JOURNAL OF HYDROLOGY (N.Z.) Vol. 14, No. 1, 1975

COMPRESSION OF TIME-SERIES DATA

R. P. Ibbitt*

ABSTRACT

When water-level data are stored on a computer the amount of storage space needed can be reduced by eliminating data that can be provided approximately by linear interpolation. Benefits from reduced computer time are also obtained since fewer values have to be retrieved from the file. The aim of the compression process is to filter out small and abrupt water-level fluctuations without affecting the information content of the record. To avoid over-compressing the data it is desirable to have some method by which an upper limit on the 'compression range' can be determined. A method is presented by which this limit can be obtained under different assumptions about the accuracy of a rating curve.

INTRODUCTION

Water-level movement is a continuous process. Instruments that record water-level movement do so either as a continuous trace (chart recorders) or by sampling the water level at various points in time. When storing water-level records on a digital computer, it is necessary to have them in numerical form. Instruments that sample the water-level record provide a numerical record that can be readily fed into a computer. Traces from chart recorders, however, have to be turned into a numerical form by 'digitizing' them, an operation which effectively samples the continuous trace to give a record which is similar to that which would have been obtained with the numerical punched-paper-tape recorders.

Implicit in all continuous records that are finally stored by a digital computer, is an assumption concerning the variation of the water-level record between the stored points. The usual assumption is one of linearity.

* Water and Soil Division, Ministry of Works and Development, Wellington.

Punched-tape recorders that record values at regular time intervals tend to record more points than are necessary to adequately represent changes in water level. For example, Chandler and Patterson (1970) report over-recording for one type of punched-tape recorder as being as high as 95 percent. Over-recording of this magnitude obviously means that any data-processing system must carry a large overhead unless some means can be found to reduce it. This is done for data filed by the Ministry of Works and Development by the use of a process called compression.

Compression is achieved by not storing the middle point of three points if it falls within a preset amount of a straight line joining the two extreme points. The preset amount is called the compression range, and its estimation by an objective method is the subject of this paper.

THE EFFECTS OF COMPRESSION

Compression eliminates from the record those points that can be estimated satisfactorily by linearly interpolating between the values that are retained. Thus, when estimating an instantaneous stage value at a particular time, the amount of error introduced by compressing the record will vary from zero at a stored value to a maximum of $\pm r$ between two stored values, where r is the compression range. The error in the instantaneous water level will introduce an error into the corresponding instantaneous flow value.

The effects of compression have been portrayed as being more complex than those just described because confusion has arisen when compressed stage data are used to calculate mean flow values. For example, it has been argued that if the stage value varies linearly throughout the period for which the mean is to be calculated, then the mean flow value estimated using all the recorded stage values is more reliable than that calculated using only the compressed data, i.e. a stage value at either end of the period. This is because most rating curves are non-linear. This argument is invalid because the reliability of a mean flow value depends mainly upon how many stage value are used in its calculation, not upon their recorded frequency. By using linear interpolation to obtain extra stage values and then applying the rating curve to all the stage values, a mean flow value can be calculated that is as good as one obtained by using all the originally recorded stage data.

Since the effects of compression on the estimation of mean flows are clouded by such factors as how many stage values should be used with a non-linear rating to obtain acceptable mean flow

values, the rest of this paper will concentrate upon the effects of compression on instantaneous flow values. The errors so found will represent extreme values, since the operation of calculating mean flows in many cases causes the errors to compensate. In cases where the value is steadily rising or falling over the period of the mean, the errors will compensate least. In this case compression will introduce positive bias, the magnitude of which will be of the same order as the mean of the errors introduced into instantaneous flow values by compression of the instantaneous stage data.

Besides the economic benefits, compression offers significant numerical advantages when data are being used in conjunction with differential equations. In calculations of such quantities as lake inflows from lake outflows, small and abrupt changes in water level can lead to numerical instability. Compression, by creating more gradual changes in water-level movement, lessens the potential for numerical problems.

