SELECTION, CALIBRATION, AND TESTING OF HYDROLOGIC MODELS

chapter 11

11

SELECTION, CALIBRATION, AND TESTING OF HYDROLOGIC MODELS

by L. Douglas James, Director, Utah Water Research Laboratory, Utah State University, Logan, UT 84322, and Stephen J. Burges, Associate Professor, Department of Civil Engineering, University of Washington, Seattle, WA 98195.

INTRODUCTION

Hydrologic modeling is an important tool for estimating and organizing quantitative information for water resources planning, design, and operation. The first two chapters of this monograph provide background on how to model by presenting the principles of deterministic and stochastic system synthesis. The next five chapters apply these principles to modeling precipitation, infiltration, runoff, evapotranspiration, and subsurface flow. Chapter 10 describes various ways these process models can be combined into a general hydrologic model, and chapter 13 inventories currently available models.

Planners or designers who require hydrologic information can choose from a variety of modeling alternatives. They can follow the principles outlined in the preceding chapters to develop a new model to meet their needs, or they can select from among available models. The purpose of this chapter is to provide guidance on how to select a model for a given application, how to calibrate the model selected to get the best information, and how to test the adequacy of the results. The guidelines begin with qualitative aids for screening to determine which models are conceptually sound for a given use and follow with quantitative measures for completing the selection, calibration, and testing.

THE BASIC PRINCIPLE OF MODEL SELECTION

The goal in hydrologic modeling is to estimate rates of water movement overland, underground, or within streams; amounts of water stored in the soil or in natural water bodies; or how these rates and amounts vary with time. Engineers and water planners see the careful representation of hydrologic processes necessary for hydrologic modeling as promising better

Acknowledgment: The authors appreciate the help of Courtney Bates, Boeing Computer Services, Kent, Washington, in preparing this paper. He generously agreed to having his work on time series error analysis directly included in the text. The authors are also grateful for the review comments received on earlier versions from David Bowles, David Dawdy, Willard Snyder, C. T. Haan, and an anonymous reviewer.

estimates of hydrologic quantities for management decisions. Those who attempt to model, however, soon find that the physical laws that govern these water movements and storages are so complex and that the parameters that must be measured for their representation are so variable in space and time that construction of a reliable model is no easy matter. Consequently, considerable research must go into developing understanding of physical laws and acquiring data on physical situations. At the present time, hydrologic science still falls far short of a universal model that can be generally recommended for all user applications, and a user must consider alternatives in deciding how best to model for a given application.

As users make these applications, the science of hydrology is also advanced. One of the most important contributions of model building toward advancing hydrologic science is through the discipline of data collection, system representation, and expression of results required for quantitative representation of complex hydrologic processes. When model results are compared to measured data, careless work quickly becomes evident as results prove to be scientifically invalid and unreliable for practical application. Through careful modeling, the scientist seeks better understanding of the causes of observed physical events, and the planner seeks more reliable estimates of design flows and of how those flows are affected by planning alternatives. The scientist wants to explain observed patterns of infiltration, overland flow, and channel flow; the planner wants to design reservoirs, delineate flood plains, and size culverts.

The fundamental criterion for the planner in selecting, calibrating, and testing hydrologic models must be how well the model performs as a sound basis for planning and management decisions. The fundamental criterion for the scientist must be how well the model contributes to a hydrologic understanding that will eventually help planners make even better decisions. By applying scientific contributions, those making planning and management decisions are helped as they can be surer of the hydrologic consequences of the alternatives they consider and can have greater confidence in the hydrologic estimates for the technical designs they recommend. The primary decision-makers, hence the primary users of hydrologic estimates, have been professional planners and design engineers. Recently, shifts toward public participation in selecting government projects, and toward individual implementation of many small-scale (flood proofing for example) and nonstructural measures, are increasing the need for hydrologic models that can produce information for specific locations (flood hazard on land parcels for example) whenever someone wants it, yet be easily understood by the less technically sophisticated public (James, 1973).

A hydrologic model represents runoff processes in a manner that can be used to estimate (a) how the prototype physical system would respond to sequences of external stimuli (usually precipitation), (b) the frequency distribution of responses of interest (high or low flows), or (c) how changes to the system would alter the response. The representation may be a physical model providing scaled-down duplication of the prototype system, an analog model using another physical system that is more economically constructed but has similar response characteristics, or a mathematical model that represents the physical processes by a set of equations programmed for a digital computer. Physical and analog models have definite advantages where governing physical laws cannot be expressed by succinct and readily-

solvable equations (or the precision desired exceeds that of the equations; e.g., hydraulic modeling of spillways) or where the analogy is inexpensive and accurate. Digital computer modeling, however, now dominates hydrology because of: (a) its relatively low cost while skilled labor required by other model types is scarce and expensive, (b) its ease of transferability from one computer or from one prototype configuration to another, (c) greater flexibility for modelers who would like to make changes. This chapter therefore is restricted to digital computer models.

A FRAMEWORK FOR COMPARING DIGITAL MODELS

Application of a digital computer model to meet a given planning need requires an inventory and evaluation of available models, adopting or adapting (few can afford to start from scratch) the one judged most suitable in terms of reliability and cost (both factors depend highly on user familiarity), and calibrating the selected model to estimate the hydrologic information of interest. Chapter 13 provides an inventory of currently available models and general information for preliminary screening. To establish screening criteria, it is helpful to review the alternatives in model development because the choices a modeler makes on these alternatives govern the suitability of his product for any given use.

Hydrologic modeling requires (a) identification of the hydrologic quantities important to the user and therefore to be estimated by the model, (b) identification of hydrologic processes within the prototype system that need to be modeled to estimate the desired quantities adequately, (c) selection of equations to represent each identified component physical hydrologic process (chapters 3 to 7), (d) structuring the equations to differentiate coefficients treated as constants in all model applications from parameters which are varied in calibrating the model for specific watersheds, (e) synthesizing the selected process equations into a computational framework (the mathematical model) adequately reproducing the desired hydrologic responses for the system as a whole (chapter 10), (f) selecting values for the parameters that provide the best estimates of hydrologic response for a specific watershed, (g) testing the adequacy of the model estimates in the desired applications, and (h) communicating the results to decision-makers.

