Intercomparison of conceptual rainfall-runoff models using data from semi-arid research catchments

A.H.M. GÖRGENS
Hydrological Research Unit, Department of Geography, Rhodes University, Grahamstown 6140, South Africa

Abstract

Using a seven-year data set derived from three well-instrumented semi-acid research catchments in South Africa, the performance of four conceptual rainfall-runoff models of differing degrees of complexity was tested. The data set was divided into consecutive warm-up, calibration and verification periods. To evaluate the adequacy of the models for use in water resources studies where storage-yield determinations are based on monthly data, a comprehensive set of statistical tests that measure the ability of the models to reproduce the observed monthly flow series was executed. The results indicate that the relatively simple model based on daily input performed quite as satisfactory, and in some ways, even better than the two complex models which operate on hourly input. The verification period performance of all the models was fairly poor, with the two hourly models respectively yielding best and worst results. None of the four models was entirely convincing in simulating the vastly variable range of runoff generation conditions in semi-arid catchments – however, the general performance of the daily model is encouraging.

Introduction

The past two decades have witnessed the devotion of much research effort by hydrologists and water resources engineers to the development of conceptual rainfall-runoff models for use in the study of the land phase of the hydrological cycle and for the prediction of runoff from rainfall on a continuous basis. For some time now there has been a continuing need that more of these models should be objectively tested and compared.

Noteable intercomparisons of conceptual rainfall-runoff models that have been reported internationally have been those executed by the World Meteorological Organization, 1975 (ten models; six catchments from six different countries; different climates); Moore and Mein, 1975 (three models; four Australian catchments; different climates): Pitman, 1978 (four models; three South African catchments; similar climates); Cembrowicz, Hahn, Plate and Schultz, 1978 (eleven models; unspecified catchments); Weeks and Hebbert, 1980 (five models; three Australian catchments; similar climates); O'Connell and Clarke, 1981 (four models; flood events from a Welsh catchment). In South Africa comparative model studies have also been reported by Cousens and Burney (1976), Pitman (1977) and Roberts (1978).

To enable a potential user of rainfall-runoff models to gain maximum benefit and guidance from a reported intercomparison of models on the same data set, the intercomparison should ideally meet two *minimum* requirements:

- The complete intercomparison should be executed by *persons other than the developers of the models* – to ensure true objectivity and independence and to give an indication of the effort required to obtain and transfer knowledge about each of the models.
- The comparison of model performance must at least partly be based on a *verification data set*, i.e. representative simultaneous input and output data series not used in the calibration or optimization of the models.

Obviously, further requirements can be formulated, but even measured by only these two requirements, very few reported model comparisons would meet the information needs of a potential user searching, for example, for a time-efficient model to use in water resources studies where storage-yield determinations based on monthly runoff data are of prime importance.

It is in the latter context of general water resources studies that this paper examines the performance of four conceptual multiple-storage rainfall-runoff models of differing complexity on each of the three semi-arid Ecca research catchments in the Eastern Cape Province of the Republic of South Africa. Three of the models were developed by Pitman (1973, 1976, 1977) at the Hydrological Research Unit (HRU) at the University of the Witwatersrand, Johannesburg, and in this paper are prefixed by HRU. The first HRU model requires monthly inputs, the second is a daily model and the third operates on hourly input data. The fourth model tested is a version of the well-known Stanford Watershed Model.

This study can be regarded as an extension of work initiated by Pitman (1977, 1978), who compared his own three models with the Stanford model. He used four years of calibration data from three temperate catchments, but particular inadequacies in instrumentation and data consistency hampered his study. No verification data set was used by Pitman. In the present study the model comparison is done by an independent user, relying only on published information about the four models, using data of generally higher reliability, consistency and areal representativeness than Pitman's, and employing both calibration and verification data sets. Furthermore, it is well recognised that the use of semi-arid catchment data creates more stringent testing conditions for conceptual models than data from catchments that are generally wet or generally arid (Chapman, 1975).

Considering the fact that more than 50 per cent of South Africa experiences a semi-arid climate and that there is a serious paucity of reliable streamflow data for this region (Görgens and Hughes, 1982), the results of the study reported in this paper may hold special significance for the evaluation and planning, through modelling, of the surface water resources of these parts of South Africa.

