Journal of Hydrology. 20 (1973) 121–136

Research Papers

ASSESSING SYSTEMATIC ERRORS IN RAINFALL–RUNOFF MODELS

A.P. AITKEN

Snowy Mountains Engineering Corporation, Cooma North, N.S.W. (Australia)

(Accepted for publication March 12, 1973)

ABSTRACT

Aitken, A.P., 1973. Assessing systematic errors in rainfall–runoff models. *J. Hydrol.*, 20: 121–136.

Techniques were examined for measuring the errors in estimated monthly flows from seven deterministic rainfall–runoff models, using results published in the literature. Various commonly used statistical tests were examined, but these did not disclose systematic errors which, by the application of other tests, were often found to exist. It was therefore concluded that hydrologists should test their results by applying a simple sign test and a measure of the reproduction of the residual mass curve referred to as the residual mass curve coefficient.

INTRODUCTION

In recent years many research workers have been engaged on the development od deterministic rainfall–runoff models whose object is to simulate the complex physical relationships occurring on a catchment during the rainfall–runoff phase of the hydrologic cycle. In a number of cases, where existing stream-flow records have been too short to form a reliable basis for storage analysis, these models have been used to extend the historical records. Hydrologists have used a variety of statistical measures and tests to express the agreement or disagreement between the computed and observed flows. It appears, however, that insufficient attention has been given to the important distinction between random and systematic errors.

This paper is concerned with the techniques of model analysis and while specific models are mentioned as illustrations, it is not the author's purpose to examine their relative merits nor to recommend the use of a particular one. Uniformity in the comparisons was achieved by using the observed and estimated monthly flows for all models, as this is the output usually required for storage analysis in water resources engineering.

DISCUSSION OF ERRORS AND TESTS

The outputs of all rainfall–runoff models are subject to errors which may be random or systematic. Random errors occur when the model shows no tendency to over- or underestimate for a number of successive time intervals. Systematic errors occur when the sign of the error tends to persist over a series of time intervals. Both types of error may be caused by imperfections in the structure of the model.

The existence of data errors must also be recognised. Included in this category, are the errors in the input data to the model including rainfall, evapotranspiration and various parameters used in equations describing the physical catchment processes such as interception and infiltration. The gauged streamflow which is used for comparison with the estimated flows may also contain errors. Random errors in the data will undoubtedly produce random errors in the output. Systematic errors in the data on the other hand, will probably not be apparent as errors (differences between observed and estimated flows) in the output, but will be reflected as incorrect values in the parameters of the model.

If a model is used for synthesizing flows for the solution of storage problems, the existence of systematic errors may be very much more serious than the effects of random errors, yet the analysis of most hydrologic models fails to distinguish between the two kinds.

Fitting criteria

The agreement between two time series may be tested by computing and comparing certain statistical parameters of the two series, or more directly, by computing dimensionless coefficients of agreement, e.g., the coefficient of determination. Some of these criteria distinguish between random and systematic errors, others do not. Those commonly used in testing hydrologic models are in the latter category.

Mean and standard deviation

Generally a first requirement of a model should be the ability to reproduce the mean and standard deviation of the observed runoff. It is essential that the mean of the computed flows agrees closely with the observed record. Agreement in the standard deviation of flows is also a useful criterion, but it should be noted that neither of these criteria indicate how well individual, estimated values fit the observed values, nor do they distinguish between random and systematic errors.

Coefficient of determination

The coefficient of determination is very commonly used for measuring the degree of association between observed and estimated flows. It is defined in the following form to permit a comparison with the next definition:

$$D = \frac{\Sigma(q_c - \overline{q}_c)^2 - \Sigma(q_c - q_{est})^2}{\Sigma(q_c - \overline{q}_c)^2} \qquad (1)$$

where D = the coefficient of determination; q_c = observed runoff; $\overline{q}_c$ = mean of the observed runoffs; q_{est} = estimated runoff obtained from the regression line of q_c on q_e; and q_e = estimated runoff.