The first five subtasks produce the model; the sixth calibrates it; the seventh tests it; and the eighth, the often overlooked aspect of depicting technical information so that the user can understand and will believe it, is critical to making the modeling effort worthwhile.

A person, selecting from among a set of models programmed by others, needs to review the decisions made by the modelers during the first five subtasks. **Selection** should be based on the user's determination that (a) the model provides the kinds of information needed, (b) the watershed characteristics represented by the model parameters indeed govern watershed response in the intended application, (c) the equations used are correct in light of the state-of-the-art, available data, and available computer facilities, and (d) the model will provide results, which are suitable for the intended use, of acceptable quality at a reasonable cost within a required time frame. **Calibration** should seek an economical and reproducible method for estimating parameters. **Testing** should judge the adequacy of the model for decision-making.

The decisions made by the modeler that a potential user should consider in model selection, calibration, and testing may be classified into the areas of (a) modeling strategy, (b) model conceptualization, (c) model structure, (d) model calibration, (e) model testing and verification, (f) model development dynamics, and (g) display and communication of model results. These seven issue areas are explored qualitatively in the next seven sections of this chapter.

ISSUES IN MODELING STRATEGY

The issues in modeling strategy relate to basic goals of model construction and to the best type of model for achieving those goals. The four broad strategy alternatives are:

1 Simulation of water movement through the component physical processes and then integrating the results through water accounting within a general hydrologic model. The Stanford Watershed Model (Crawford and Linsley, 1966), as an example, provides continuous information on water storages and movements over time. Some of the hydrologic event models (chapter 10) provide this kind of accounting during storm periods.

2 Simulation of the total flow (or soil moisture, evapotranspiration, or some other water movement or storage) for a selected hydrologic event or events (a given date and given location) through relationships derived by statistical techniques seeking maximum correlation between observed inputs (e.g., precipitation, antecedent moisture, and basin size) and output, (e.g., streamflow). Traditional procedures for estimating storm runoff from information on storm characteristics and antecedent moisture (Linsley et al., pp. 265-274), a variety of statistical approaches (Beard, 1962), and some of the "event" models (chapter 10) are examples.

3 Estimation of the statistical properties of the time series of events through relationships derived by statistical techniques seeking maximum correlation of the selected statistics (e.g., the mean or standard deviation of a recorded series of annual floods, computed 10- or 100-yr flood peaks, etc.) with watershed characteristics (drainage area, forest cover, slope, soil type, etc.). The procedures developed by Santa Clara County, CA, for estimating flood peaks on ungaged watersheds (Saah et al., 1976) and by Fletcher et al. (1976) for estimating flood flows for desired frequencies for highway culvert design are illustrations. Flow statistics estimated in this manner are used in stochastic methods (chapter 2) to estimate reservoir yield.

4 Optimization of system design or operation by programming the model to integrate hydrologic information with relevant economic, social, and environmental factors and produce comparisons of management alternatives. This modeling strategy is illustrated in models selecting the optimum combination of structural and nonstructural flood control measures (James, 1970) or facilitating other aspects of hydrologic design (Fleming, 1975, pp. 257-312).

The first strategy, simulation based on component physical processes, requires subdivision of the watershed into separable spatial units and subdivision of the period of analysis into discrete time units. Data must be collected to depict the precipitation and potential evaporation stimulating hydrologic activity in each spatial unit by time period and to depict the physical characteristics controlling the response of each spatial unit. Such modeling is limited by the impossibility of obtaining complete information on the small space-time grid required for comprehensive physical representation and by the computer time required for repeated solution (once

for each space-time unit) of the differential equations that provide accurate process representations. Consequently, operational models use areas of a size matching rain gage density rather than small homogeneous areas and use time aggregations of total water movement over hours or days rather than continuous instantaneous flow rates. The greater grid coarseness in time and space, however, causes the parameters in the equations to be less correlated with directly measurable physical watershed characteristics. Values for the parameters, therefore, have to be estimated by statistical methods based on criteria of best match of model results to measured data. A model calibrated in this way can only be used for conditions covered by the data used to calibrate it; it cannot be used to extrapolate beyond the range of the calibration data.

The second strategy provides the user a simpler model by sacrificing replication of water movement through physical processes. The statistical techniques used to estimate output (streamflow) from correlations with input (precipitation) and watershed characteristics are tested on the basis of goodness of match of the end results without any provision for intermediate process checks. A typical example is to use streamflow measured at scattered gages to develop a regional relationship between flows and watershed characteristics to estimate flows at ungaged sites. Greater separation of these methods from physical processes suggests greater difficulty in obtaining reliable estimates for watersheds not well represented in the data base and greater scatter between predicted and measured results.

Examples of the third strategy (a regional relationship to estimate a design parameter such as a 100-yr flood instead of a flow during a dated historical event) are found in estimating reservoir yield (Haan and Allen, 1972) and in regional flood studies. The Santa Clara County flood study used records from 23 gaging stations in and near the county to derive the relationship between the 100-yr flood peak and watershed characteristics:

$$Q_{100} = 19.4\, A^{0.95}\, P^{0.58}\, S^{-0.25} \quad \ldots\ldots\ldots\ldots [11.1]$$

where Q_{100} is the 100-yr peak in cfs (0.0283 m^3/s), A is the drainage area in square miles (2.59 km^2), P is the mean annual precipitation in inches (0.0254 m), and S is the slope of the main stream (Saah et al., 1976, p. 36). The equation has a coefficient of determination (R^2) of 0.904 and a standard error of estimate for Q_{100} of 0.218. The information on the design parameter or dependent variable used to derive such a relationship is ideally estimated from long sequences of measured flows but can be approximated from data series constructed through one of the first two strategies.

While equation [11.1] is very helpful to Santa Clara County officials needing to estimate flood peaks at ungaged locations, it also provides a good example to illustrate cautions required in using this sort of model. Obviously equation [11.1] should not be applied outside the county, except perhaps at locations just across its borders. In addition, a location map showing the 23 gaged watershed shows virtually all the gages to be in the western half of the county, and the eastern portion is significantly different in climate, geology, and topography. Even within the western half, many ungaged watersheds fall outside the gaged range of the dependent variables or are influenced by other variables which do not appear in the equation because

they did not happen to vary much among the watersheds that were used. All of these issues illustrate dangers in extrapolating outside the range of available data, and all of them should be carefully considered.