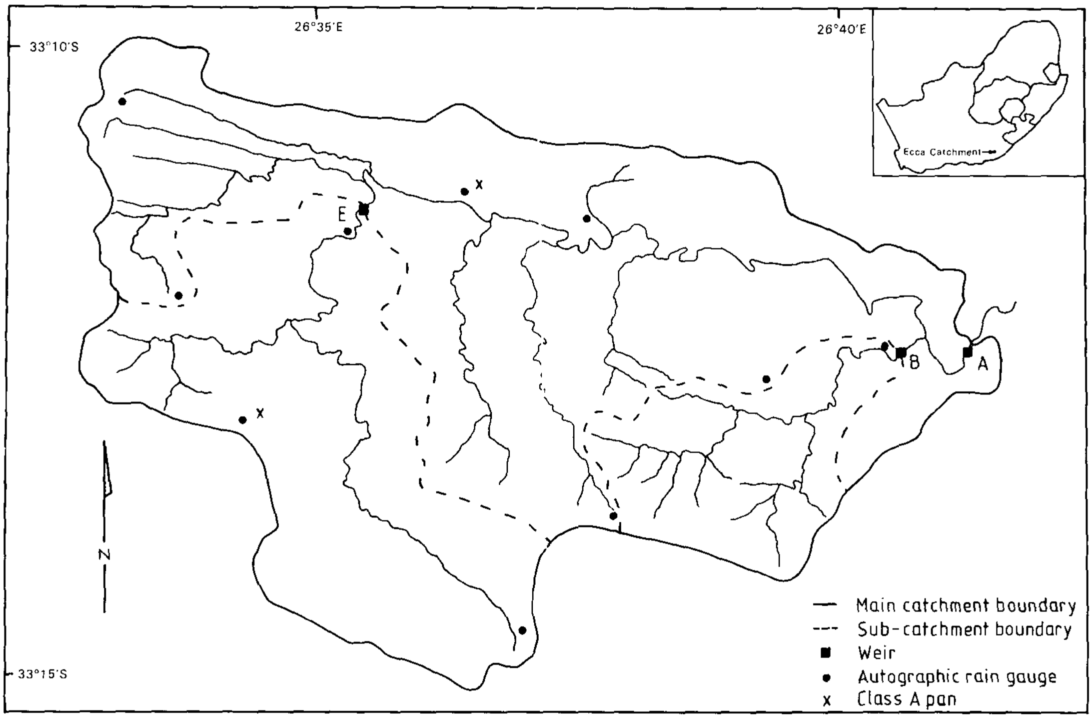

Figure 1
The Ecca research catchments

The Ecca Research Catchments and Data Set

The location of the three Ecca River research catchments, A, B and E, as well as the hydro-meteorological instrumentation network are shown in Fig. 1. A summary of catchment characteristics appears in Table I. The Ecca River is a lower tributary of the Great Fish system which drains the eastern flank of the semi-arid Karoo region of South Africa. Since 1974 the Ecca catchments have been systematically instrumented and continuously monitored by the Hydrological Research Unit of the Geography Department at Rhodes University, Grahamstown, with the express purpose of establishing a growing data bank of reasonable quality on *semi-arid* rainfall-runoff processes.

TABLE I
SUMMARY OF CATCHMENT CHARACTERISTICS

Catchment	A	B	E
Area (km²)	73,1	9,1	24,0
Measured MAP (mm), 1975-1981	487	478	520
(Range)	347-649	346-637	371-682
Measured MAR ($10^3 m^3$), 1976-1981	1691	198	310
(Range)	30-8563	3-1007	0,3-1732
Measured MAR/MAP (%), 1976-1981	4,59	4,38	2,42
Class A pan evaporation/yr (mm)	1846	1813	1840
No of effective autographic raingauges	10	3	5
Zero-flow months (%), 1976-1981	41,7	58,3	81,9
Proportion of area with slopes > 20% (%)*	31	48	21

*According to Roberts (1978).

The climate of the research area is harsh, with large differences between winter and summer extreme and average temperatures. The estimated long-term mean annual precipitation (MAP) of the Ecca varies from about 520 mm in the southwest to about 370 mm in the northeast, with an overall average of 420 mm (Roberts, 1978). Summer precipitation (60% of total) occurs mostly in the form of convectional thunderstorms and winter rain (40% of total) is usually associated with cold fronts from the southwest of the African continent.

Total basin relief of the Ecca catchments is about 570 m and soils are shallow and stony on ridgetops and slopes, with deeper colluvial deposits in the valleys. Vegetation consists of sparse sub-succulent woodland which thins to low succulent scrub on the flatter areas. There is a high proportion of bare ground through most of the year. The research catchments are uncultivated and are used exclusively for small livestock farming.

All streams in the research area are ephemeral with very low proportions of rainfall actually resulting in runoff as Table I shows. The likelihood of real differences in hydrological response among the individual Ecca catchments, suggested by Table I, has been borne out by analyses made by Görgens (1980), and Murray and Görgens (1981).

The hydro-meteorological network of relevance to this study consists of ten autographic raingauges, two Class A evaporation pans and three sharp-crested multiple notch weirs (Fig. 1; Table I). Data available for the study consist of the following: hourly rainfall totals for the period January 1975 to December 1981; daily Class A pan evaporation values at the more northerly site for the period January 1975 to December 1981 and at the more southerly site for the period February 1980 to December 1981; instantaneous hourly flow rates at all the weirs for the period

January 1976 to December 1981. Hourly rainfalls were averaged by Thiessen polygon methods to yield catchment rainfall for input to the hourly models, while daily or monthly inputs to the daily or monthly models were merely totalled from the hourly catchment averages.

Estimates of potential evapotranspiration were made by applying smoothed regional pan factors to the gross evaporation measured at the Class A pan sites. Average daily pan values for the overlapping period at the two pan sites were based on the proportion of each catchment area below (northerly pan values) and above (southerly pan values) the 426 m contour.

TABLE 2
COMPARISON OF MODEL CHARACTERISTICS

Model	Number of conceptual storages	Number of parameters	Input/ Output	Execution time* (s)	Average no. calibration runs per catchment
HRU monthly	2	12	1 month	7	33
HRU daily	3	13	1 day	41	19
HRU hourly	4	15	1 hour	169	49
Stanford	4	19**	1 hour	204	31

*CPU time on ICL 1904 computer for a 6 year calibration run.
**Excluding snowmelt parameters which are not applicable in the study area.