The term $\Sigma(q_c - \overline{q}_c)^2$ is referred to as the initial variation and the term $\Sigma(q_c - q_{est})^2$ the residual variation or unexplained variation. The coefficient of determination will always be less than unity. A high value of D indicates good results from a model; a low value poor, or even statistically insignificant results. This coefficient is a good measure of the degree of association between the observed and estimated values. It does not, however, reveal systematic errors should these exist.

Coefficient of efficiency

The term efficiency to describe the degree of association between observed and estimated flows was introduced by Nash and Sutcliffe (1970). It is analogous to the coefficient of determination as can be seen from the following equation, but it is not identical:

$$E = \frac{\Sigma(q_c - \overline{q}_c)^2 - \Sigma(q_c - q_e)^2}{\Sigma(q_c - \overline{q}_c)^2} \qquad (2)$$

where E = coefficient of efficiency of the model.

As with the coefficient of determination, the term $\Sigma(q_c - \overline{q}_c)^2$ represents the initial variation and the term $\Sigma(q_c - q_e)^2$ the residual or unexplained variation. Again the value of this statistic will always be less than unity. If the results from a model are highly correlated but biased, that is, they do not plot randomly around the 45° line on a graph of observed versus estimated events, then the value of the coefficient of efficiency will be lower than the coefficient of determination.

Serial correlation coefficient

Hydrologists sometimes test whether the first-order (and maybe other orders) serial correlation coefficient of the estimated and observed runoffs are significantly different. This test is intended to indicate whether systematic errors are present in the estimated flows, but it appears doubtful whether the test is sufficiently powerful for this purpose. Further, occasions arise where no significant serial correlation occurs between monthly runoffs.

Sign tests

Though they do not appear to have been applied by hydrologists, sign tests provide a very simple method of testing whether the estimated time series contains systematic errors. One such test involves listing the observed and estimated flows side by side and allocating a plus sign to each over-estimate of monthly runoff and a negative sign to each under-estimate. The number of runs of plus and minus signs can then be counted and compared with the expected number. If a Chi-square test indicates that the number of runs is significantly less than that expected for random errors, then it is concluded that the model introduces a systematic error.

Maximum range of the residual mass curve

Engineers have commonly used a method referred to as "residual mass curve analysis" when determining the required sizes of storages for water resources projects. The residual mass curve is computed by first subtracting the mean monthly flow from each individual (monthly) flow to obtain the residuals, which are summed sequentially. The series of monthly values so obtained form a curve which commences and ends on the abscissa (e.g., Fig. 1).

The ordinate of the residual mass curve, at any point in time, depends on the history of preceding events. Comparison of the residual mass curves for observed and estimated flows may therefore reveal the existence of systematic errors in the estimated flows. Although this type of comparison was used by Brown (1961) in an examination of various correlation methods of streamflow estimation in the Snowy Mountains in New South Wales, it has apparently not been used for rainfall–runoff models.

As a simple direct test, the percentage error in the maximum range of the estimated residual mass curve can be used. This gives an indication of the error in the computed storage requirement for full regulation of the streamflow should this calculation be based on the estimated rather than the historical record.

Residual mass curve coefficient

A statistic is now introduced which measures the association between the observed and estimated residual mass curves. This statistic is referred to as the residual mass curve coefficient and is defined as follows:

$$R = \frac{\Sigma(D_c - \bar{D}_c)^2 - \Sigma(D_c - D_e)^2}{\Sigma(D_c - \bar{D}_c)^2} \quad (3)$$

where R = residual mass curve coefficient; D_c = departure from the mean for the observed residual mass curve; $\bar{D}_c$ = mean of the departures from the mean for the observed residual mass curve; and D_e = departure from the mean for the estimated residual mass curve.

As with eq. 1 and 2 a value of R equal to unity indicates perfect agreement. In practice, however, the value of R will always be less than unity. This statistic is thought to have an important advantage over those of eq. 1 and 2 in that it measures the relationship between the sequence of flows and not simply the relationship between individual flow events. If the flow sequence contains systematic errors this coefficient should indicate their presence.