The fourth strategy integrates hydrologic modeling with various information on the physical, economic, and environmental consequences of water management alternatives to help decision-makers understand (a) the consequences of physical changes, occurring within the watershed, on runoff (Lumb and James, 1976) or (b) the benefits and costs of structural changes under construction. The integration expands the scope of modeling to portray implications of hydrologic change rather than have the user look to other methods for applying the output of the hydrologic model.

The first strategy has its primary application when one wants a time series of hydrologic responses or information on how hydrologic responses will be altered by watershed changes. It is enhanced by quantitative information on physical watershed characteristics and becomes progressively more attractive as estimates of more types of hydrologic response (floods, low flows, runoff volumes, soil moisture, etc.) are desired. The second strategy has its primary application in estimating a single type of hydrologic response for specific events (e.g., flood peaks on a given date) and becomes progressively more attractive as estimates are required for a number of storms on the same watershed or for the same storm on a number of adjacent watersheds. The third strategy is more economical than the first two for producing design estimates but achieves the economy at a greater risk of error because of the further separation of the approach from physical causes. The fourth strategy extends modeling by interfacing with non-hydrologic factors but does so at the risk of becoming too broad (interdisciplinary) to be easily understood and of combining data of greatly varying reliability and precision without explicit recognition of the uncertainties resulting. Progression from the first to the fourth strategy moves the modeling effort from scientific hydrology through statistical analysis to multiple objective planning; a potential user needs to select a model that matches where he needs to be on this spectrum.

ISSUES IN MODEL CONCEPTUALIZATION

The basic issue of model conceptualization is the degree of detail to use in making model results reflect physical watershed characteristics within the framework of the selected modeling strategy. Specific examples, strategy by strategy, are (a) which physical processes to model, the detail and level of sophistication to use for those selected, and the accounting detail to use in routing water through the model, (b) which outputs to have the model estimate (for flood hydrology the scale may vary from peak flows to complete flood hydrographs), the storm characteristics and watershed conditions to use as independent variables, and the statistical techniques to use, (c) which distribution to use to represent a hydrologic time series and which parameters of that distribution to estimate, and (d) which economic, environmental, and social considerations to combine with the hydrology to guide decision-making.

The scientific bias in modeling strategy is to assume that any effort that will make a model theoretically more correct or empirically more reliable is desirable. The planning reaction is to ask whether the extra effort is justified. The practical justification is in terms of contribution to better

water resources planning and management decisions. Specifically, will the value received from the additional information justify the additional cost of a more comprehensive model? Factors relevant to this determination include the contribution the additional or more reliable results will make to better planning decisions or more economical engineering design, the strength of the theory and the reliability of the data required for the additional computations, the cost of assembling any additional data in the required form, and the capacity of available computer facilities to handle additional computations at a reasonable cost. A user selecting a model should carefully review the data and funds he has available and the objectives that he is trying to achieve to avoid either a greater level of sophistication than can be justified by his need or an approach that does not produce key inputs to his decision-making.

A number of examples might be used to illustrate the care that should go into selecting an appropriate level of model sophistication; one is the selection of a method of streamflow routing. Methods available range from using time-area historgrams, adding linear reservoirs, applying Muskingum routing parameters derived from recorded hydrographs, to computing from kinematic or full dynamic routing equations. Hydraulic theory supports dynamic routing as the most accurate, but the additional accuracy is only gained at the cost of collecting additional data and executing a more complex model. Justification of these additional costs depends on the model application. As examples, justification is highly likely for simulating flood hydrographs for urban drainage design but very unlikely for estimating reservoir yield for water supply design; vastly different time increments are involved.

The tendency has been to add theoretical advances to hydrologic models without first exploring the advantages and disadvantages. One result has been models that have an order of magnitude of greater precision in representing some processes than in representing others. This lack of balance can only be justified if the results desired from the model are sensitive to the more precisely represented hydrologic process by a comparable order of magnitude. The sensitivity of the results should be explored before adding theoretically justified but computationally more expensive model elements.

Coleman and DeCoursey (1976) built on previous work by McCuen (1973) to provide a methodology for employing model sensitivity analysis to estimate the reduction in variance achieved by model modification; the achieved reduction in variance provides a basis for determining whether the modification is justified. The method defines sensitivity (S) as the derivative of model results (R) with respect to a parameter (P) of interest. Therefore,

$$S = \partial R/\partial P \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [11.2]$$

Relative sensitivity (S_r) is defined as

$$S_r = \frac{\partial R}{\partial P}\frac{P}{R} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [11.3]$$

A number of numerical measures of model performance are presented later in this chapter. P may be taken as a defined model parameter or as a discrete change to the model (shifting from Muskingum to kinematic routing).

These sensitivity estimates can be applied at various levels of sophistication. Relative sensitivity provides a basis for comparing various parameters and concentrating research and data collection on the more sensitive. Such comparisons contribute a great deal toward logical priorities in making model changes and working toward a product that does not consume a great deal of computer time in refined calculations whose contribution is negligible because the model combines the results with very rough estimates from other computations. Equation [2] can be combined with the rules of economic optimization (James and Lee, 1971) to make explicit modeling choices. For this purpose, the economic sensitivity (S_e) can be defined by

$$S_e = \frac{\partial R \; M_b}{\partial P \; M_c} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [11.4]$$

where M_b is the marginal benefit of better results to the model user and M_c is the marginal cost of parameter refinement. A value of S_e exceeding unity indicates a profitable change. A user may find a model change that he can justify solely from his intended application ($S_e > 1$ with only M_b to himself counted). Modelers often would consider marginal benefits to all users to justify additional work. In a completely balanced model, one would achieve equal marginal rates of substitution among parameters and equal marginal rates of transformation among results (James and Lee, 1971, pp. 74-82). Such refinement, however, should be taken as more of a conceptual ideal than a working guide.

