analysis of an observed flow record for the river,
(2) by analysis of similar rivers in the region,
(3) by analysis of the monthly water balance.

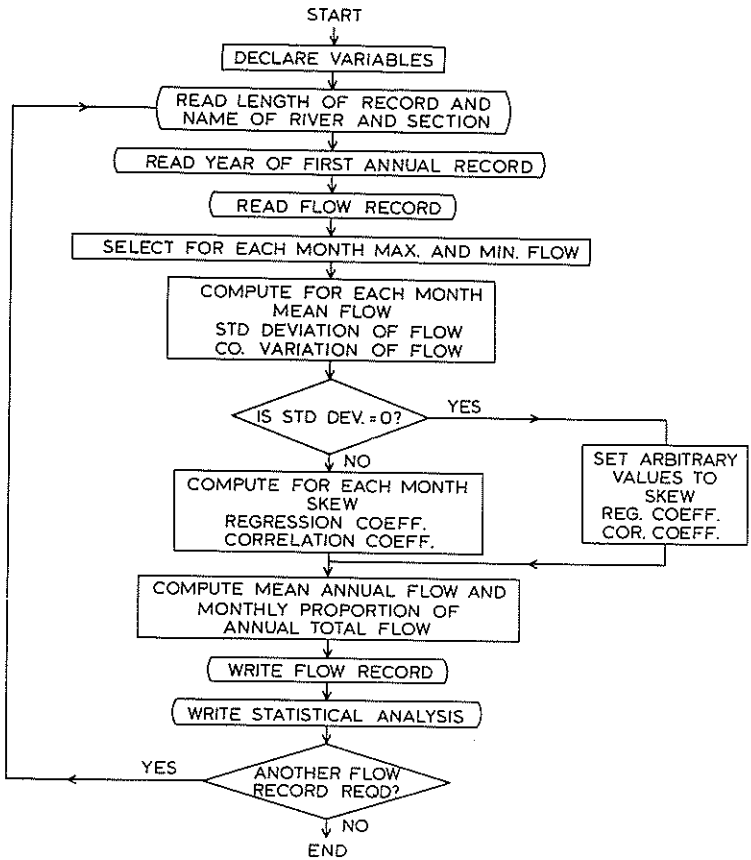

FIG. 2 — Simplified flow diagram for programme DASTAT.

Analysis of Observed Flow Records

Analysis of an observed monthly flow record may be performed by a computer programme called DASTAT. For each month DASTAT selects or calculates the following statistics of the flow record:

(1) mean,
(2) monthly proportion of annual total,
(3) standard deviation,
(4) coefficient of variation,
(5) coefficient of skew,
(6) regression coefficient,
(7) correlation coefficient,
(8) maximum and minimum.

Mean annual flow and annual mean flows are also calculated.

For the first six of the above the standard error of the estimate and the two-thirds confidence limits are calculated. Confidence limits are printed out as "value plus standard error" and "value minus standard error". Since the distribution of the correlation coefficient is very skewed, especially near unity, it is not normal practice to compute its standard error (Weatherburn, 1957), but two-thirds confidence limits are computed.

Calculation of the statistics is described by Weatherburn (1957), Crow *et al.* (1960), and Matalas and Benson (1968).

Regional Analysis

Regional analysis is a valid approach provided that the region which contains the catchment and other gauged catchments is both meteorologically and topographically homogeneous. Benson and Matalas (1967) discuss this approach.

The author has calculated parameters for each of five gauged rivers within a 60-mile-diameter region in the English Midlands (Pearson, 1967). It was possible to select a set of regional parameters which adequately described all five rivers. Regional parameters were used to generate a synthetic flow record for an ungauged river in the region.

Analysis of Water Balance

A water-balance approach may be adequate for some preliminary studies. No gaugings would be required; the only information needed is rainfall and evaporation data and some fairly superficial knowledge of the catchment to enable deduction of its frozen storage, surface storage, and ground-water storage properties by comparison with other gauged catchments. The approach is fairly obvious and is not detailed here because the exact procedure will depend upon the quality of the results required.