DISCUSSIONS OF MODELS EXAMINED

Seven models were chosen for examination for the following reasons: (1) results from the application of these models to various catchments have been published; (2) the chosen models are generally the better known ones; (3) most of these models have been used in practice to solve engineering problems. Table I lists the catchments and models examined.

In examining these results it is in no way the intention of the author to demonstrate the comparative merits of the models. To do so would, at least, involve testing all the models on the same group of catchments.

Number of parameters

In general the greater the number of parameters which the hydrologist can adjust (usually, within certain liberal ranges) either automatically within the computer or manually prior to each computer run, the greater the possibility of adjusting the model to simulate more closely the observed runoff events.

Table II summarises the approximate number of parameters in each of the seven models studied. It will be noticed that the number of parameters ranges from only 4 in the case of the Boughton-Jones Model up to a maximum of 43 in the USDAHL Model. Although the number of parameters in a model

TABLE I

Details of the catchments and models examined

No.	Catchment name	Area (km^2)	Model	Reference
1	Russian River (U.S.A.)	939	Stanford IV	Crawford and Linsley (1966)
2	Russian River (U.S.A.)	2,050	Stanford IV	Crawford and Linsley (1966)
3	Russian River (U.S.A.)	3,480	Stanford IV	Crawford and Linsley (1966)
4	French Broad River (U.S.A.)	766	Stanford IV	Crawford and Linsley (1966)
5	South Yuba River (U.S.A.)	28.0	Stanford IV	Crawford and Linsley (1966)
6	Napa River (U.S.A.)	211	Stanford IV	Crawford and Linsley (1966)
7	Beargrass Creek (U.S.A.)	47.9	Stanford IV	Crawford and Linsley (1966)
8	Maribyrnong River (Aust.)	865	Porter-McMahon	Porter and McMahon (1971)
9	Dandenong Creek (Aust.)	270	Porter-McMahon	Porter and McMahon (1971)
10	Ramu River (New Guinea)	881	Ribeny-Brown	Ribeny and Brown (1968)
11	Wagga Research Catchment (Aust.)	0.081	Boughton	Boughton (1966)
12	Parwan Weir (Aust.)	0.85	Boughton-Jones	Jones (1970)
13	River Ray (United Kingdom)	19	SM 2 D (b)	Mandeville et al. (1970)
14	Little Mill Creek (U.S.A.)	18.5	USDAHL	Holtan and Lopez (1970)
15	Upper Taylor Creek (U.S.A.)	255	USDAHL	Holtan and Lopez (1970)
16	Beaver Creek (U.S.A.)	115.0	USDAHL	Holtan and Lopez (1970)
17	Brushy Creek (U.S.A.)	109.8	USDAHL	Holtan and Lopez (1970)

must have a great effect on its potential accuracy this question is not examined in this paper.

Time interval

Generally the time interval of the input data to the model, such as the rainfall and potential evaporation, is the same as that of the various computational processes in the model. Some of the models tested do, however, vary the time interval depending on the amount of "hydrological activity". Within reason the shorter the time interval of a model the greater its potential ac-

TABLE II

Number of model parameters and time interval

No. *1	Catchment model	Number of parameters	Time interval
1–7	Stanford	19*2	1 h
8, 9	Porter-McMahon	15	1 day
10	Ribeny-Brown	8	7 days
11	Boughton	10	1 day
12	Boughton-Jones	4	1 day
13	Institute of Hydrology	8	3 h
14–17	USDAHL	43	$\frac{1}{4}$ h *3

*1 Reference number used in Table I.
*2 Does not include the snowmelt parameters.
*3 Could also be less than 15 min if desired.

curacy as it permits a closer simulation of the physical processes which are occurring over short intervals of time. Further, the smaller the catchment under study, the smaller the time interval required for accurate simulation of the catchment processes.