Other considerations in tying model results to physical watershed characteristics are the extent to which those characteristics can be expected to change over time and the control the user has over those changes. If watershed changes result from influences outside the user's control, the model can be used in advance to estimate flow sensitivity to these changes. A flood control planning agency, concerned that upstream ubranization and channelization will increase design flows for downstream structures, needs a model that will relate upstream land use and channel changes to downstream flow for flood control plan formulation. If the user regulates certain watershed characteristics (e.g., channel dimensions or detention storage capacity), the more directly those characteristics are used as input data to the model, the easier it will be for him to use the model for planning. A user who thinks in terms of specific watershed characteristics should select a model whose parameters relate to those characteristics.

ISSUES IN MODEL STRUCTURE

The basic issue of model structure is which method to use in putting the conceptualized model components together. Specific issues under this general heading are (a) how the level of aggregation (the size of the spatial units and the length of the time intervals) used in the model match the spatial and time detail at which information is required by the user, (b)

how well the equations and computational framework contained within the model represent the state-of-the-art for estimating the hydrologic responses of interest to the user, (c) the robustness of the model as reflected by its ability to make reliable estimates if the user faces limited data availability, and (d) the capability of the user to collect and manage data for model calibration and testing. Most hydrologic models are structured with particular users in mind. As additional users apply a model, the structure becomes generalized.

A person desiring to use a hydrologic model to obtain needed information may either start from scratch and build a specialized model or adopt a more general model. There is some chance of finding a general model that will work directly, but, more likely, it will have to be modified to handle the peculiarities of the new special situation. A new specialized model can often produce better estimates of a few selected hydrologic responses but requires much greater expertise and cost. The total process of developing a general model requires much greater investment in research and development but generates net savings if the investment is divided among more applications. A major component of a general model is its data management structure. Many special purpose models are poorly programmed, and little thought is given to the development of data files for easy manipulation within the model. Developers of more general models have generally been forced to recognize that as much as 80 percent of the effort involved in modeling a watershed is spent in data management; consequently, they have made considerable effort to simplify data management for the user. The quality and scope of the data management component of a model is an extremely important factor in model selection and use. Development of a few general models would therefore seem to be more in the interest of hydrologic science and practice than for each user to develop his own specialized one.

A user in the process of selecting the model with the best structure for a particular application normally has in mind a minimum acceptable level of accuracy. He also would like to improve on that minimum level to the degree better results justify the extra time and cost (equation [11.4]). For example, a user that needs to know flood peaks on a small watershed needs to identify models (e.g., from those in chapter 13) with a fine enough spatial grid to represent that watershed and a fine enough time grid not to miss the flow peak by a significant amount. A selection can then be made among qualifying models on the bases of theoretical reputability, computational economy, probable public credibility of the modeling approach, and the time required to become familiar with the model and establish the necessary data files.

Model builders generally major on theoretical reputability with a minor in computational economy. They are less likely to have structured their models to minimize user difficulties in making applications. One user difficulty modelers are very likely to overlook is that many water resources planners are required by law or by administrative regulation to incorporate certain considerations in their analysis and cannot use a model whose programming they do not understand well enough to know for sure that it meets those requirements. Other users encounter difficulty in running a computer program or in interpreting the results and will return to a more familiar method even though theoretically it will not meet their needs as well. A great deal more attention needs to be given to the technology trans-

fer of complicated hydrologic models; one survey could not find one user that was able to apply a complex model successfully without first taking a course on the model from someone who was previously trained (James et al., 1975, pp. 157-160).

ISSUES IN MODEL CALIBRATION

Once a model is selected, it must be calibrated for a watershed. There are two calibration criteria: model results should match recorded data, and the estimates of parameter values should be consistent with watershed characteristics. The second criterion should not be interpreted too literally. For example, conceptual watershed models often represent infiltrated water movement by means of several conceptual soil moisture storages. These storages cannot be tied to any specific watershed characteristic; they represent integrated, areally averaged conceptual approximations to subsurface moisture storage.

A model that produces only an estimate of the 100-yr flood peak is almost impossible to check without a long history of gaged records. A model that produces annual runoff volumes provides one value to check per year. A continuous streamflow model provides a continuous time series of flows for comparison. A model that replicates the total runoff process provides the possibility of checking soil water content, depression storage, and other observable watershed responses. A model whose parameters are related to physical watershed characteristics allows the user to compare the parameter estimates that do the best job of simulating hydrologic output with such measurable physical watershed characteristics as impervious area, soil permeability, and channel flow times. Such models provide greater opportunity to gain confidence in the results.

The calibration process requires a procedure to evaluate the success of a given calibration and another procedure to adjust the parameter estimates for the next calibration. The criterion of success may be a subjective judgment on adequacy, some statistic selected as measuring goodness of fit, or some multi-objective function combining several statistics. The adjustment procedure may be a subjective determination of what parameter changes seem likely to be desirable, a set of rules derived from sensitivity studies (equation [11.2] or [11.3]) on the various parameters, or a systematic variation of parameter values designed for efficient examination of a response surface (James and Lee, 1971, pp. 486-488). The statistical models are more likely to use a systematic search because their objective function is more easily defined whereas the many items for which synthesized results can be compared with measured values in the more general models make selection of criteria of best fit more difficult. At the present state-of-the-art, users wanting to estimate flood peaks, low flows, and soil water content may legitimately make three different calibrations of the same general model for the same watershed.

Statistics that might be examined in deciding if a given calibration is acceptable include (a) statistics computed from the complete set of simulated flows (e.g., minimum sum of the squares of daily flow differences), (b) statistics computed from only those flows of particular interest (e.g., minimum sum of the square of daily flow differences during low flow periods for a waste water dilution study) (c) statistics on other types of measurements such as soil water content or known physical characteristics of the watershed such as impervious area, (d) statistics on changes in flow from

one time period to the next, (e) magnitudes and patterns of systematic and random "errors," and (f) sensitivity data on the impacts of uncertainty in initial conditions, of errors in precipitation and evaporation data (Dawdy and Bergman, 1969), or of defective modeling equations. Any one of these statistics might be selected to guide the model user to an acceptable calibration, or several might be combined in a more complicated objective function. No matter which statistic or combination of statistics is chosen, the calibration goal is to minimize errors (differences between recorded and simulated values), and alternatives error measures are discussed in a later section.

