CALIBRATION AND VALIDATION OF HYDROLOGICAL MODELS

INTRODUCTION

Hydrological models are the mathematical models having some unknown co-efficients known as parameters. Model calibration means the estimation of those parameters from historical input-output records. Model validation means judging the performance of the calibrated model over that portion of historical records which have not been used for the calibration.

For model calibration the methods, which have been commonly used, include (i) manual parameter assessment using 'Trial and Error' procedure, (ii) automatic parameter assessment using numerical optimisation procedure and (iii) a combination of (i) and (ii). For the model validation, various validation criteria, developed based on the observed and computed output records, are used.

In this lecture the following aspects of the hydrological modelling have been discussed in brief:

(i) Hydrological processes considered in stream flow simulation models
(ii) Hydrological Modelling procedures
(iii) Concept of deterministic mathematical modelling and sources of uncertainity.
(iv) Goodness of fit and accuracy criteria
(v) Model Calibration and validation methods
(vi) Model validation including the schemes for Systematic Validation of simulation models
(vii) Sensitivity analysis; and
(viii) Extrapolation from calibration conditions etc.

HYDROLOGICAL PROCESSES CONSIDERED IN STREAM FLOW SIMULATION MODELS

Various stream flow simulation models generally consider the following hydrological processes to simulate the time series of stream flow.

(a) Land Surface Processes
  (i) Interception
  (ii) Infiltration
  (iii) Overland flow
  (iv) Evapotranspiration
  (v) Snow accumulation and Melt

(b) Sub-surface Processes
  (i) Interflow
  (ii) Soil moisture storage and Movement
  (iii) Ground water storage and flow

(c) Channel Processes
  (i) Channel flow
  (ii) Flood plain storage
  (iii) Lakes, Reservoirs and Diversions

HYDROLOGICAL MODELLING PROCEDURES

The following procedures are usually followed for Hydrological Modelling:

- Develop a suitable model structure to simulate various component processes keeping in mind the quantity and quality of the data available and nature of the problems for which the modelling is required.
- Calibrate the developed model using the historical records.
- Validate the model using the historical records which have not been considered for calibration.
- Perform sensitivity analysis study to identify the most sensitive parameters of the model which require proper investigation before arriving at the final parameter values.
- Use the calibrated and validated model for solving the specific hydrological problem for which the development of the model is intended for.

CONCEPT OF DETERMINISTIC MATHEMATICAL MODELLING AND SOURCES OF UNCERTAINTY

The concept of deterministic simulation can be illustrated as in Fig.1, where the physical system, in this case a catchment, is shown on the left, and the mathematical model is shown on the right. The model is a simplified representation of the physical system.

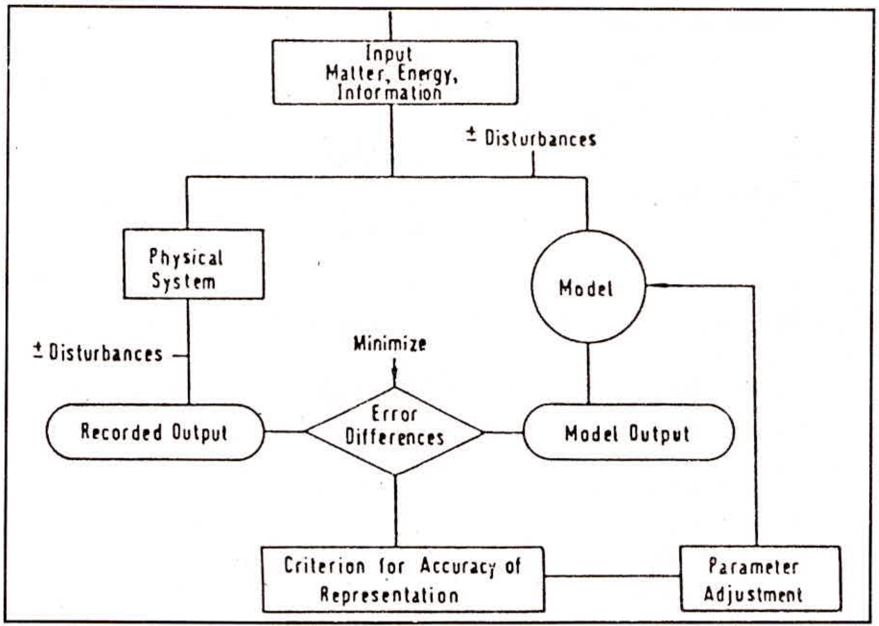

Fig. 1 : The concept of deterministic mathematical modelling (After Fleming, 1975).