Table II also summarises the time intervals generally adopted for each of the models studied. The time interval varies from 1 week in the case of the Ribeny-Brown Model down to an interval of several minutes for the USDAHL Model. The effect of the time interval on model errors was not considered in this paper but it is known to be a very significant factor.

APPLICATIONS OF FITTING CRITERIA

Traditional tests

Mean and standard deviation

The mean and standard deviation of the observed and estimated monthly flows on the catchments studied are shown in Table III. It is apparent that there is good agreement between these two statistics on all the catchments. Tests were carried out to determine whether the means of observed and estimated discharges differed significantly at the 0.05 level. In all cases, the results were negative.

Coefficient of determination

Table III shows the coefficient of determination between observed and estimated monthly flows. It will be noticed that the range in this coefficient is from 0.41 to 0.99 (which corresponds to a correlation coefficient ($r = \sqrt{D}$) of from 0.64 to 0.99). All these results are highly significant.

TABLE III

Results of commonly applied statistical tests*

No.	No. of months	Mean		Standard deviation		Coefficient of determination (D)	Coefficient of efficiency (E)
		observed	estimated	observed	estimated		
1	96	70.4	70.3	91.0	88.4	0.98	0.97
2	96	63.6	63.6	91.9	94.6	0.99	0.99
3	96	59.1	58.1	94.1	91.6	0.99	0.99
4	72	98.3	99.3	51.1	54.0	0.91	0.90
5	60	84.5	83.0	108.2	109.5	0.94	0.93
6	60	25.6	25.6	61.0	58.1	0.99	0.98
7	48	44.6	45.0	52.6	48.0	0.94	0.94
8	132	7.26	7.41	12.95	11.07	0.85	0.80
9	132	17.2	18.1	21.5	19.8	0.76	0.76
10	120	91.6	94.1	57.4	56.9	0.78	0.76
11	96	2.03	1.72	5.18	4.58	0.41	0.35
12	105	2.04	2.08	4.10	5.82	0.46	– 0.40
13	49	15.9	17.3	20.2	20.3	0.90	0.89
14	96	24.3	24.4	31.2	29.9	0.81	0.80
15	60	44.0	43.5	60.1	46.8	0.92	0.88
16	95	7.11	6.60	16.4	20.0	0.88	0.86
17	90	8.66	8.48	18.2	19.9	0.81	0.75

* All runoffs are expressed in millimetres and are for a period of one month.

TABLE IV
Results of sign tests

No.	Result of significance test	Expected No. of runs	Observed No. of runs
1	N.S.*1	49	41
2	N.S.	49	36
3	0.05*2	49	34
4	0.05	37	23
5	N.S.	31	23
6	0.005	31	13
7	N.S.	25	21
8	0.005	67	37
9	0.025	67	46
10	0.025	61	42
11	N.S.	27.5	23
12*3	0.025	45	28
13	0.025	25.5	13
14	0.025	49	31
15	0.005	31	15
16*3	N.S.	26	18
17*3	N.S.	26.5	22

*1 N.S. not significant at the 0.05 level.
*2 Significant at 0.05 level.
*3 Zero estimated and recorded flow ignored in counting runs.

Coefficient of efficiency

Table III shows that for most of the catchments studied the coefficient of efficiency is very similar to that of the coefficient of determination. In several cases it is slightly lower than D probably due to a small bias in the model. For catchment 12 a value less than zero was obtained because of the large bias existing for low flows in this catchment with the model estimating 61 months of no flow whereas only 36 months of no flow were observed.