The Models

A brief summary of the characteristics of the models used in this study is given in Table 2. The so-called Anderson version of the Stanford Watershed Model (Parmele, 1972; Roberts, 1978) was used in its single-catchment, hourly input mode. This version differs from that reported by Crawford and Linsley (1966) in that the infiltration function was made more flexible at the cost of an additional calibration parameter, and the channel routing subroutine was modified to accept a time delay histogram of quarter hour increments – a necessity for smaller catchments.

The monthly and daily versions of the HRU models are described by Pitman (1973, 1976), in which the model programs (in FORTRAN) are also provided. However, the computer program of the hourly version had to be extracted from a flood-forecasting study by Pitman and Basson (1979), where it is reported in a forecast mode, in tandem with the daily version. This hourly version was then changed into a simulation mode, so that it could be run independently of the daily version. As all three HRU models were designed to operate on regional longterm mean monthly Symons pan evaporation totals, the evapotranspiration/soil moisture status functions of all three models had to be modified to operate on continuous A-pan evaporation records based on daily values. Care was taken to keep these modifications faithfully close to Pitman's (1973, 1976) basic assumptions.

Model Calibration and Verification Procedures

After some intial experimentation it was decided to follow a "three-stage" calibration/verification strategy in this study. The year 1975, for which no runoff data exist for any of the catchments, was used as a "warm-up" year for all four models. Data for the years 1976 to 1980, inclusively, were used as a calibration sample in all cases and the year 1981 as a verification sample. A fair number of runoff events occurred in 1981, which therefore represented a reasonable verification sample on its own. In this way, it was hoped, robust model parameters could be ensured by the largest possible calibration sample and that an extraordinary sequence of very high rainfall in 1979 (649 mm for catchment A) which included two multiple-day winter rainstorms of large return period, followed by a very dry 1980 (347 mm for A) which produced only one runoff event towards December, would especially strengthen the model calibrations.

Calibration of the models was undertaken both manually and by automatic optimization routine. The latter is part of a larger study still in progress and this paper reports only on the manual calibration approach. Manual calibration is deemed the most likely approach by a potential conceptual model user in a non-research environment, because of the computer time costs associated with automatic optimization.

The general manual calibration strategy was to focus on subjective comparisons of simulated and observed monthly totals and, sometimes, individual hydrographs, while all the time attempting to keep certain objective criteria of model performance within an acceptable range of their optimum values. These objective criteria were: mean annual runoff (MAR); variance of monthly totals; coefficient of efficiency based on individual monthly means (Garrick, Cunnane and Nash, 1978); and residual mass curve coefficient (Aitken, 1978).

The procedure used for each model was to choose initial parameter values for catchment A according to guidance provided in the relevant model documentation. Once an approximate fit was achieved on A, the new parameters were transferred to B and E and approximate model fits on these cathments were sought. Final calibrations on all three catchments were then attempted more or less simultaneously for each model. It was hoped that the above procedure would make maximum use of any possible internal consistency of each model's parameters among such "similar" catchments and that such a "feed-back" loop based on the similitude of the three modelling situations would ensure sound, if not optimum, final parameters.

The verification test was done by running each model with the full seven years of input data using the above final parameters and assessing the models' performance separately for each of the calibration period, the verification period and the total runoff period. It is now well accepted that the performance of conceptual rainfall-runoff models can best be evaluated by assessing a variety of different statistics and tests that measure the ability of the models to reproduce the observed flow series, followed by a subjective decision made on the basis of this assessment. In this study a set of 11 criteria were utilized, calculated on the basis of simulated and observed *monthly* totals. These criteria are described in the Appendix.

Comparison of Model Performance

A visual impression of the nature of the observed flow record used for the calibration and verification periods, respectively, as well as the one-to-one fit achieved by the four models, may be gained by inspection of Figs. 2 and 3. These diagrams show the typically erratic nature of the semi-arid runoff regime and highlight the magnitude of the 1979 winter flows compared to the rest of the available record. All the models seem to fit various parts of the calibration record for catchment A with varying