Basically four sources of uncertainty occur in deterministic simulation, the disagreements between recorded and simulated output resulting from:

1. Random or systematic errors in the input data, e.g. precipitation, temperature, or evapotranspiration used to represent the input conditions in time and space for the catchment.

2. Random or systematic errors in the recorded output data, e.g. water level or discharge data used for comparison with the simulation output.

3. Errors due to non-optional parameter values.

4. Errors due to incomplete or biased model structure.

Thus, during the calibration process only error source 3 is minimized, whereas the disagreement between simulated and recorded output is due to all four error sources. The measurement errors and errors source 2 serve as a 'background noise' and give a minimum level of disagreement below which further parameter or model adjustments will not improve the results. The objective of a calibration process is then to reduce the error source 3 until it is insignificant compared with the data error sources 1 and 2.

During a calibration process it is of the utmost importance to ensure that a clear distinction is drawn between the different error sources, so that it is not attempted to compensate for errors for one source by adjustment within another source, e.g. compensate for a data error by parameter adjustments. Otherwise the calibration will degenerate to curve fitting, which may result in a reasonable fit within the calibration period but will inevitably give poor simulation results for other periods. In the following five examples it would be physically incorrect and fatal for future predictions to try to compensate for the following discrepancies between recorded and simulated flows using parameter adjustments:

* Both flood peak and runoff volume for a hydrograph are under predicted, owning to an underestimation of the average precipitation, Error source-1.

* Discrepancies are observed between simulated and recorded flow in a period where the recorded flow is known to be very uncertain owing to problems with the rating curve. Error source 2.

* A flood peak is under predicted as a result of embankments being breached whereas the model has been developed assuming non-breaching embankments. Error source 4.
* Travel time for high flows is smaller than the travel time for low flows but the routing model is linear with the travel time independent of flow regime. Error source 4.
* The base flow in low flow periods decreased during the calibration period owing to ground water abstraction and lowering of the ground water tables but ground water abstruction cannot be accounted for directly in the applied model. Error source 4.

GOODNESS OF FIT AND ACCURACY CRITERIA

During the calibration procedure an accuracy criterion can be used to compare the simulated and measured outputs. This enables an objective measure of the goodness of fit associated with each set of parameters to be obtained and the optimum parameter values to be identified. However, selection of an appropriate criterion is greatly complicated by the variation in the sources of error discussed in the last section. It further depends on the objective of the simulation (e.g. to simulate flood peaks or hydrograph shape) and on the model output variable, e.g. phreatic surface level, soil moisture content, stream discharge or stream water level. No single criterion is entirely suitable for all variables and even for a single variable it is not always easy to establish a satisfactory criterion. Hence a large number of different criteria has been developed. The most widely used criterion is the sum of the squares of the deviations between recorded and simulated values of a variable:-

$$F^2 = \sum_{i=1}^{n} (QOBS_i - QSIM_i)^2 \qquad (2)$$

where

F^2 = index of disagreement, or objective function
$QOBS_i$ = recorded value at time step i
$QSIM_i$ = Simulated value at time step i
n = number of values (time steps) within the considered time period

All values of $QOBS_i$ and $QSIM_i$ are based on a time step which may be e.g. one hour, one day or one month. One disadvantage with this criterion is that F^2 is dimensional (e.g. $(m/s)^2$). Therefore, the following nondimensional form is often used:

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(QOBS_i-\overline{QOBS})^2-\frac{1}{n}\sum_{i=1}^{n}(QOBS_i-QSIM_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(QOBS_i-\overline{QOBS})^2} \qquad (3)$$

where

$$(\overline{QOBS}) = \frac{1}{n}\sum_{i=1}^{n} QOBS_i$$

R^2 is often denoted the coefficient of determination, the explained variance or the model efficiency. R^2 can vary from -0 to +1, where $R^2 = +1$ represents a complete agreement between recorded and simulated values. It is noted that the simple one parameter model $QSIM_i = \overline{QOBS}$ will give $R^2=0$. Although the R^2 criterion is a dimensionless measure it depends heavily on the variance in the recorded series. Thus comparison of R^2 values for different catchments or even for different periods in the same catchment makes no sense.