Serial correlation coefficient

This test was carried out for catchments 8, 9, 10, 11 and 12 but was not helpful in detecting any difference between the time series of estimated and

TABLE V
Tests of the residual mass curves

No.	Error (%)*1	R*2
1	1.8	0.78
2	− 1.2	0.99
3	6.6	0.98
4	5.4	0.90
5	0.7	0.97
6	2.7	0.97
7	1.7	0.94
8	31.0	0.31
9	− 13.9	0.53
10	− 41.6	− 1.11
11	19.6	0.76
12	− 16.2	0.58
13	1.0	0.89
14	19.2	0.11
15	29.4	0.59
16	− 14.3	0.45
17	1.5	0.90

*1 Error (%) = observed range − estimated range/observed range × 100.
*2 R = residual mass curve coefficient.

observed flows. For the first three catchments the difference in the serial correlation coefficients was not significant at the 0.05 level and on the other two the serial correlation coefficients in themselves were not significant. The remaining catchments were not tested in this manner because for most of these the coefficient of determination (see Table III) was much higher and hence it was even less likely that a significant difference could be detected between the estimated and observed values of the serial correlation coefficient.

Tests which reveal systematic errors

Sign tests

The results of the application of the sign test are shown in Table IV. It will be noticed that the number of runs on all seventeen catchments studied is less than that expected on the basis of random errors. In ten cases a Chi-square test found the difference was significant at the 0.05 level or higher. It therefore appears that these models frequently contain systematic errors even when examined on the relatively long time period of one month.

Maximum range of the residual mass curve

The residual mass curves were computed for the observed and estimated monthly flows on all catchments and the percentage difference in the maximum range of the two curves determined. These differences varied from 1% to over 40% as shown in Table V and in eight cases exceeded 10%.

Residual mass curve coefficient

The residual mass curve coefficient for each catchment studied is also included in Table V. This coefficient gives a very much higher value for three of the catchments than that obtained for D or E. On the other hand the values for six of the catchments are very much smaller than those obtained for the other two coefficients. These results indicate the greater sensitivity of this coefficient to the measurement of the fit between the observed and estimated residual mass curves.

DISCUSSION OF RESULTS

Statistical tests

The commonly used statistical tests are not effective in determining systematic errors in models. This may be of little consequence when a model is very precise (say with D equal to or greater than 0.95). However, for lower values of D systematic errors may exist which are more clearly seen by estimating the value of R (e.g., catchments 10 and 14). On the other hand, if the general seasonal trend of the runoff is simulated by the model, the value of R may be greater than that of D or E (e.g., catchments 11 and 12). In these circumstances the value of D and E may penalise the results of the model too severely.

While the percentage difference in the ranges of the observed and estimated residual mass curves are of interest this test bears no resemblance to a standard statistical measure. It is therefore felt that the residual mass curve coefficient which serves a similar objective is a more appropriate measure.

To determine quickly the existence of a systematic error in a model the simple sign test is the most suitable approach and should always be used.