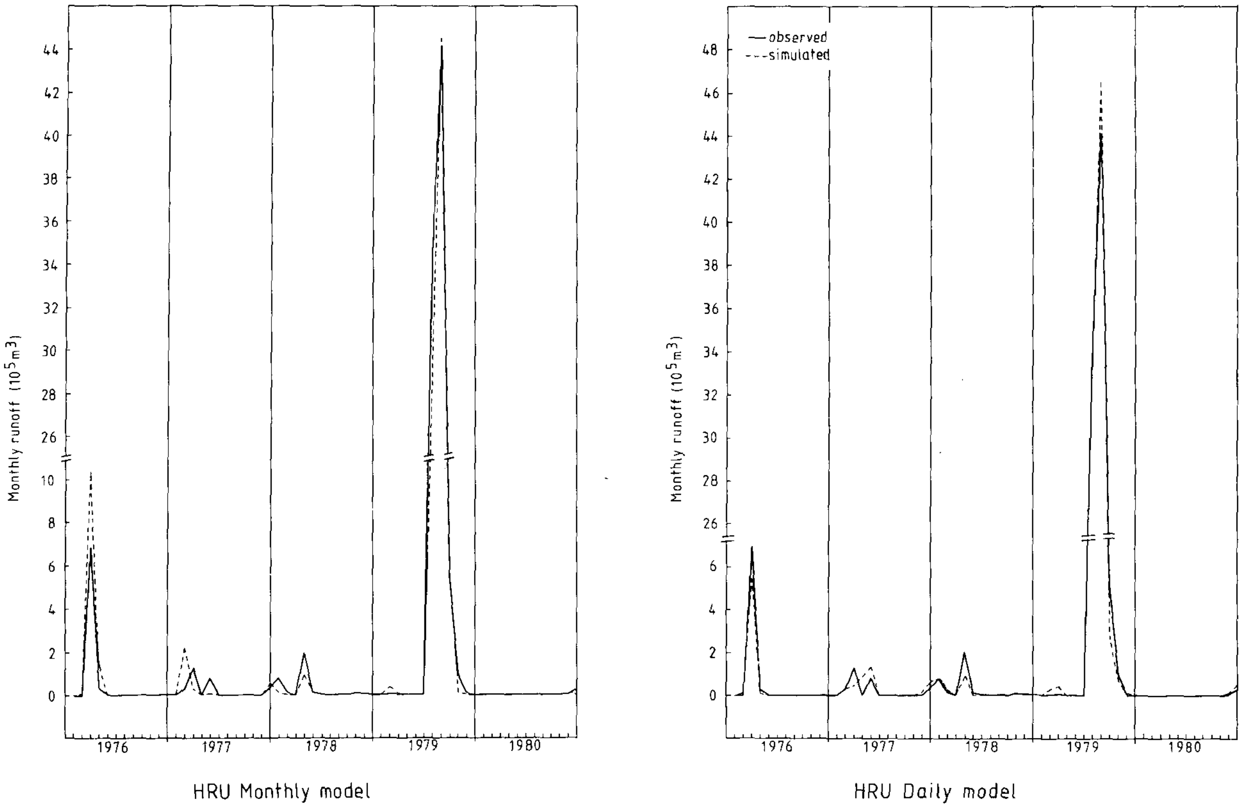

Figure 2(a)
Catchment A – calibration period 1976-1980: comparison of observed and simulated flows: HRU monthly and daily models.

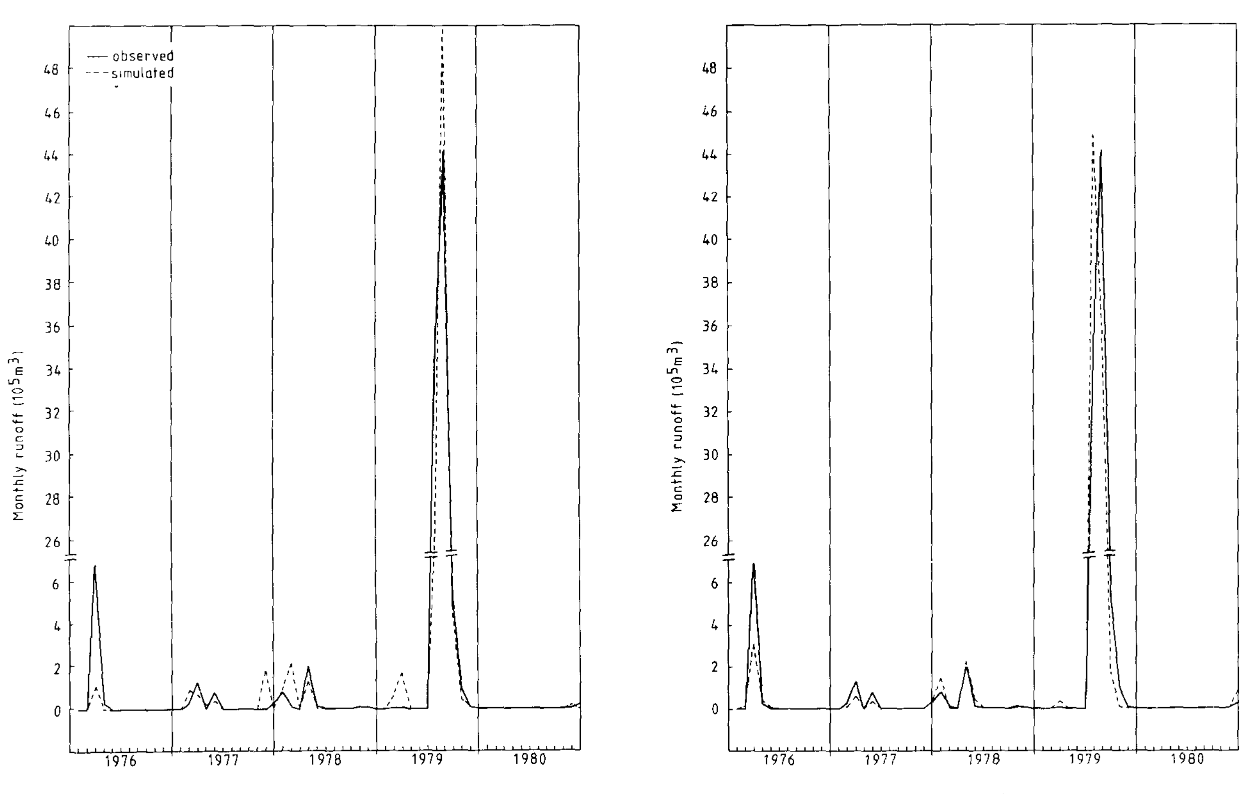

Figure 2(b)
Catchment A – calibration period 1976-1980: comparison of observed and simulated flows: HRU hourly and Stanford models.