Among the other numerical criteria often used are the following:

$$F=\sum_{i=1}^{n}|(QOBS_i-QSIM_i)| \qquad (4)$$

which is a measure of the accumulated deviation (absolute) between recorded and simulated values;

$$F^2_{\log}=\sum_{i=1}^{n}(\log\ QOBS_i-\log\ QSIM_i)^2 \qquad (5)$$

which does not focus as much peak matching as done the F^2 criterion; and

$$R = \frac{\sum_{i=1}^{n} (QOBS_i - \overline{QOBS})\ (QSIM_i - \overline{QSIM})}{\sum_{i=1}^{n} (QOBS_i - \overline{QOBS})^2 \ \sum_{i=1}^{n} (QSIM_i - \overline{QSIM})^2} \qquad (6)$$

which is the linear correlation coefficient between the simulated and the recorded series:

It is perfectly feasible to calibrate a model by optimizing just one of the available criteria. However, a calibration based on 'blind' optimization of a single numerical criterion risks producing physically unrealistic parameter values which, if applied to a different time period, will give poor simulation results. In the same vein it should be remembered that the criteria measure only the correctness of the estimates of the hydrological variables: generated by the model and not the hydrological soundness of the model relative to the processes being simulated. It is therefore recommended that, in a calibration, numerical criteria be used for guidance only.

In general it is recommended that a combination of the following four conditions be considered in determining goodness of fit:

1. A good agreement between the average simulated and recorded flows, i.e. and good water balance.
2. A good agreement for the peak flows, with respect to a volume, rate and timing.
3. A good agreement for low peaks.
4. A good overall agreement for hydrograph shape with emphasis on a physical correct model simulation.

These four conditions can be optimized numerically or subjectively through interactive computer graphics. In cases where all four criteria cannot be optimized simultaneously the priority depends on the objective of the project in question.

Finally, although the use of numerical criteria has been emphasized above, the value of graphical comparison of simulated and observed hydrograph should not be overlooked. Although analyzed more subjectively, a graphical plot provides a good overall impression of the model capabilities, is easily assimilated and may impact more practical information than does a statistical function. Graphical comparison should always be included in any examination of the goodness of fit of a simulated hydrograph.

MODEL CALIBRATION

Model calibration in general involves manipulation of a specific model to reproduce the response of the catchment under study within some range of accuracy. In a calibration procedure an estimation is made of the parameters, which cannot be assessed directly from field data. All empirical (black box) models and all lumped, conceptual models contain parameters whose values have to be estimated through calibration. The fully distributed physically-based models contain only parameters which can be assessed from field data, so that in theory a calibration should not be necessary if sufficient data are available. However, for all practical purposes the distributed, physically-based models also require some kind of calibration, although the allowed parameter variations are restricted to relatively narrow intervals compared with those for the empirical parameters in empirical or lumped, conceptual models.

CALIBRATION METHODS

In principle three different calibration method can be applied:

a. 'Trial and Error', manual parameter assessment
b. Automatic, numerical parameter optimization
c. A combination of (a) and (b)

The trial and error method implies a manual parameter assessment through a number of simulation runs. This method is by far the most widely used and is the most recommended methods, especially for the more complicated models. A good graphical representation of the simulation results is a prerequisite for the trial and error method. An experienced hydrologist can usually achieve a calibration using visual hydrograph inspection within 5-15 simulation runs.

Automatic parameter optimization involves a numerical algorithm which optimizes or minimizes a given numerical criterion. The objective of automatic parameter optimization is to search through the many combinations and permutation of parameter levels to achieve the set which is the optimum or 'best' in terms of satisfying the criterion of accuracy. Several optimization techniques have been used for calibration of hydrological models. A decade ago the most popular was Rosenbrocks's method (Rosenbrock, 1960).

The advantages of automatic parameter optimization over the trial and error method are:

* Automatic optimization is quick, because almost all work is carried out by computer.

* Automatic optimization is less subjective than the trial and error method, which to a large degree depends on visual hydrograph inspection and the personal judgment of the hydrologist.

Disadvantages of automatic parameter optimization include:-

* The criterion to be optimized has to be a single numerical criterion based on a single variable; as discussed in earlier section, though, selection of an appropriate criterion under these constraints is a complicated task.

* If the model contains more than a very few parameters the optimization will probably result in a local optimum instead of the global one.

* The theories behind the search algorithms assume that the model parameters are mutually independent. This assumption is usually not satisfied in practice.