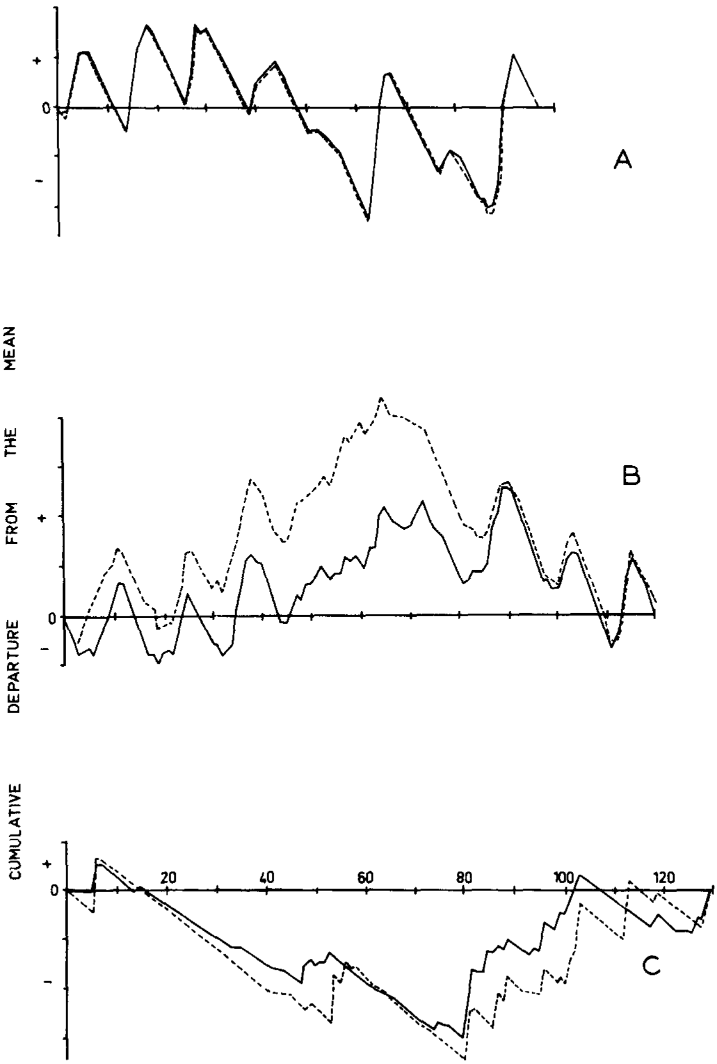

Fig. 1. Residual mass curves for observed and estimated monthly runoffs. A. Russian river at Healdsburg (No. 2); B. Ramu River (No. 10); C. Parwan Weir (No. 12); —— = observed, ---- = estimated.

Graphical displays

Hydrologists have used a range of graphical displays of their results to convey a picture of the models' performance. These displays can apparently either enhance or discredit the results from some models.

Fig. 1, 2 and 3 show the residual mass curves, scatter diagrams and flow duration curves for catchments 2, 10 and 12 and cover the range of results from all the catchments studied.