ERRORS IN THE ESTIMATION OF FLOW VALUES

The compression process is carried out on water levels, since these are the data that are stored. Normally, however, flow values rather than water levels are required and questions arise about the possible effect of compression of water-level data on the flow values. The answers to such questions lie in the rating curve used to obtain flow values from levels. Any rating curve is only an estimate of the true stage–flow relationship, so that errors in flow values are invariably introduced when a rating curve is used. Provided that the changes in flow values brought about by compression are small compared with those inherent in the use of a rating curve, then there is no basis for choosing the uncompressed value as being any more correct than the compressed value.

The rating curve represents the results of what amounts to a non-linear regression. For a given level, h, on the staff gauge it allows the corresponding flow to be estimated within limits that can be calculated. These limits allow for all the errors that could be present in an estimate of the flow based on a particular staff-gauge value. When flow values are estimated from the recorded stage values, the limits that can be placed upon the flow estimate are widened because of uncertainties introduced into the assessment of recorded stage value by such factors as recorder precision.

If all flow values were calculated from the rating using external staff-gauge readings, the confidence limits that could be placed upon the resulting flow are calculated as follows:

1. Form the differences between actual gauged flows q_i and the values predicted by the rating q_i' (Table 1, column 4).
2. Since, in absolute terms, the differences calculated in step 1 increase with the value of the flow they should be scaled to a common datum to avoid biasing the results in favour of large flows. This can be done by dividing the difference $q_i - q_i'$ by q_i' (Table 1, column 5).
3. Form the sum of the squares of the proportional differences formed in step 2.

$$\sum_{i=1}^{m} \left(\frac{q_i - q_i'}{q_i'} \right)^2$$

where m is the number of gaugings used to construct the rating.

TABLE 1 — Calculation of proportional rating variance.

Line	*(1) Stage (m)*	*(2) Gauged flows (m^3/s)*	*(3) Flows from rating (m^3/s)*	*(4) Col. 2 − Col 3*	*(5) Col. 4 ÷ Col. 3*	*(6) Col. 5 squared*	*(7) Cumulative sum of Column 6*
1	1.366	2.106	1.773	0.333	0.187	0.03531	0.03531
2	1.426	2.120	1.964	0.156	0.079	0.00632	0.04164
3	1.798	2.341	3.379	−1.038	−0.307	0.09440	0.13605
4	1.850	4.098	3.609	0.489	0.135	0.01833	0.15438
5	0.939	0.732	0.706	0.026	0.036	0.00136	0.15574
6	0.866	0.538	0.574	−0.036	−0.062	0.00387	0.15962
7	0.978	0.733	0.782	−0.049	−0.063	0.00398	0.16361
8	2.730	7.990	8.730	−0.740	−0.084	0.00717	0.17079
9	1.340	1.630	1.693	−0.063	−0.037	0.00139	0.17219
10	2.540	7.080	7.425	−0.345	−0.046	0.00216	0.17435
11	2.010	4.500	4.367	0.133	0.030	0.00092	0.17527
12	2.190	5.870	5.311	0.559	0.105	0.01108	0.18636
13	1.850	3.930	3.609	0.321	0.088	0.00789	0.19425
14	1.721	3.080	3.053	0.027	0.008	0.00007	0.19433
15	1.542	2.620	2.362	0.258	0.109	0.01190	0.20624
16	3.581	15.280	15.949	−0.669	−0.041	0.00175	0.20800
17	3.115	11.320	11.715	−0.395	−0.033	0.00113	0.20913
18	4.260	24.220	23.356	0.864	0.036	0.00136	0.21050
19	5.334	38.350	38.133	0.217	0.005	0.00003	0.21053
20	4.505	25.600	26.394	−0.794	−0.030	0.00090	0.21144
21	6.675	65.500	61.971	3.529	0.056	0.00324	0.21468
22	5.819	44.100	46.057	−1.957	−0.042	0.00180	0.21649
23	7.490	81.400	79.441	1.959	0.024	0.00060	0.21710
24	7.160	73.190	72.092	1.097	0.015	0.00023	0.21733
25	3.630	14.930	16.434	−1.504	−0.091	0.00837	0.22570
26	7.767	89.590	85.898	3.692	0.042	0.00184	0.22755

Proportional rating variance $= (26-1)^{-1} \times 0.22755 = 0.0091$.