* An automatic routine cannot distinguish between the different error sources mentioned earlier. Therefore, an automatic optimization algorithm will try to compensate, e.g. for data errors by parameter adjustments, with the results that the parameters values often become physically unrealistic and give poor simulation results when applied to a period different from the calibration period.

Combination of the trial and error and automatic parameter optimization method could involve, for example, initial adjustment of parameter values by trial and error to delineate rough orders of magnitude, followed by fine adjustment using automatic optimization within the delineated range of physically realistic values. The reverse procedure is also possible, first carrying out sensitivity tests by automatic optimization to identify the important parameters and then calibrating them by trial and error. The combined method can be very useful but does not yet appear to have been widely used in practice.

Finally, given the large number of parameters in a physically based distributed model like the SHE, it is not realistic to obtain an accurate calibration by gradually varying all the parameters single or in combination. A more sensible approach is to attempt a coarser simulation using oniy the few parameters to which the simulation from sensitivity analysis. However, experience suggests that the soil parameters will usually require the most attention because of their role in determining the amount of precipitation which infiltrates and hence the amount which forms overland flow.

The above methods of calibration consider single objective function. In case multi objective function is required to be considered, then two types of approaches, viz. classical approach and Pareto approach may be utilised. In classical approach a combined objective function is desired assigning the weights to the various objective function depending upon the user requirement. In pareto approach a set of parameter values are determined using search algorithm in such a way that the global optima is achieved considering the multi objective function.

MODEL VALIDATION

If the model contains a large number of parameters it is nearly always possible to produce a combination of parameter values which permits a good agreement between measured and simulated output data for a sort calibration period. However, this does not guarantee an adequate model structure or optimal parameter values. The calibration may have been achieved purely by numerical curve fitting without considering whether the parameter values so obtained are physically reasonable. Further, it might be possible to achieve multiple calibrations or apparently equally satisfactory calibrations based on different combinations of parameter values. In order to find out whether a calibration is satisfactory, or which of several calibrations is the most correct, the calibration should therefore be tested (validated) against data different from those used for the calibration (e.g. Stephenson and Freeze, 1974). Klemes (1986) states that a simulation model should be tested to show how well it can perform the kind of task for which it is intended. Performance characteristics derived from the calibration data set are insufficient as evidence of satisfactory model operation. Thus the validation data must not be the same as those used for calibration but must represent a situation similar to that to which the model is to be applied operationally.

Klemes (1986) further noted that a central question is: what are the grounds for credibility of a given hydrological simulation model? Usually they concern the goodness of fit of the model output to the historical record in a calibration period, combined with an assumption that the conditions under which the model will be used will be similar to those of the calibration period. Clearly, though, this is insufficient for a physically-based distributed model which is designed specially to simulate conditions different from those likely to be available for calibration, e.g. when simulating the impact of a future land-use change. In that case a demonstration of model transposability is required. Initially transposability referred to geographical transposability within one hydrologically homogeneous region. However, its scope has since been broaden to include transposability from one land use type to another, from one region to another and, recently, from one climate to another.

SCHEMES FOR SYSTEMATIC VALIDATION OF SIMULATION MODELS

The heirarchical scheme proposed by Klemes (1986), should be referred to here. The scheme is briefly discussed below:

The scheme is called hierarchical because the modelling tasks are ordered according to their increasing complexity, and the demand of the test increase in the same direction. Two major categories are proposed for the process to be simulated, in particular:

1. Stationary conditions, and
2. Non stationary conditions

Each of them being sub-divided into two hierarchical sub-groups according to whether the simulation is to be done for:

a. the same station (basin) which was used for calibration or
b. a different station (basin)

Here, the term stationary is used to denote physical conditions that do not appreciably change with time:

Typical examples of modelling tasks in these four classes of increasing difficulty are as follow:

1a. Filling in a missing segment of, or extending a stream flow record
1b. Simulation of a stream flow record in an ungauged basin
2a. Simulation of streamflow record in a gauged basin for conditions after a change in land use, climate or both
2b. Simulation of a streamflow record in an ungauged basin for conditions after a change in land use, climate or both

The following tests are recommended as a minimum standard for operational testing of models for the above four levels of difficulty of the simulation task:

1a. Split sample test
1b. Proxy basin test
2a. Differential split sample test
2b. Proxy basin differential split sample test.