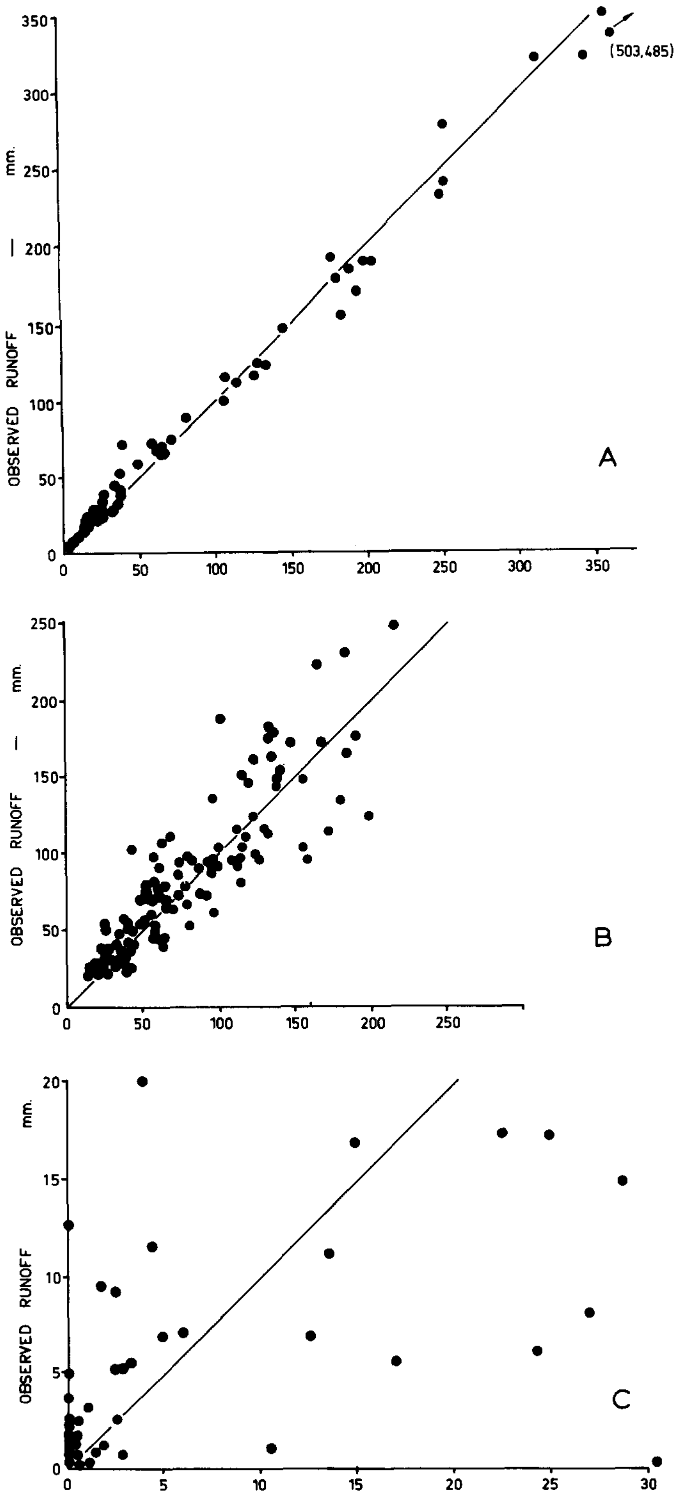

Fig. 2. Scatter diagrams of observed vs. estimated monthly runoff. A. Russian river at Healdsburg (No.2); B. Ramu River (No. 10); C. Parwan Weir (No. 12).

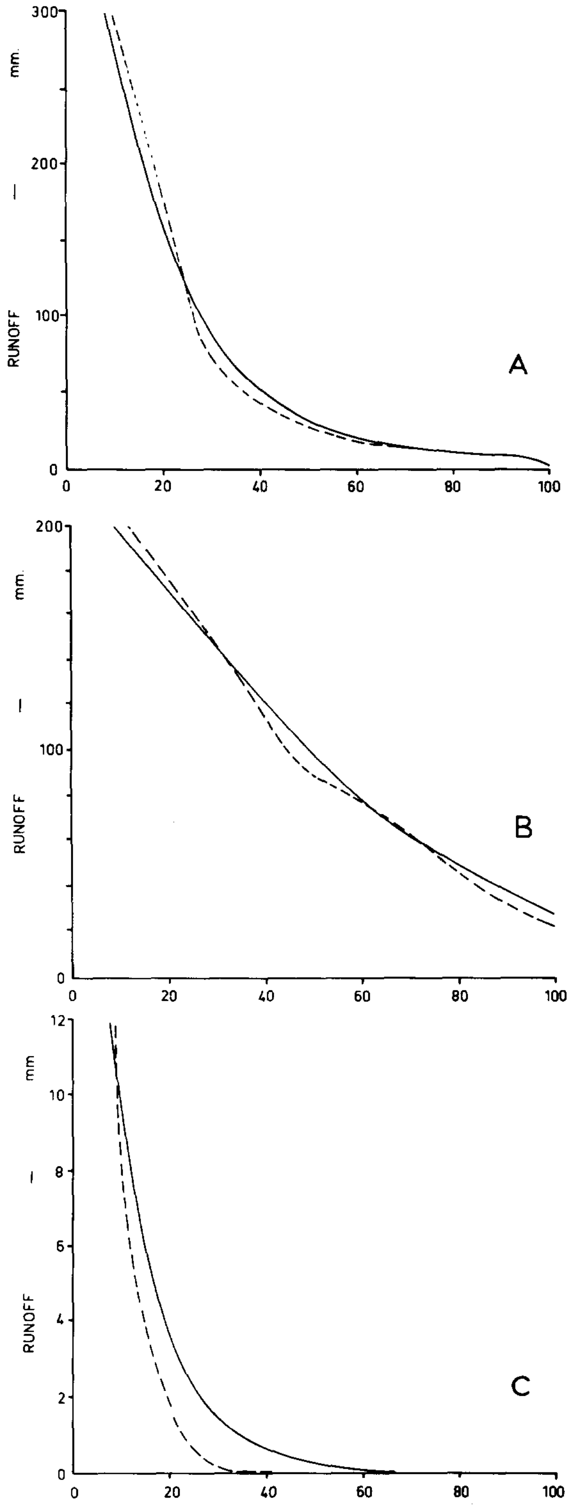

Fig. 3. Flow duration curves for observed and estimated monthly runoffs. A. Russian river at Healdsburg (No. 2); B. Ramu River (No. 10); C. Parwan Weir (No. 12); —— = observed, - - - - = estimated.