4. The proportional variance of the distribution of flow values for a given water level, which will be referred to as the proportional 'rating variance', is then given by

$$\frac{\mathrm{Var}'(q)}{q^2} = \frac{1}{(m-1)} \sum_{i=1}^{m} \left(\frac{q_i - q_i'}{q_i'} \right)^2 \qquad (1)$$

5. If an assumption is then made about the distribution of flow values (the usual one is that they are normally distributed, and this has been adopted in the rest of this paper), confidence intervals can be placed upon any flow predicted from the rating.

To assess the consequences upon the calculation of a flow value using a recorded stage value instead of the external staff-gauge value, the likely differences between the recorded water level and the external staff-gauge value need to be quantified and then converted into the effect upon the flow. For a given external staff-gauge reading there will be a range of values that the recording instrument could register. This range depends upon such things as recorder precision and response time.

For a particular instrument and installation, sources of error should be identified and numerical values placed upon their magnitude. The method for dealing with each numerical value depends upon the confidence that can be placed on it. Those values that it is considered will rarely be exceeded, e.g. less often than once per hundred 'observations', should be treated as maximum errors – while figures over which there is some doubt, e.g. half the errors are likely to exceed the given value, should be treated as probable errors.

If the errors are normally distributed, then maximum errors at the 1-percent level of confidence can be equated to $2.57\sigma_i$ while probable errors can be equated to $0.67\sigma_i$ where σ_i is the standard deviation of the error distribution of the ith source of error. Remembering that variances are the square of standard deviations and that variances are cumulative, the variance of the distribution of the differences between the staff gauge readings and recorder readings, $\mathrm{Var}(h)$ is given by

$$\mathrm{Var}\,(h) = \sum_{i=1}^{k} \left(\frac{\sigma_i}{a_i} \right)^2 \qquad (2)$$

where there are k sources of error and a_i can have values of 2.57 or 0.67 only, depending upon whether or not σ_i results from a maximum error or a probable error.

To determine the variance of the distribution of flows the proportional rating variance, equation (1), must have the 'stage variance' Var(h) added to it, after Var(h) has been scaled into flow units. The appropriate scaling factor (see Appendix 1) is the square of the rate of change of flow with stage. (Note that this is the reciprocal of the 'slope of the rating'.)

The effect of compression is to introduce an additional error source into equation (2). Since the effects of the compression can be no greater than the compression range, this extra term should be treated as a maximum error.

Total proportional variance of the distribution of the flow values, $\mathrm{Var}(q)/q^2$, can then be obtained from equations (1) and (2) as

$$\frac{\mathrm{Var}\,(q)}{q^2} = \frac{\mathrm{Var}'\,(q)}{q^2} + A\,\mathrm{Var}\,(h) \qquad (3)$$

where $$A = \frac{1}{q^2}\left(\frac{\partial q}{\partial h}\right)^2 = \frac{1}{q^2}\left(\frac{1}{s}\right)^2$$

s being the 'slope of the rating' (see Appendix 1).

From equation (3) the standard deviation of the distribution of flow values for a given value of q can be calculated, and by comparing the confidence limits with and without the effects of compression introduced via the term Var(h), the consequences of different compression ranges can be examined.

EXAMPLE

Twenty-six gaugings from the Taueru River at Te Weraiti are shown in Fig. 1. Calculation of the proportional rating variance (equation 1) is shown in Table 1.