Certain model parameters are more appropriately obtained by direct measurement before beginning the calibration than by statistical criteria of goodness of fit. Watershed drainage area, the impervious fraction of the watershed, and channel hydraulic capacity are specific examples. As models become more sophisticated and the relationships between model parameters and watershed characteristics become better known, the number of parameters that have to be estimated by minimizing error statistics become fewer, and model calibration will be able to make greater use of field data. For his selected model and application, the modeler should give careful thought in deciding which parameters to estimate from data on watershed characteristics and which to estimate by adjusting values to minimize errors. The parameters to estimate from field information are those which the model structure ties to physical watershed characteristics on which reliable descriptive data are available.

Presently, very few parameters can be identified directly from field measurements. Watershed area can be readily obtained for rural watersheds; however, delineating the watershed boundary on a topographical map can lead to substantial errors in urban watersheds if field checks are not made to adjust for human activities that have altered drainage patterns. If variable Muskingum-type coefficients are required, they can be obtained readily from stream-gage measurements. Unless one is modeling an extremely small area (about a square yard or a square meter per measurement) field infiltration data are virtually worthless for direct estimation of model parameters. The enormous spatial variability of infiltration, land slopes, etc., necessitate using some average values representative of the level of the process. Similarly, use of an average permeability for groundwater flow is necessary; the permeability coefficient has been known to range over three orders of magnitude in a distance as short as 200 ft (61 m) in materials that are geologically uniform. The modeler should be very careful in using physical field measurements for most parameter values.

Calibration can simultaneously deal with the full set of parameters or can separate out subsets that are relatively independent of one another and deal with each subset separately as is done in one self-calibrating version of the Stanford Watershed Model (Liou, 1970). The chief advantage of dealing with subsets separately is that some watershed characteristics have much greater effect on certain flows than on others. For example, recorded information on the shape of the flood hydrograph can be used to calibrate parameters related to channel hydraulics but is of little worth in calibrating parameters related to subsurface flow. The parameter subsets used by Liou are those dealing with:

1 Streamflow recession for which parameters are calibrated on flow sequences between storm periods.

2 Division of precipitation between surface runoff and infiltration and of infiltration between that contributing to evapotranspiration and that contributing to interflow or baseflow. Liou calibrates this parameter subset on annual runoff volume and the distribution of that volume among the months of the year.

3 Flood routing as calibrated by the shape and timing of the rising and crest portions of flood hydrographs.

4 Interflow as calibrated by flows in periods immediately after larger flood events.

After deciding which parameters to assign values directly from information on watershed characteristics, the user should group the remaining parameters into relatively independent subsets for joint calibration. The recession rates of Liou's first group can be estimated directly from recession data; two major parameter subsets for estimation by calibration remain: (a) parameters that determine runoff volume and (b) parameters that determine flood hydrograph shape. These two subsets should be calibrated separately.

One method for calibrating these subsets of parameters is by the systematic pattern search method presented by Hooke and Jeeves (1961), applied to hydrologic model calibration by Munro (1971), and refined by Lumb et al. (1975). Some error measurement (possible equations given later in this chapter) is selected for minimization as a criterion of goodness of fit. A set of initial (X_0) for the parameters to be calibrated is established by arbitrarily selecting plausible values for each, and a simulation is made. The results are associated with some error (E_0). The value of the parameter to which the error measurement is most sensitive is then increased by a relatively small amount Δ while holding all other parameter values constant. If the error increases with a positive Δ, a negative Δ is tried. After these two trials, the first parameter is returned to its X_0 value, and the process is repeated for each of the remaining parameters one at a time until one or two of these local excursions have been made for each parameter. The search then makes a pattern move from X_0 to a new set (X_1) by changing the value for each parameter in the direction found to reduce the error measurement. The amount of change for a parameter is increased as repeated local excursions indicate adjustment in the same direction. The search continues in this fashion from one set of parameters to the next (X_0, X_1, X_2 ... X_n) until the error measurement is no longer being reduced appreciably ($E_n \cong E_{n-1}$).

An alternative adjustment procedure is to use either the judgment developed through model building experience or a sensitivity analysis to determine which particular flows (or other simulated output) vary most with a change in a given parameter. Each parameter is then only adjusted as indicated by error measurements originating from related sensitive flows. Factors to consider in choosing between the two approaches include: (a) systematic search takes more computer time for many trials and is therefore impractical for a large number of parameters; (b) judgment search depends on adjustment rules that must be gained by experience, vary with local hydrologic conditions, are much more difficult to program, and are more sensitive to data errors; and (c) systematic search will lead to minimum error measurements while judgment search will lead imperfect models to parameter values that make better sense when compared with known watershed characteristics. Either method can be programmed into a model, but

repeated trial simulations are expensive, and one should try to keep the cost as low as possible by beginning from a good first approximation and by using a simulation algorithm that is as economical as possible.

One very important goal in calibration is that the results be reproducible and independent of the person performing the calibration. Subjective calibration works reasonably well for a one time application by an individual user but becomes quite unsatisfactory when separate calibrations are used to compare hydrologic responses among watersheds or time differences in response of a given watershed. Deterministic errors caused by judgment differences dominate physical differences.

ISSUES IN MODEL TESTING AND VERIFICATION

Once the calibration process has been used to estimate the best values for the model parameters, the outcome needs to be evaluated to determine if the results provide adequate information for answering the questions facing the decision-makers. Answers may be unsatisfactory either because the model provides faulty estimates of needed hydrologic quantities or because the selected outputs do not provide information the decision-makers consider relevant. The accuracy problem is more widely discussed by hydrologic model builders, but the relevancy problem may well generate greater potential-user skepticism. Hydrologic information is imperfect. The acceptability of results from a given calibration generally depend on whether the insufficiencies justify additional effort to upgrade the model, to improve the calibration, or to turn to an alternative method for estimating the desired information.