When a complete set of statistical data has been calculated or selected, each parameter can be plotted over a 12-month period. If two-thirds confidence limits are also plotted, a smooth curve can usually be obtained such that this curve lies statistically within the confidence bounds of the monthly values. The statistical test is that about two-thirds of the ordinates of the smoothed curve should pass between the two-thirds confidence limits of the original plot. Smoothing can be done either by eye or by a smoothing equation which weighs each value to a predetermined extent with its neighbour, or neighbours, on each side.

The justification for smoothing is that it appears reasonable to expect some degree of continuity of the characteristics of flow from month to month. Erratic changes probably arise from the statistical inadequacy of the flow data in a short-term record. The reason for smoothing is to make use of this assumed continuity to improve the quality of the estimates of the statistical parameters of flow.

A synthetic record of monthly flow can be generated from data comprising monthly values of each of the five parameters and the length of the required record. If the synthetic record is analysed, the statistics of the analysis will have expected values equal to the respective values in the data.

A discussion of the skew is relevant. Values for the coefficient of skew depend very largely on the frequency of high versus low flows observed for each month, and in turn influence the magnitude of extreme flows in the synthetic record.

For each month, synthetic flow is distributed as the skewed gamma distribution, the order of gamma being selected for the required degree of skew. Matalas (1967) describes the mathematics of the use of gamma for this purpose.

Since the magnitude of the coefficient of skew depends upon the magnitude of extreme flow, its value will necessarily be subject to a large standard error because the recurrence interval of extreme events cannot be determined with certainty from a short record. Further, for each month, the value of the observed skew coefficient is usually more strongly influenced by higher than lower extreme flows. Consequently, low flows may be poorly reproduced in the synthetic record.

For low-flow studies it may be unsatisfactory to select a set of values for the skew coefficient from the values of an observed record. This is supported by the fact that, if skew values from the observed record are used, negative flows may be generated — which is physically impossible.

The alternative, for low-flow studies, is to select values of the skew coefficient determined by the low-flow characteristics of the observed record. A simple expression can be derived for the value

of the skew coefficient which will fit the minimum flows of the synthetic record to minimum-flow values determined from the observed record. The expression is derived in Appendix 1.

In this application, minimum flow is determined as the asymtopic value of the Type-III distribution of Gumbel (1958) and can be calculated for each month by well-known graphical methods.

For problems other than low-flow studies the user may decide to adopt observed skew values for generation of the synthetic record. This appears to be a suitable procedure for most storage studies.

It is possible to use skew values based mainly upon the observed record, only modifying those that would otherwise give rise to zero or negative synthetic flows. This may be done by imposing the condition that the coefficient of skew should not be less than a certain value which is related to other flow statistics. The derivation of this relation is given in Appendix 1. If this condition is satisfied, zero or negative flows should not appear in the synthetic record.

If the objective is to determine the range of storage required to maintain a certain regulated draw-off, the actual production of a flow record may not be necessary; a satisfactory answer may be available with less computation from range theory, as described by Yevjevich (1967).

QUALITY OF THE SYNTHETIC FLOW RECORD

A synthetic flow record is only as good as the flow statistics used to produce it. A user of a synthetic flow record should be satisfied that the statistics comply with the following conditions:

(1) they should be adequately defined,
(2) they should describe the characteristics of flow relevant to the purpose to which the synthetic record is to be put,
(3) they should be modified if necessary to account for any future changes in the hydrological properties of the catchment.

The first two conditions will be considered further.

Adequate Definition

If the statistics of an observed record are not adequately defined, the synthetic flow record may not adequately represent the behaviour of the river. Quantitatively, expected errors in the synthetic flow record can be estimated by the magnitude of the standard error of each value of each statistic, relative to the magnitude of the statistic itself.

For this reason mean monthly flows rather than mean daily flows are generated, since daily flows for a given date will show more variation from year to year than corresponding mean monthly flows. For example, the standard error of mean flow on the first day of January will be greater than the standard error of the mean January flow.

Relevance of Described Characteristics

It is useful to examine the physical significance of each of the statistics in order that the user can satisfy himself that the synthetic flow record should reproduce those characteristics of the historical record which are of interest.