(1a) *Split-sample test* : The available record should be split into two segments one of which should be used for calibration and the other for validation. If the available record is sufficiently long so that one half of it may suffice for adequate calibration, it should be split into two equal parts, each of them should be used in turn for calibration and validation, and results from both arrangements compared. The model should be judged acceptable only if the two results are similar and the errors in both validations runs acceptable. If the available record is not long enough for a 50/50 splitting, it should be split in such a way that the calibration segment is long enough for a meaningful calibration, the remainder serving for validation. In such a case, the splitting should be done in two different ways, e.g. (a) the first 70% of the record for calibration and the last 30% for validation; (b) the last 70% for calibration and the first 30% for validation. The model should qualify only if validation results from both cases are acceptable and similar. If the available record cannot be meaningfully split, then only a model which has passed a higher level test should be used.

(1b) *Proxy-basin test* : This test should be required as a basic test for geographical transposability of a model, i.e. transposability within a region. If streamflow in an ungauged basin C is to be simulated, two gauged basins A and B within the region should be selected. The model should be calibrated on basin A and validated on basin B and vice versa. Only if the two validation results are acceptable and similar can the model command a basic level of credibility with regard to its ability to simulate the streamflow in basin C adequately.

This kind of test should also be required when an available streamflow record in basin C is to be extended and is not adequate for a split-sample test as described above. In other words, the inadequate record in basin C would not be used for model development and the extension would be treated as simulation in an ungauged basin (the record in C would be used only for additional validation, i.e. for comparison with a record simulated on the basis of calibrations in A and B).

Consider geographical transposability between regions I and II. If streamflow needs to be simulated in an as yet unspecified ungauged basin C (or on a number of such basins) in region II the procedure should be as follows. First, the model is calibrated on the historic record of a gauged basin D in region I. Streamflow measurements are started on at least two different substitute basins, A and B, in region II and maintained for at least three years. Then the model is validated on these three-year records of both A and B and judged adequate for simulation in a basin C if errors in both validation runs, A and B, are acceptable and not significantly different. After longer records in A and B become available, these two basins can be used for model development and subjected to the simpler test for transposability within a region as described above, using A and B as proxy basins for C. Of course, the substitute basins A and B, would not be chosen randomly but would be selected so as to be representative of the conditions in region II, and, as far as possible, with due consideration of future stream gauging needs.

(2a) *Differential split-sample test* : This test should be required whenever a model is to be used to simulate flows in a given gauged basin under conditions different from those corresponding to the available flow record. The test may have several variants depending on the specific nature of the change for which the flow is to be simulated.

For a simulation of the effect of a change in climate, the test should have the following form. Two periods with different values of the climate parameters of interest should be identified in the historic record, e.g. one with high average precipitation, the other with low. If the model is intended to simulate streamflow for a wet climate scenario then it should be calibrated on a dry segment of the historic record and validated on a wet segment. If it is intended to simulate flows for a dry climate scenario, the opposite should be done. In general, the model should demonstrate its ability to perform under the transition required: from drier to wetter conditions or the opposite.

If segments with significantly different climatic parameters cannot be identified in the given record, the model should be tested in a substitute basin in which the differential split-sample test can be done. This will always be the case when the effect of a change in land use, rather than in climate, is to be simulated. The requirement should be as follows: to find a gauged basin where a similar land-use change has taken place during the period covered by the historic record, to calibrate the model on a segment corresponding to the original land use and validate it on the segment corresponding to the changed land use.

Where the use of substitute basins is required for the testing, two substitute basins should be used, the model fitted to both and the results for the two validation runs compared. Only if the results are similar can the model be judged adequate. Note that in this case (two substitute basins) the differential split-sample test is done on each basin independently which is different from the proxy-basin test where a model is calibrated on one basin and validated on the other.

A differential split-sample test can arise by default from a simple split-sample test if the only meaningful way of splitting an available record is such that the two segments exhibit markedly different conditions.

(2b) *Proxy-basin differential split-sample test* : This test should be applied in cases where the model is supposed to be both geographically and climatically (or land-use-wise) transposable. Such universal transposability is the ultimate goal of hydrological modelling, a goal which may not be attained in decades to come. However, models with this capability are in high demand and hydrologists are being encouraged to develop them despite the fact that thus far even the much easier problem of simple geographical transposability within a region has not been satisfactorily solved.