Faulty results may stem from a variety of causes with the appropriate remedial action (where remedial action is possible) depending on the problem. Common problems and possible remedial measures include:

1 Errors in the data used in calibration. Both the data used as input to the model and the data used to check model output should be checked. Application of correlation techniques (Haan, 1977, pp. 222-235) to identify cases where a relationship among data departs from its normal pattern is one approach, and such graphical techniques as the double mass curve (Linsley et al., 1975, p. 81) are often a practical substitute. Data with large errors should not be used for calibration, particularly with computerized search techniques that do not temper results with hydrologic judgment. The effects of smaller data errors can be balanced over time by using a longer calibration period. Seldom are records of concurrent precipitation, stream flow, evapotranspiration, etc., complete; therefore, some gaps in the records exist. If one is calibrating a continuous simuluation model, these gaps must be filled in. The user should not be too concerned if simulated and observed streamflow do not match very well during periods when filled-in data have been used; he should not use the filled-in portion of the record (fill-in techniques are very crude) to influence his assessment of the model calibration. As another warning, streamgage records for overbank flow are less accurate than within channel flow estimates. Flow rates for large floods are seldom measured but rather extrapolated from stage-discharge relationships. Difficulties in matching recorded high flows cannot necessarily be blamed on the model.

2 Use of a period of record that does not contain enough events with the physical processes needed to calibrate key parameters. For example, the hydraulic characteristics of overbank flow differ substantially from those

of main-channel flow; consequently, overbank flow parameters cannot be estimated from a period of record during which no such flows occurred. The calibration period should be reviewed to make sure that all important event types were recorded.

3 Inadequate or misrepresentation by the model of hydrologic processes found in the watershed (for example, a failure to represent the impervious characteristics of frozen ground in watersheds where significant storms occur after freezing periods). Model results should be compared visually with recorded data series to look for consistent variations. Possible physical causes of those detected should be explored, and the model should be corrected to deal with problems where the effort seems justified.

4 Insufficient output for decision-making. Any failure to develop desired information normally becomes obvious when model results are presented to users. Sometimes the additional information can be obtained by outputing quantities already computed internally, but often the modeling process itself must be revised.

The basic issue in model testing is to determine if the hydrologic estimates (residual error) achieved by the calibration are acceptable. The basic process is to run the calibrated model for a time period other than that used for calibration and evaluate the results. The modeler would normally test with the same error measurements used for calibration, but a new user may consider different sorts of error more important and select different statistics to evaluate a previous calibration.

The user is interested in the expected error, the distribution of possible errors, the consequences of using erroneous information, and the prospects for improving the estimates. The acceptability of a calibrated model ultimately depends on judgment as to whether or not the estimates can be improved (a) for a cost justified by reducing the consequences of using erroneous information and (b) before the critical decisions must be made. Hydrologic modelers can contribute to better water resources planning through frank discussion of these issues with decision-makers.

ISSUES IN MODEL DEVELOPMENT DYNAMICS

Many hydrologic modelers are more familiar with the issues in model selection, calibration, and testing than they are with the issues in communicating model results to users to provide a useful water resources planning tool. These issues include how best to (a) transfer an understanding of the technology of applying a complicated model to others, (b) organize libraries of available models and keep them updated for ready access and reference by users, (c) facilitate communication of user-discovered model deficiencies to researchers to take corrective action, (d) facilitate the exchange of experiences among model builders to minimize duplication of effort, and (e) decide when to use a previously developed model and when to develop a new one.

Because these issues have not been studied in anywhere near the depth given to the topics addressed in the previous sections, criteria for efficient model development dynamics have not been established. Basic principles of effective communication, however, suggest needs for (a) carefully documented user manuals and other descriptive material on each model so users can make informed decisions whether a particular model is applicable for a situation and make that application should the decision be favorable,

(b) libraries where users can go for manuals, source programs, and contacts with experts who can provide detailed help, and (c) effective feedback channels for exchanging modeling experiences. Development of specific guidelines for achieving effective programs in all three areas would be very worthwhile.

Model documentation should include a frank presentation of model strategy, conceptualization, and structure and specific recommendations to facilitate model calibration and testing. The reader needs to learn model strengths and weaknesses with respect to the application that he has in mind. He needs to be told of the theoretical and empirical constructs incorporated within the model so that he can decide whether he concurs, if they also hold for the use he has in mind; and if they are in harmony with the design policies of his agency. He needs a flow chart to understand how the program is organized, a listing for detailed analysis, and a dictionary defining the major variables so that he is in a position to make changes that he feels desirable. He needs guidance on the sensitivity of model output to input data so that he can concentrate his data collection efforts productively and on sensitivity to parameter values so that he can calibrate more effectively.

Two items of concern to all model users are learning and implementation time. Most model developers do a poor job of conveying this information. The lack of a concise manual which states what the model can and cannot do, its accuracy, data needs, user training time, implementation time, etc. is a strong recommendation against use of the model. Model builders who do not provide this information are only permitting their models to be used by a few skilled persons. User demand for a careful summary of model capabilities from model salespersons will improve the overall level of practice.

ISSUES IN DISPLAY AND COMMUNICATION OF MODEL RESULTS

Most hydrologic models have been designed to communicate to scientific hydrologists or to design engineers, and very little effort has been spent in developing the potential of modeling for displaying hydrologic information to land use planners, property managers, water users, and others not trained in water resources engineering. Such user groups require more explanatory labeling of model output, conveyance of that output in a manner that generates credibility through its intuitive reasonableness, greater ability to produce information at specific spots of special interest to particular individuals (translation of streamflow measurements to a graphic portrayal of what the flood will look like as it crosses a specified land parcel), and quick response to information needs whenever they arise. These needs contrast with the common operating scenario in which hydrologists communicate a single set of design information to a design group.

Hydrologic models have great potential for communicating hydrologic information to the public. One can imagine a videoscreen portraying a reach of a river and adjacent buildings in three dimensions. By interacting with a computer terminal, one might ask to see the 100-yr flood moving through the reach and be shown the dynamics of the rising and falling hydrograph and the damage left in its wake. He might then ask how this hydrograph (or any other historical storm hydrograph) would be affected by new upstream development, a storage reservoir, a culvert enlargement, or any other factor of interest. The potential for communicating to the public is tremendous and untapped. Complex models need not create communication barriers.

Nevertheless, the modeler needs to keep in mind the importance of conveying the uncertainties inherent in his work as well as the results. All models are imperfect. It is important to show and explain the uncertainty bounds on all outputs. Display techniques might range from plotting levels of confidence on both sides of prediction curves to showing the low, most probable, and high estimates on a videoscreen system.