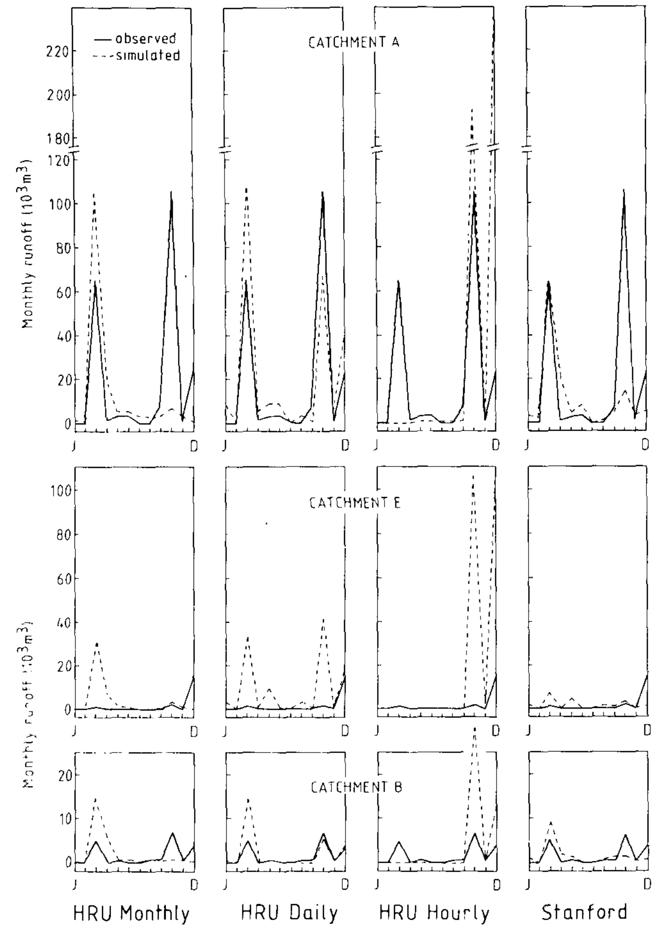

Figure 3
All catchments – verification period 1981: comparison of simulated and observed flows.

degrees of success. (This observation also holds for the other two cathments). However, Fig. 3 shows distinct differences in the calibrated models' individual abilities to predict runoff for the verification year. Based on visual comparison for all three catchments it may be concluded that for the calendar year 1981 the HRU hourly model (as calibrated) shows the least predictive ability, with the Stanford model performing slightly better than the HRU daily and monthly versions.

The numerical criteria of goodness-of-fit of all the model/catchment combinations for the complete flow record period are shown in Table 3. Selected criteria for the verification period are given in Table 4. As the relative performance of the models according to the criteria for the calibration period is little different from that for the total flow-record period, only the latter period's criteria are shown, for reasons of space economy.

Judging by the mean values in Table 3 of the respective criteria for each model, it can be seen that the performance of the Stanford and HRU daily models are more or less equivalent, while the HRU hourly and monthly models are comparable. The numerical goodness-of-fit values for the Stanford and HRU daily models are generally "better" than those for the other two, but the nominal differences between the two sets of values for the two pairs of models are not dramatic. Table 3 shows that on average all models reproduce the overall water balance well enough. However, the HRU monthly model tends to underestimate the monthly variance (due to generation of too high baseflows), and the Stanford model exaggerates the montly variance (high flows too peaked). The CE and k values indicate high overall one-to-one correspondence for all the models on catchments A and B but a much lower quality of fit for the HRU hourly and monthly models on catchment E. At this juncture, it should be pointed out that the CE and k values may be disproportionately influenced by the closeness of the one-to-one-fit on the high winter flows of 1979, mentioned earlier, and should not be afforded too much importance.

The small persistence in model residual errors suggested by the high R and low P values in Table 3 is deceptive, as the

TABLE 3
RESULTS OF STATISTICAL TESTS* ON MONTHLY FLOWS FOR THE COMBINED CALIBRATION AND VERIFICATION PERIODS**

Model	Catchment	M(%)	V(%)	CE	k(%)	R	P	r(%)	I(%)	MD(%)	ST significant at 0,05	DT significant at 0,05
HRU monthly	A	-3,5	-9,6	0,98	1,2	0,97	1,71	-7,0	-5,3	0,0	no	no
	B	-1,6	-1,9	0,94	5,9	0,99	1,01	-3,1	-1,8	0,0	yes	no
	E	-7,1	-14,5	0,69	10,1	0,92	1,39	-19,4	-17,3	-28,2	no	no
	mean	-4,1	-8,7	0,87	5,7	0,96	1,37	-9,8	-8,1	-9,4		
HRU daily	A	-2,2	3,7	0,99	0,7	0,99	1,09	-1,4	5,2	39,3	yes	no
	B	-3,7	7,9	0,95	5,3	0,99	0,86	-3,9	5,8	0,0	yes	no
	E	8,8	-3,5	0,86	6,9	0,97	0,36	-13,4	-19,6	-35,9	no	no
	mean	1,0	2,7	0,93	4,3	0,98	0,77	-6,2	-2,9	1,1		
HRU hourly	A	-3,7	-1,6	0,94	1,9	0,97	1,11	-4,8	-1,0	-14,3	yes	no
	B	1,4	5,3	0,86	8,5	0,98	1,07	-1,7	-4,2	0,0	no	no
	E	11,4	-2,7	0,73	9,5	0,92	0,56	-7,9	-27,9	-5,1	no	no
	mean	3,0	0,3	0,84	6,6	0,96	0,91	-4,8	-11,0	-6,5		
Stanford	A	-7,4	0,4	0,92	2,2	0,99	0,61	-3,5	11,2	0,0	no	no
	B	-8,0	14,1	0,91	6,9	0,99	0,83	-5,6	14,9	0,0	yes	no
	E	0,0	12,3	0,97	3,1	0,99	0,71	-1,6	7,5	0,0	yes	no
	mean	-5,1	8,9	0,93	4,1	0,99	0,72	-3,5	11,2	0,0		