The test to demonstrate such a general transposability may have different forms depending on the specific modelling task involved. In the simplest case of geographical and climatic transposability within a region (e.g. for a model intended for assessment of impact of climatic change in an ungauged basin C), the test should have the following form. Two gauged basins, A and B, with characteristics similar to those of basin C are selected and segments with different climatic parameters, e.g. w for wet and d for dry, are identified in the historic records of both of them. Then, for an assessment of the impact of a dry climate scenario, the model is first calibrated on Aw and validated on Bd, and then calibrated on Bw and validated on Ad. It is judged adequate if errors in both validation runs Ad and Bd are acceptable and not significantly different. By analogy, a model intended for an assessment of the impact of a wet climate scenario would have to be calibrated/validated on Ad/Bw, and on Bd/Aw, and judged adequate if results from Bw and Aw are adequate and similar.

SENSITIVITY ANALYSES

Analysis of the sensitivity of the simulation results to changes in parameter values and analysis of parameter stability can served as model tests. Such analyses can be carried out in different ways. The influence of the length of the calibration period on parameter uncertainty as well as parameter stability with time can also be evaluated from such analysis.

EXTRAPOLATION FROM CALIBRATION CONDITIONS

If the calibration is based on a narrow range of data, the model, even of physically-based, may not be applicable outside this range. For example, if the data based contains only small floods, the model, even if properly validated in the operational sense, cannot be trusted to simulate very large floods adequately. The calibration/validation exercise should therefore be based on as wide range of conditions as possible. This approach can also be useful in eliminating incorrect calibrations in cases where it has been possible to achieve multiple calibrations based on different combinations of realistic parameter values. The incorrect calibrations are less likely to support acceptable simulations based on data outside the range used for calibration.

REMARKS

The model calibration and validation are the important aspects of the hydrological modelling proper calibration and validation of the hydrological model is necessary before using the model for simulation. For the validation of the models, the hierarchical scheme discussed in the lecture may be adopted. In order to ascertain the uncertainity in the parameters as well as parameter stability the sensitivity analysis must be carried out.

BIBLIOGRAPHY

1. Abbott, M.B., J.C. Bathurst, J.A. Cunge, P.E. O'Connell and J. Rasmussen, 1986. *An Introduction to the European Hydrological System - System Hydrologique European "SHE". History and Philosophy of Physically based distributed modelling system.* Jour. of Hydrology, 87: pp. 45-59.

2. Aitken, A.P., 1973. *Assessing Systematic Errors in Rainfall Runoff Model.* Journal of Hydrology, 20, 121-136.

3. Bathurst, J.C., 1986. *Physically-based distributed modelling of an upland catchment using the Systeme Hydrologique European.* J. Hydrol., 87(1/2), 79-102.

4. Bergston, S., 1976. *Development and application of a conceptual runoff model for Scandinavian Catchments.* Department of Water Resources Engineering, University of Lund. Bulletin Series A No. 52.

5. Betson, R.P., R.L. Tucker and F.M. Haller, 1969. *Using Analytical Method to Develop a Surface Runoff Model.* Water Resources Research, Vol. 5, No. 1, pp. 103-111.

6. Carrigan, P.H., Jr., 1973. *Calibration of U.S. Geological Survey Rainfall-Runoff Model for Peak Flow Synthesis - Natural Basins.* Report No. U.S.G.S.-W.R.D.-73-026, U.S. Geological Survey-WRD, Reston, Va, U.S.A.

7. Crawford, N.H. and Linsley, 1966. *Digital Simulation in Hydrology.* The Stanford Watershed Simulation Model 'IV'. Technical Report No. 39, Department of Civil Engineering, Stanford Univ., Stanford, California.

8. DHI, 1988. *Lecture notes and Exercises - Training course on Hydrological Computerized Modelling System (SHE).*

9. Fleming, G., 1975. *Computer Simulation Techniques in Hydrology.* Elsevier.

10. James, L.D., 1972. *Hydrologic Modelling, Parameter Estimation and Watershed Characteristics.* Journal of Hydrology, Vol. 17, pp. 283-307.

11. Klemes, V., 1986. *Operational Testing of Hydrological Simulation Models.* Hydrol. Sci. J., 31(1), 13-24.

12. NIH, 1982-83. *Modelling of Daily Runoff for Kasurnala Basin Using Betson and USGS Models.* Report No. CS-2, National Institute of Hydrology, Roorkee.

13. Rosenbrock, K.H., 1960. *An Automatic Method for finding the greatest or least value of a function.* The Computer Journal, Vol. 7(3).

14. Woolhiser, D.A., 1973. *Hydrologic and Watershed Modelling - State of the Art.* Transactions of the ASAE, 16, pp. 533-559.

** *** **