ORGANIZING FOR MODEL CALIBRATION

General Principles

Once the user has selected a hydrologic model based on the principles outlined above, he needs to calibrate it. Space does not permit discussion of the specific intricacies of calibrating all the various types of models discussed above and catalogued in Chapter 13, but deterministic rainfall-runoff model calibration will be discussed in some detail and used to provide a foundation for some general principles. These models were chosen for illustration because they are probably the most difficult to calibrate and because the principles described can generally be extended readily to other model types. The intent is to describe a general approach to rainfall-runoff model calibration and avoid becoming involved in such specialized issues as modeling snowmelt, river freezing and thawing, groundwater recharge, interbasin tranfers, flash flood events, and desert storms. No matter what aids to calibrate are programmed into a model or recommended in user manuals, the user should follow a systematic calibration approach based on general principles.

Model calibration should begin by deciding what information will be important when the calibrated model is applied. The user needs to translate these decisions into criteria of goodness of fit, data to collect, and pitfalls to avoid. These issues are extremely well articulated by Clark (1973); his paper should be read along with this chapter before applying any model.

The calibration process should begin from a clear realization that calibration is data demanding. The less frequent a given hydrologic event, the more difficulty will be encountered in calibrating a model for that type of event. For example, extreme flood events are more difficult to replicate than are ordinary flows.

A Philosophy for Dealing with Error

No hydrologic modeler nor model user can afford to forget the errors inherent in mathematical representation. Rainfall-runoff models convert a precipitation time series into a series of stream flows. Error exists in precipitation measurement and in the use of gaged data to estimate precipitation over a watershed area. Any model contains error in its representation of hydrologic processes; and the time series of recorded flows, against which model-simulated events are compared, also contains measurement errors.

Ideally, the modeler can best serve the model user by incorporating within the programming means for computing both (a) the error associated with the estimated values of the parameters, particularly for assessing the calibration, and (b) the error associated with the quantities estimated by the model, particularly for assessing the results. Few hydrologic models, however, are programmed with either capability. Hydrologists need to pursue application of state estimation techniques, specifically Kalman and Extended Kalman filters (Gelb, 1974) for these purposes. Estimates of both parameters and flows and their respective associated errors could with such a system be updated as new data become available.

The crux of the Kalman filter technique (Kalman, 1960) is a combination of two independent estimates of system state to produce an optimal or minimum variance estimate. The measurements of the prototype and the predictions from the model provide the two independent estimates. The weighting used for combining the two estimates is based on the relative uncertainties in the two processes. For the measurements, this uncertainty is represented by the vector $\underline{V}$ (t) for values for m measurement items. For the model, the uncertainty is represented by estimation error propagated by the system model and specified by vector $\underline{W}$(t). Larger model uncertainty necessitates placing less weight on model predictions and more on the measurements.

A simple form of the Kalman filter represents the model of the system by the first order differential equation:

$$\frac{d\underline{X}}{dt} = \underline{f}\ (\underline{X},t) + \underline{W}(t) \quad \ldots\ldots\ldots\ldots\ldots\ldots [11.5]$$

where $\underline{X}$ is a vector of n state variables (perhaps different moisture storages or flows), $\underline{f}$ is a driving function for estimating values for the n state variables at time t given their values at time t-1, and $\underline{W}$(t) represents model error with respect to each state variable. $\underline{W}$(t) thus accounts for such varied modeling errors as neglected parameters, linearization of nonlinear functions, incorrect input data, etc. Measurements of the system are represented by:

$$\underline{Y}\ (t) = \underline{h}\ (\underline{X},t) + \underline{V}\ (t) \ldots\ldots\ldots\ldots\ldots\ldots [11.6]$$

where $\underline{Y}$(t) is a vector of m measured variables (those of the n state variables that are measured, $\underline{h}$ is function for estimating variables from measurements of the m variables, and $\underline{V}$(t) is the noise vector for the m measurement variables. For hydrologic modeling, the m measurements will be less than the n states because of storages or flow points included in the model but for which no measured data are available for comparison.

Equation [11.6] could, for example, represent estimation of metered flows $\underline{Y}$ from stage-discharge relationship $\underline{h}$, where $\underline{V}$ represents the uncertainty involved in converting stage to discharge. Equation [11.5] could then represent a modeling of flows $\underline{X}$ by function f from input parameters, initial conditions, and precipitation, where $\underline{W}$ represents model uncertainty. Given $\underline{h}$, one could estimate $\underline{X}_1$ the vector of flows at time t = 1, from equation [11.6] with error $\underline{V}$. Given parameter values and input data, one could estimate $\underline{X}_2$ from equation [11.5] by integrating over the time interval with error W. These two estimates would then be weighted according to respective error magnitudes for an estimate of $\underline{X}$. From these results, one could obtain, through equation [11.5], associated errors in estimated parameter values.

As examples of successful applications of models of the form of equations [11.5] and [11.6], considerable success has been achieved in flood forecasting using relatively simple (Todini and Wallis, 1978) and more complicated models (Kitanidis and Bras, 1978). These applications use measurements of precipitation and streamflow as they become available to update model parameters, noise component estimates, and short-term forecasts. Models

may be calibrated using, e.g., recorded precipitation and streamflow; the model error covariance matrix is automatically obtained. Predictions with this type of model using a reduced measurement vector (e.g. precipitation) yield streamflow plus a measure of prediction error. Lettermaier and Burges (1976) and Bowles and Grenney (1978) report other applications. Traditional approaches do not provide comparable prediction errors. Much work is underway with these types of models.

The conventional approach to model calibration does not account for uncertainties included in the filter models explicitly; values for coefficients (called parameters) in deterministic models are estimated by assuming no error in any of the inputs, in the model, or in the quantities compared with model output. One, for example, calibrates the Stanford Watershed Model (Crawford and Linsley, 1966) by choosing a time period having concurrent measurements of model input and output time series (precipitation, evapotranspiration, streamflow, etc.) and forcing the model output to as close a match as possible with the measured physical system output. The quality of the calibration can then be determined via model testing, i.e., examining differences between measured and predicted outputs for a different period of record. It is this testing that measures the validity of the calibrated model.