Mean flow is related to the annual pattern of rainfall excess modified by changes in various types of storage.

The non-dimensional form of the standard deviation, namely the coefficient of variation, is the most convenient form for discussion. Coefficient of variation is related to the reliability of flow; it is low when mean monthly flow varies little from year to year, which is likely to occur for those months where monthly mean flow changes little from one month to the next. Conversely, variable flow can be expected for the times of the year when monthly mean flow is changing most rapidly.

The coefficient of skew is related to the occurrence of occasional extreme flows. Its absolute value is likely to be highest either at the time of the year when flooding is most frequent or, less often, when occasional droughts occur. The coefficient of skew cannot model both types of extreme simultaneously and will usually be determined by the high monthly flows.

The regression between one month's flow and the next may be explained by ground-water storage (Harms and Campbell, 1967) and other forms of storage. Some parts of the regression may be due to persistency of rainfall. The regression coefficient is likely to be highest for catchments where changes in storage are a significant proportion of monthly flow or where climatic conditions favour persistency of rainfall. High coefficients will occur for those months of the year when changes of storage are greatest or when rainfall tends to be most persistent from month to month.

The correlation coefficient measures the strength of the regression, and the annual pattern of the correlation coefficient is likely to be similar to that of the regression coefficient. This can be seen by examination of the equations for both coefficients and from the observation that the regression coefficient is unlikely to change erratically from month to month.

COMPUTATION

Programmes GENSYN and DASTAT are written in Fortran IV for the IBM 360 computer. A version written in Algol 60 for the English Electric KDF9 is also available. These programmes, along with user instructions and sample output results, are held by the Systems Laboratory, Ministry of Works, Wellington.

CONCLUSION

The proven usefulness of the technique of stream flow generation should be extended by the device of selecting a value of the skew coefficient such that the parts of the synthetic record that are of greatest interest for any particular application will have the greatest accuracy.

Consideration has been given to the probable accuracy of the synthetic record, and the physical significance of the statistical parameters of flow have been explained.

ACKNOWLEDGMENTS

Permission to publish this work was granted by the Commissioner of Works. Computer programmes were written by Mr J. L. Fairweather.

The author wishes to thank Mr M. D. Johnson for help in the preparation of this paper.

REFERENCES

Benson, M. A.; Matalas, N. C. 1967: Synthetic hydrology based on regional statistical parameters. *Water Resources Research 3* (4) : 931–935.

Crow, E. L.; Davis, F. A.; Maxfield, M. W. 1960: *Statistics manual taken from ordinance development.* New York, Dover. 288 pp.

Gumbel, E. J. 1958: Statistical theory of floods and droughts. *J. Inst. Water Engineers 12*: 157.

Harms, A. A.; Campbell, T. H. 1967: An extension to the Thomas-Fiering model for the sequential generation of stream flow. *Water Resources Research 3* (3) : 653–661.

Matalas, N. C. 1967: Mathematical assessment of synthetic hydrology. *Water Resources Research 3* (4) : 937–945.

Matalas, N. C.; Benson, M. A. 1968: Note on the standard error of the coefficient of skewness. *Water Resources Research 4* (1) : 204–205.

Pearson, F. H. 1967: *A water balance and other hydrological investigations for the River Anker.* (Unpublished M.Sc. dissertation, University of Newcastle-upon-Tyne.) 136 pp.

Thomas, H. A.; Fiering, M. B. 1962: Mathematical analysis of streamflow sequences in analysis of river basins by simulation. In: A. Maass *et al. Design of water resource systems.* Cambridge, Harvard University Press. p. 466.

Weatherburn, C. E. 1957: *A first course in mathematical statistics.* Cambridge, Cambridge University Press. 277 pp.

Wisler, C. O.; Brater, E. F. 1959: *Hydrology.* 2nd ed. Tokyo, Toppan. p. 333.

Yevjevich, V. W. 1967: Mean range of linearly dependent normal variables with application to storage problems. *Water Resources Research 3* (3) : 663–671.