In estimating the stage variance Var(h) (equation 2), the major source of error would appear to be instrument precision, the recorder being a 12-m Foxboro. A reasonable value for the magnitude of the recorder precision would be 0.025 m. (This corresponds to being able to read the trace ± 0.25 mm.) As it is likely that errors greater than ± 0.25 mm will frequently occur in reading values from such a chart, the value of 0.025 m should be treated as a probable error, the appropriate value of a_i in equation (2) being 0.67. Assuming that the other source of error would be that caused by the compression at range r mm, use of equation (2) leads to

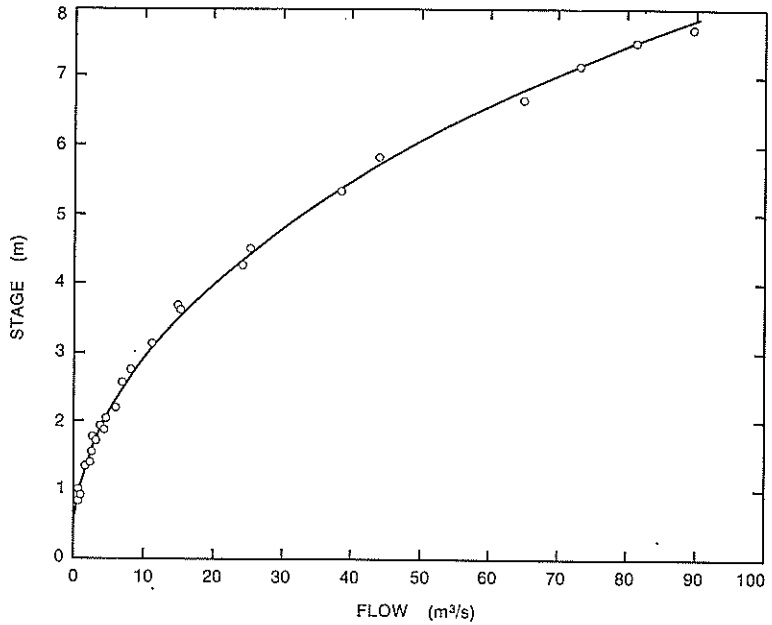

FIG. 1 — Rating curve for Taueru River at Te Weraiti.

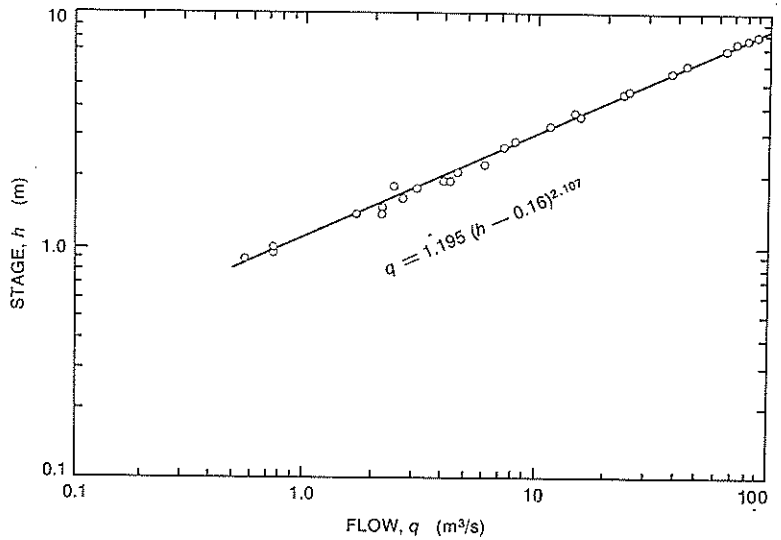

FIG. 2 — Rating curve for Taueru River at Te Weraiti plotted on logarithmic scale.

$$\mathrm{Var}(h) = \left(\frac{0.025}{0.67}\right)^2 + \left(\frac{r}{25.7\times 1000}\right)^2 = 0.00139 + 0.000000151r^2 \qquad (4)$$

Direct estimation of slopes from curves is notoriously inaccurate. To improve the slope estimate the rating was plotted on logarithmic graph paper (Fig. 2) to give the equation

$$q = 1.195(h - 0.16)^{2.107} \text{ for } 0.8\text{ m} < h < 8.0\text{ m} \qquad (5)$$

where q is in m^3/s and h is in m. By using Appendix 2 with $k = 1.195$, $n = 2.107$ and $a = -0.16$, a value for A (equation 3) can then be calculated.