*See Appendix for explanation of symbols
**1976-1981, inclusive

TABLE 4
SELECTED STATISTICS OF MODEL PERFORMANCE FOR THE VERIFICATION PERIOD*

Model	Catchment	M(%)	V(%)	k(%)	r(%)
HRU monthly	A	- 24,9	- 23,9	23,6	7,9
	B	39,4	226,1	134,4	122,0
	E	166,4	333,4	207,7	127,7
	mean	60,3	178,5	121,9	85,9
HRU daily	A	19,3	- 0,7	12,8	8,8
	B	41,6	241,4	100,1	75,1
	E	562,3	1021,2	313,8	204,5
	mean	207,7	420,6	142,2	96,1
HRU hourly	A	108,8	536,0	51,1	260,4
	B	159,0	1510,1	273,0	364,7
	E	1204,1	9162,4	820,2	1084,5
	mean	490,6	3736,2	381,4	569,9
Stanford	A	- 35,9	- 72,4	20,2	- 30,2
	B	3,6	18,4	82,4	16,6
	E	32,6	- 72,2	97,4	- 53,5
	mean	0,1	- 42,1	66,7	- 22,4

*Calendar year 1981

residual mass curve plots in Fig. 4 show. These diagrams reveal error persistence in at least the HRU hourly and monthly model simulations. All the models consistently underestimate the range, r, of the residual mass curves, but show little consistency and a low success rate in estimating the index of seasonality, I. The three HRU models all produce one or more serious errors in estimating the maximum deficient flow period. A presence of systematic errors is indicated by the sign test for at least one catchment in the case of all the models, but no correlated errors can be detected.

The apparently poor predictive performance of the models (as calibrated) depicted in Fig. 3 is further substantiated by Table 4. The three HRU models grossly overestimate total runoff, variance and range of residual mass curve, while the Stanford model shows random variation in errors among the catchments. The best single prediction is achieved by the HRU daily model on catchment A, but the earlier suggestion that the Stanford model shows better predictive ability for 1981, based on overall performance, is further underlined by the values in Table 4.

Discussion

In the light of the apparent adequacy of the performance of all the models on most of the catchments suggested by Table 3, the

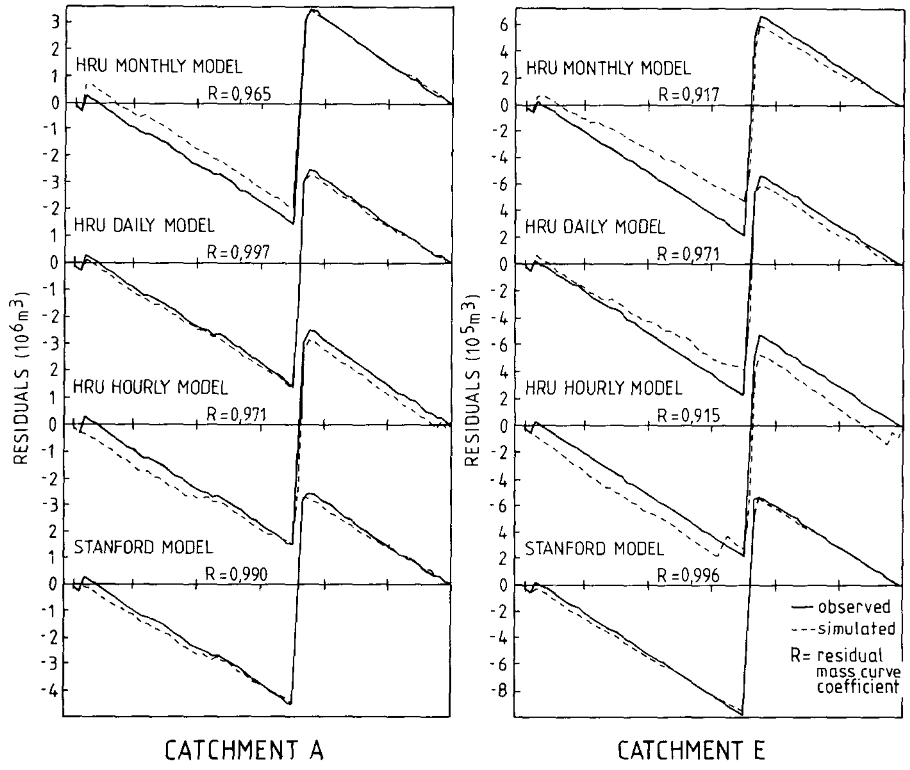

Figure 4
Catchments A and E – combined calibration and verification period 1976-1981: comparison of simulated and observed residual mass curves.

generally poor predictions for 1981 are somewhat surprising, if not puzzling. Some or all of the following arguments and counter-arguments will constitute an explanation of the verification results.