APPENDIX 1

Minimum Flow Studies

This procedure is for the selection of a value of the skew coefficient such that the frequency distribution of the synthetic record will be asymtopic to a predetermined absolute minimum flow.

The flow in any month $t+1$ is given by the expression:

$$q_{t+1}=\bar{q}_{t+1}+b_t(q_t-\bar{q}_t)+s_{t+1}I_{t+1}(1-r_t^2)^{\frac{1}{2}}$$

with the notation described earlier in this paper.

The following additional variables are introduced:

M_{t+1} and M_t are absolute minimum mean monthly flows for months $t+1$ and t.

K_{t+1} is the skew of the distribution of synthetic flows for month $t+1$ to be chosen such that the distribution of flows for month $t+1$ is asymtopic to M_{t+1}. The distribution of synthetic flows is modelled on the gamma distribution.

m_{t+1} is the order of the gamma distribution used to model synthetic flows for month $t+1$.

$Z(m_{t+1})$ is the gamma distribution used to model synthetic flows for month $t+1$. $Z(m_{t+1})$ is calculated by the relation:

$$Z(m_{t+1})=\sum_{P=1}^{m_{t+1}}[r(P)]^2$$

where P is a counter, and $r(P)$ is a normal random deviate with zero mean and unit variance.

A series of values of $r(P)$ for varying P may be calculated by a standard computer procedure. If m_{t+1} is nonintegral, the final term in the series of $Z(m_{t+1})$ is multiplied by the fractional part of m_{t+1}. For instance, if $m_{t+1}=1.5$, then

$$Z(1.5)=[r(1)]^2+0.5[r(2)]^2.$$

I'_{t+1} is the minimum value to which the distribution of I_{t+1} is asymtopic. I_{t+1} is given by

$$I_{t+1}=[Z(m_{t+1})-m_{t+1}]/\sqrt{(2m_{t+1})}.$$

Since the minimum value of $r(P)$ is zero, the minimum value of $Z(m_{t+1})$ is also zero, and the minimum value of I_{t+1}, viz. I'_{t+1}, is given by

$$I'_{t+1}=-\sqrt{m_{t+1}}/2.$$

Also $$K_{t+1}=\sqrt{8}/m_{t+1},$$

since the skew of the distribution of synthetic flows is equal to both the skew of the distribution of I_{t+1} and the skew of the distribution of $Z(m_{t+1})$,

i.e. $$I'_{t+1}=-2/K_{t+1}.$$

Note that K_{t+1} must be greater than zero. That is, the distribution of synthetic flows for each month must be positively skewed. Incidentally, $\bar{q}_{t+1}$, $\bar{q}_t$, b_t, s_t, and r_t must also all be positive.

The conditions for minimum q_{t+1} (i.e. $q_{t+1}=M_{t+1}$) are:

(1) q_t must be a minimum (i.e. $q_t=M_t$),

(2) I_{t+1} must be a minimum (i.e. $I_{t+1}=-2/K_{t+1}$).

Under these conditions the following relation applies:

$$M_{t+1}=\bar{q}_{t+1}+b_t(M_t-\bar{q}_t)-2s_{t+1}(1-r_t^2)^{\frac{1}{2}}/K_{t+1},$$

$$\text{i.e.}\quad K_{t+1}=\frac{2s_{t+1}(1-r_t^2)^{\frac{1}{2}}}{\bar{q}_{t+1}-M_{t+1}-b_t(\bar{q}_t-M_t)}.$$

If M_{t+1} and M_t are both zero,

$$K_{t+1}=\frac{2s_{t+1}(1-r_t^2)^{\frac{1}{2}}}{\bar{q}_{t+1}-b_t\bar{q}_t}.$$

If, further, r_t (and therefore b_t) are zero,

$$K_{t+1}=\frac{2s_{t+1}}{\bar{q}_{t+1}}$$

$$=2\times\text{coefficient of variation.}$$

This final relation may be useful for generating records that are not serially correlated and that are asymtopic to zero; for instance rainfall records which are shown not to be serially correlated may be extended, using skew values equal to twice the coefficient of variation.