Substitution of the appropriate values into equation (10) (Appendix 2) gives the A values, Table 2, to be substituted into equation (3). Not surprisingly, the greatest value of A (8.91) occurs for the least stage value – a property that could have been deduced from equation (10).

TABLE 2 — A factors.

Line	*Stage height (m)*	*A from eqn (3)*
1	1.366	3.05
2	1.426	2.77
3	1.798	1.65
4	1.850	1.55
5	0.939	7.32
6	0.866	8.91
7	0.978	6.63
8	2.730	0.67
9	1.340	3.19
10	2.540	0.78
11	2.010	1.30
12	2.190	1.08
13	1.850	1.55
14	1.721	1.82
15	1.542	2.32
16	3.581	0.38
17	3.115	0.51
18	4.260	0.26
19	5.334	0.17
20	4.505	0.24
21	6.675	0.10
22	5.819	0.14
23	7.490	0.08
24	7.160	0.09
25	3.630	0.37
26	7.767	0.08

Substituting the values of A and the different values of $\mathrm{Var}(h)$ obtained for different values of the compression range r into equation (3) gives the proportional variances of the flow distribution. The percentage confidence limits for any value of q can then be readily found for any compression range r and for any desired level of confidence by substituting the appropriate value of r into equation (4) and C into the expression:

$$C\sqrt{\mathrm{Var}(q)}/q$$

where, on the assumption of normally distributed errors, C takes the values of 2.57, 1.96, and 1.64 for confidence limits with expectations of being exceeded of 1 percent, 5 percent and 10 percent respectively.

To avoid large tables attention has been focussed on the 6th line of Tables 1 and 2, since this represents the worst case. For the same reason, only values for $C=1.96$ (i.e. 5-percent confidence limits) have been given. Table 3 shows how the errors in a flow value accumulate. The major source of error is seen to be the rating curve, but it is closely followed by the instrument precision. In this case the instrument is a particularly insensitive one. To assess the improvement to be gained from using a recording instrument with a precision of ± 3 mm, $\mathrm{Var}(h)$ was recalculated for equation (4) using 0.003 instead of 0.025. The results are shown in Table 4.

DISCUSSION

The figures given in Tables 3 and 4 illustrate two points: (a) that with better instrument precision, compression at a given range has greater potential effect; (b) that for the rating in the example the basic rating error is the dominant source of error in flow calculations.

It should, however, be pointed out that the procedure given above treats the case in which errors are likely to be greatest. Also, to put the size of the figures in Tables 3 and 4 into perspective requires a careful statement of what the figures actually mean. For instance, the figure of 19.36 percent given at the foot of Table 4 means that on average 5 percent of the instantaneous flow values calculated from the stage–time data using the given rating will differ from the true value by more than ± 19.36 percent. Usually, mean flows would be expected to be more accurate than 19.36 percent because of compensating errors.

TABLE 3 — Percentage errors in a flow estimate for 5-percent confidence limits.

Basic error from rating:	18.70
Additional error from instrument precision:	10.03
Additional error from compression at ranges of —	
5 mm:	0.02
10 mm:	0.11
15 mm:	0.20
20 mm:	0.35
30 mm:	0.81
50 mm:	2.16
Total error at compression range 20 mm:	18.70+10.03+0.35=29.08

TABLE 4 — Percentage errors in a flow estimate for 5-percent confidence limits, with improved instrument precision.

Basic error from rating:	18.70
Additional error from improved instrument precision:	0.12
Additional error from compression at ranges of —	
5 mm:	0.03
10 mm:	0.14
15 mm:	0.31
20 mm:	0.54
30 mm:	1.20
50 mm:	3.17
Total error at compression range 20 mm:	18.70+0.12+0.54=19.36

Since the estimate of flows from recorded stage values using the rating is very dependent upon the accuracy of the rating, this warrants closer study. The rating affects the flow calculations through the proportional rating variance, equation (1). From equation (1) it will be seen that doubling the number of gaugings halves the proportional rating variance. However, since it is the square root of the proportional rating variance that is used to calculate confidence limits, quadrupling the number of gaugings is needed to halve the width of the confidence limits for a given probability of their being exceeded. This fact, which is almost certainly not new, should form an important part in the formulation of gauging policy.