The calibration of the models may have been biased towards the reproduction of high runoff events by the presence of the winter 1979 flows in the calibration record, while 1981, the verification year, was a "normal" runoff year: This implies that the calibration sample was too small or too dominated by a few extreme events. The predictive performance of the Stanford model seems to negate this explanation, but perhaps the study reveals that the HRU models need a much longer calibration period than the Stanford model for the same quality of optimization on semi-arid data.

The calibration itself may have been inexpertly executed, resulting for some catchment/model combinations in non-optimum parameter sets: The verification experience with the Stanford, HRU daily and HRU monthly models on catchment A appears to refute this point. Also, the calibration procedure described in the previous section does seem intuitively sound.

The verification sample may include rainstorm types poorly represented in the calibration sample: Of the bigger runoff events of 1981 the ones earlier in the year were the result of widespread frontal rainfall while the events later in the year were caused by high-intensity convectional thunderstorms. Although these storm types are both well represented in the calibration sample it was noted that localised thunderstorm runoff was poorly simulated by all the models at some points during the calibration period. The verification plots (Fig. 3) suggest that the HRU hourly model was calibrated to be *hypersensitive* and the HRU monthly and the Stanford models to be *insensitive* to high-intensity rainfall.

The models may inherently be inadequate for the simulation of rainfall-runoff processes in the semi-arid environment: All model/catchment combinations yielded one or more dramatic simulation failures for specific rainfall input series. Also, the three HRU models tended to sustain baseflow for too long while the Stanford model produced too little baseflow.

Conclusions

A comprehensive set of statistical tests on simulated monthly flows for a combined calibration and verification period of six years indicated that there is no great difference in performance of the simpler HRU monthly and daily input models as compared with the more complex HRU hourly and Stanford models. In fact, the HRU daily model along with the Stanford model, can be regarded as yielding the best overall results.

All the models showed a serious divergence in values of selected goodness-of-fit criteria moving from the calibration to the verification period. For the verification period (calendar year 1981) the Stanford model showed the highest predictive ability, followed by the HRU daily and monthly models.

The large errors yielded by the verification tests may be attributed to any one or a combination of factors such as calibration sample too small, inadequate calibration causing poor ability to simulate runoff from high-intensity rainstorms, and inherent model inadequacies.

None of the four models was entirely convincing in simulating the vastly variable range of runoff-generation conditions in semi-arid catchments, and all model/catchment combinations yielded one or more dramatic simulation failures for specific rainfall input series. However, the general performance of the relatively simple HRU daily model is encouraging.

As a general conclusion, it can be stated that the results point out the foolishness of blindly adopting the most complex available model for a water resources study where, for example, storage-yield determinations are based on monthly flow data. For a semi-arid catchment, a daily input model is probably quite adequate on condition that a comprehensive calibration of the model is possible.

Acknowledgements

This study forms a part of a long-term hydrological research project funded by the Water Research Commission of South Africa. The author is indebted to Jenny Gibson of the Hydrological Research Unit at Rhodes University for invaluable assistance during the data preparation and calibration phases of the study, for the drawing of the diagrams and for typing the manuscript.

References

AITKEN, A.P. (1973) Assessing systematic errors in rainfall-runoff models. *J. Hydrol.* **20** 121-136.

CEMBROWICZ, R.G., HAHN, H.H. PLATE, E.J. AND SCHULTZ, G.A. (1978) Aspects of present hydrological and water quality modelling. *Ecological Modelling* 5 39-66.

CHAPMAN, T.G. (1975) Trends in catchment modelling. In: T.G. Chapman and F.X. Dunin (Eds.) *Prediction in catchment hydrology*, Aust. Acad. Sci., Canberra, pp. 459-482.

COUSENS, D.W.H. and BURNEY, J.R. (1976) Small catchment flood modelling. *Water S.A.* **2**(4) 150-155.

CRAWFORD, N.H. and LINSLEY, R.K. (1966) Digital simulation in hydrology: Stanford Watershed Model IV. Dep. Civ. Eng., Stanford Univ., Stanford, Calif., Tech. Rep. No. 39.

GARRICK, M., CUNNANE, C. and NASH, J.E. (1978) A criterion of efficiency for rainfall-runoff models. *J. Hydrol.* **36** 375-381.

GÖRGENS, A.H.M. (1980) Engineering implications of results of rainfall-runoff research in semi-arid catchments. Proceedings of the 1980 Symposium on Agricultural Engineering, S.A. Institute of Agricultural Engineers, pp. B5-B11.

GÖRGENS, A.H.M. and HUGHES, D.A. (1982) Synthesis of streamflow information relating to the semi-arid Karoo Biome of South Africa. *S. Afr. J. Sci.* **78**.

MOORE, I.D. and MEIN, R.G. (1975) An evaluation of three rainfall-runoff models. Proceedings of the Hydrology Symposium, Institution of Engineers of Australia, Armidale, pp. 122-126.