For catchment 2 which has values for *D, E* and *R* all equal to 0.99, the three types of graphical displays are all most favourable.

For catchment 10 which has values for *D, E* and *R* equal to 0.78, 0.76 and less than zero, respectively, the scatter diagram and the flow duration curve appear satisfactory but the residual mass curve reveals the existence of a marked systematic error. This therefore, illustrates that, although scatter diagrams and flow duration curves are both commonly used to display results, it is important to remember that they obscure the time sequence of events. (In the case of catchment 10 there is evidence (Ribeny and Brown, 1968, Tables II and IV) that the use of five rain gauges instead of the one gauge used in the derivation of the estimates on which the residual mass curve of Fig. 1 is based could lead to under-estimation or over-estimation of areal mean annual rainfalls for periods as long as 2 years.)

For catchment 12 which has values for *D, E* and *R* of 0.46, less than zero and 0.58, respectively, the scatter diagram is poor. Further, a Kolmogorov-Smirnov test detected a statistically significant difference between the observed and estimated flow duration curves. However, the model by simulating satisfactorily the seasonal flow condition indicates a reasonable fit between the observed and estimated residual mass curves.

CONCLUSION

It has been demonstrated that systematic errors are a common product of rainfall–runoff models and can not be detected by the usually adopted statistical tests. Because these errors are considered to be important in water resources engineering, especially storage analysis, it is recommended that hydrologists apply a simple sign test and compute the residual mass curve coefficient in addition to those statistical tests already commonly applied, when analysing their results.

ACKNOWLEDGEMENTS

The permission of the Snowy Mountains Engineering Corporation to publish this paper is gratefully acknowledged. Thanks are due to a number of the author's colleagues for their constructive criticisms of ideas presented in this paper. The author also wishes to thank the developers of the models used in this paper for supplying details of their results.

The views expressed in this paper are those of the author and are not necessarily endorsed by the Corporation.

REFERENCES

Boughton, W.C., 1966. A mathematical model for relating runoff to rainfall with daily data. *Civ. Eng. Trans., Inst. Eng., Aust.,* CE 8 (1): 83–97.

Brown, J.A.H., 1961. Streamflow correlations in the Snowy Mountains area. *J. Inst. Eng., Aust.,* March 1961: 85–95.

Crawford, N.H. and Linsley, R.K., 1966. Digital simulation in hydrology: Stanford watershed model IV. *Stanford Univ., Dep. Civ. Eng., Tech. Rep.,* No. 39, 210 pp.

Holtan, H.N. and Lopez, N.C., 1970. USDAHL-70 Model of watershed hydrology. *U.S.D.A., Agric. Res. Serv., Tech. Bull.,* No. 1435 (pending), 110 pp.

Jones, J.R., 1970. The estimation of runoff from small rural catchments. *Civ. Eng. Trans., Inst. Eng., Aust.,* CE 12 (2): 133–140.

Mandeville, A.N., O'Connell, P.E., Sutcliffe, J.V. and Nash, J.E., 1970. River flow forecasting through conceptual models, 3. The Ray Catchment at Grendon Underwood. *J. Hydrol.,* 11: 109–128.

Nash, J.E. and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, 1. A discussion of principles. *J. Hydrol.,* 10: 282–290.

Porter, J.W. and McMahon, T.A., 1971. A synthetic model for the simulation of continuous streamflow data. *J. Hydrol.,* 13: 297–324.

Ribeny, F.M.J. and Brown, J.A.H., 1968. The application of a rainfall–runoff model to a wet tropical catchment. *Civ. Eng. Trans., Inst. Eng., Aust.,* CE 10 (1): 65–72.