So far no mention has been made of sites at which zero flows have been recorded. The principles set out above cannot be applied without some modification. Multiplying equation (3) through by q^2 and substituting for coefficient A gives

$$\mathrm{Var}(q) = \mathrm{Var}'(q) + (\partial q/\partial h)^2 \, \mathrm{Var}(h)$$

Close to zero flow, large changes in h cause small changes in q, so that $(\partial q/\partial h)$ tends to zero. This leads to the conclusion that near

zero flow the accuracy of flow estimates depends entirely upon the rating variance – theoretically at least. The conclusion is not of much practical help when trying to decide compression ranges. To meet this contingency the same procedure as for sites with perennial flow should be followed, but with an arbitrarily defined non-zero minimum flow. A suitable value for this minimum flow would be that at which estimates of the slope of the rating can no longer be made satisfactorily. The error derived for the different compression ranges would then be assumed to apply at zero flow.

So far the discussion has avoided any suggestion as to which of the compression ranges should be used. The following guide is suggested as reasonable: that the effect of compression should not exceed 5 percent of the combined effects of the errors caused by rating inadequacies and recorder precision. For the data given in Table 3 this would imply a compression range of about 40 mm, since 5 percent of $(18.70+10.03)$ is approximately $(0.81+2.16)/2$ (the mean of the compression effects at 30 mm and 50 mm). This rule assumes both a fixed accuracy for the recording instrument and for the rating. While this may be so for the former (at least for long periods of time), it will only be true for the latter when a rating has been replaced by an entirely new rating. (Note that a replaced rating still has a period of validity, whereas an amended rating is usually applicable from the same instant of time as the rating it supersedes.) Where there are already closed ratings the basic error for a new rating can be estimated by averaging the basic errors from each of the closed ratings. For sites where there has been no previous rating, a safe assumption would be to equate the basic error to half that calculated using the gaugings available.

ACKNOWLEDGMENTS

The author wishes to thank: Mr F. Scarf formerly of the Ministry of Works and Development, Nelson Residency, for initiating the work that led to this paper; Mr R. Curry of the Ministry of Works and Development, Wellington District Office, for providing the data for the example; and Dr S. M. Thompson of the Ministry of Works and Development, Power Division, for his constructive criticism of the drafts. Permission to publish this paper was given by the Commissioner of Works.

REFERENCES

Chandler, A.; Patterson, J. E. 1970: Digital event recorders for representative and experimental basins. In: *Proceedings of the Symposium on the Results of Research on Representative and Experimental Basins, Wellington, 1970.* IASH Publication No. 96. pp. 700–707.

Kendall, M. G.; Stuart, A. 1969: *The Advanced Theory of Statistics.* vol. 1. Hafner, New York.

APPENDIX 1

Kendall and Stuart (1969: pp. 231–232) derive the following expression for the variance of a dependent variable g in terms of a series of independent variables:

$$\text{Var}\,(g[x_1 \ldots x_i \ldots x_n]) = \sum_{i=1}^{n} \left(\frac{\partial g}{\partial x_i}\right)^2 \text{Var}\,(x_i) \qquad (7)$$

where $(\partial g/\partial x_i)$ is the partial derivative (or slope) of g with respect to x_i evaluated at the expected value of x_i.

APPENDIX 2

If parts of a rating plot as a straight line on log-log paper, then the each linear portion can be represented by an equation of the form

$$q = k(h+a)^n \qquad (8)$$

where q is the flow at stage h, and k, n and a are constants. Differentiation of equation (8) with respect to h gives

$$\frac{\partial q}{\partial h} = \frac{nq}{(h+a)} = nk(h+a)^{n-1} \qquad (9)$$

whence

$$A = \frac{n^2k^2}{q^2}(h+a)^{2n-2} = \frac{n^2}{(h+a)^2}. \qquad (10)$$