MURRAY, D.L. and GÖRGENS, A.H.M. (1981) Storm runoff analyses on three semi-arid catchments. *Water S.A.* 7(4) 223-233.

O'CONNELL, P.E. and CLARKE, R.T. (1981) Adaptive hydrological forecasting - a review. *Hydrol. Sci. Bull.* **26**(2) 179-205.

PARMELE, L.H. (1972) Errors in output of hydrologic models due to errors in input potential evapotranspiration. *Water Res. Res.* **8**(2) 348-359.

PITMAN, W.V. (1973) A mathematical model for generating monthly river flows from meteorological data in South Africa. Hydrol. Res. Unit, Univ. Witwatersrand, Johannesburg, Rep. No. 2/73.

PITMAN, W.V. (1976) A mathematical model for generating daily flows from meteorological data in South Africa. Hydrol. Res. Unit, Univ. Witwatersrand, Johannesburg, Rep. No. 2/76.

PITMAN, W.V. (1977) Flow generation by catchment models of differing complexity - a comparison of performance. Hydrol. Res. Unit, Univ. Witwatersrand, Johannesburg, Rep. No. 1/77.

PITMAN, W.V. (1978) Flow generation by cathment models of differing complexity - a comparison of performance. *J. Hydrol.* **38** 59-70.

PITMAN, W.V. and BASSON, M.S. (1979) Flood forecasting for reservoir operation. Hydrol. Res. Unit, Univ. Witwatersrand, Johannesburg, Rep. No. 1/79.

ROBERTS, P.J.T. (1978) A comparison of the performance of selected conceptual models of the rainfall-runoff process in semi-arid catchments near Grahamstown. Hydrol. Res. Unit, Dep. Geog., Rhodes Univ., Grahamstown, Rep. No. 1/78.

SOROOSHIAN, S. and DRACUP, J.A. (1980). Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: correlated and heteroscedastic error cases. *Water Resour. Res.* **16**(2) 430-442.

WALLIS, J.R. and TODINI, E. (1975) Comment upon the residual mass curve coefficient. *J. Hydrol.* **24** 201-205.

WEEKS, W.D. and HEBBERT, R.H.B. (1980) A comparison of rainfall-runoff models. *Nordic Hydrol.* **11** 7-24.

WORLD METEOROLOGICAL ORGANIZATION (W.M.O.) (1975) Intercomparison of conceptual models used in operational hydrological forscasting. Operational Hydrology Rep. No. 7, W.M.O., Geneva.

Appendix

Notation – Symbols used in the appendix

A_n	area of positive or negative residual run
d	departure from the mean for the residual mass curve
$\bar{d}$	mean of the residual mass curve
I_s	range of the residual mass curve of overall mean monthly flow distribution
i	index of items included in either observed or simulated flow series, or residual mass curve series
j	index of calendar month, i.e. j = 1 to 12
N	total number of months in record
n	index of positive or negative residual run
r_a	range of residual mass curve
v_x	variance of observed flow series
v_y	variance of simulated flow series
x_i	observed monthly flows
$\bar{x}$	mean of observed monthly flows
$\bar{x}_{mj}$	mean of observed monthly flows for calendar month j
y_i	simulated monthly flows
$\bar{y}$	mean of simulated monthly flows

Definition of statistical goodness-of-fit criteria

(1) Per cent error in *mean annual runoff*, M:
$M = 100(\bar{y} - \bar{x})/\bar{x}$

(2) Per cent error in *monthly variance*, V:
$V = 100(v_y - v_x)/V_x$

(3) Two related, but different, criteria of *one-to-one fit* – the coefficient of efficiency based on individual monthly means, CE (Garrick, Cunnane and Nash, 1978):
$CE = 1 - \Sigma(x_i - y_i)^2/\Sigma(x_i - \bar{x}_{mj})^2$
and the maximum equivalent "constant" error, k (Weeks and Hebbert, 1980) which is calculated from
$C_w = k\,[C_v{}^2(N - 1)/N + 1]^{0,25}$, where
$C_w = [\Sigma(X_i - y_i)^2/N]^{0,5}/\bar{x}$ and
$C_v = v_x/\bar{x}^2$

(4) Two different criteria to measure *persistence* in model residual errors – the residual mass curve coefficient, R (Aitken, 1973):
$R = 1 - \Sigma(d_{xi} - d_{yi})^2/\Sigma(d_{xi} - \bar{d}_x)^2$
and the coefficient of persistence, P, suggested by Wallis and Todini (1975):
$P = \Sigma A_n{}^2/\Sigma(x_i - y_i)^2$

(5) Two different indexes relating to *storage requirements* for flow regulation – per cent error in the range of the residual mass curve, r:
$r = 100(r_{ay} - r_{ax})/r_{ax}$
and per cent error in the index of seasonal variability
$I = 100(I_{sy} - I_{sx})/I_{sx}$

(6) A simple measure of *low flow* simulation – percent error in maximum deficient flow duration, MD, i.e., the maximum continuous duration of flow smaller than the mean monthly flow.

(7) A measure of the existence of *systematic errors* in the simulated time series – the one sample runs test or "sign test", ST (Aitken, 1973).

(8) A measure of the existence of *correlated errors* in the simulated time series – the Durbin-Watson-statistic, DT (Sorooshian and Dracup, 1980).