Practical Help

The model user should seek about five consecutive years of continuously recorded hydrologic and meteorologic data which include low flow and high flow events reflecting the range of flow and storm phenomena experienced by that watershed. Experience and observations in nearby watersheds are useful when shorter records are all that is available at a site of interest or when unusual events of interest have not been recorded.

The user next needs to decide whether he wants to calibrate the model to represent the watershed as a whole or if it would be better to divide the total area into parts for separate calibration. If subdivision is desired, how many subwatershed divisions should be used? Recorded outputs are needed from each (or most) of the subwatersheds for calibration. Quite often a modeler is forced to treat hydrologically diverse subwatersheds as one lumped system simply because there is no practical way to estimate parameters satisfactorily in more homogeneous subwatershed units. If division of the watershed is required, at least one or two more years of record reduces compensating errors which could give erroneous parameter estimates. The art of hydrology can be used, of course, as a substitute for some additional measurements.

To begin a calibration, it is necessary to estimate initial values of water stored in various compartments of the model as well as to make initial estimates for the parameters. Starting values for initial conditions can be estimated by assigning numerical values based on observable phenomena. For example, in models having conceptual water storages representing interception, surface runoff, water in the top few inches of soil, etc., the initial values will usually be zero if it did not rain during the preceding week. An antecedent precipitation index, API, is helpful in estimating the initial moisture content of deeper storages. Starting values for the parameters are best estimated from previous calibrations on other watersheds with similar physical characteristics. With these initial storage values and parameter estimates, runoff is simulated from precipitation, simulated runoff is compared with recorded runoff, and the quality of the simulation is evaluated.

The quality of a simulation depends on the following requirements:

1 Annual flow volume for each water year should agree with the recorded values.

2 Seasonal flow volumes should agree with the corresponding recorded values.

3 Weekly and daily volumes should agree with recorded values.

4 Simulated and recorded hydrographs corresponding to specific precipitation events should have similar shapes, the same peak values, and the same time of occurrence to peak flow.

When calibrating, initial emphasis is placed on the first two requirements. Finer tuning concentrates on items 3 and 4. When all four requirements have been satisfied, further checks must be made to see that the model has not given the right answers for the wrong reasons. Such a situation is likely to result for models with multiple peaked objective function response surfaces when the selected calibration is taken from the wrong peak. Greater model credibility results if:

5 Predicted evapotranspiration (ET) is less than potential evapotranspiration for the area.

6 Modeled water storage contents fluctuate with precipitation patterns.

7 The values estimated for model parameters are in accord with physically observed watershed characteristics.

8 The modeled division between surface runoff and base flow is reasonable in light of known watershed soil and geological conditions.

It might be possible, for example, to satisfy requirements 1 to 4 yet drastically mismatch actual and simulated ET. Peak flows, for example, may have been matched by synthesizing too much surface runoff and too little base flow.

When calibrating flood hydrographs, time distributions as well as flows should be compared. If, for example, the simulated agrees with the recorded hydrograph in shape and volume but comes either earlier or later in time, one needs to look to possible causes. They are of two sorts. The rain gage, either because of improper synchronization with the stream gage, or because of its juxtaposition to the watershed and prevailing storm patterns, may be providing a faulty time relationship. The model parameters may represent channel travel times inadequately. If the model is calibrated to match recorded flows during a period of improper synchronization, major problems will occur if the results are tested for another period without that timing problem.

The most important factor in obtaining a good calibration is the representativeness of the precipitation data. It is desirable to have precipitation gages located within the watershed boundary but possible to use gages located nearby. What distance constitutes nearby and how should these gages outside the watershed be used? The answers depend on precipitation areal and temporal variability patterns in the region. For example, a recording gage at the Seattle-Tacoma International Airport is not representative of precipitation amounts ten miles to the northwest. Therefore, it is not possible to use precipitation amounts measured at the airport to calibrate accurately many of the small watersheds that drain into Lake Washington. How does one proceed in such circumstances? The most common approach is to use locally measured daily storage gage totals and the recording gage data to distribute that total over periods between storage gage readings. Precipitation at the master gage is scaled (by a different

factor for every day if necessary) to reflect actual precipitation totals on the watershed.

Which storms should be used to match recorded and simulated hydrographs? Storms resulting from major frontal activity usually provide the best results because proportional precipitation over a large area is more likely. Lesser frontal storms differ markedly in areal variability. Consistent matching of hydrographs resulting from convective storms is made hopeless by the spatial and temporal variability of such storms. The only way to overcome this problem is through a dense network of precipitation gages or, in the future, combination of one or more precipitation gages with indirect precipitation measurement techniques.

The most likely sources of data error are inoperable or non-operating equipment, erroneous transcribing of measured quantities, and systematic errors in the measurements. Precipitation data should be examined for large events. For example, the data during the calibration period should be compared with maximum hourly, 6 hourly, 24 hourly, etc., precipitation records. Moderate events should be compared with regional precipitation records and weather maps for the same date. Even after these checks, some important errors may not be recognized until the simulation is effected. For example, if a large precipitation amount is not matched with a large hydrograph, both records should be checked for the events in question. Often, a mismatch between simulated and recorded events in a model, that is otherwise doing well, points to an error in either the recorded precipitation or streamflow. For example, in one situation, it was impossible to obtain agreement between recorded and simulated flows during winter months because the recorded streamflow data had been incorrectly adjusted for the stream icing over.

GRAPHICAL CALIBRATION AIDS

Since there is no unique criterion that defines a good calibration, several summary statistics that might be compared are discussed below. The problem is compounded because shortcomings in the model or data can be accommodated by the parameters during calibration. For example, a seemingly good calibration may result from parameters adjusted in a way that compensates for a systematic measurement error. Such a calibration can be expected to prove quite poor when tested over some other time period when the measurement error did not occur.

It is advisable that many calibration aids be employed. Graphical comparisons are extremely useful, and in fact, some form of computer graphics is almost essential. Four particular displays are recommended:

1 Continuous time series plot of simulated flow superposed on the recorded flow. (A plot on cartesian paper emphasizes differences for high flows whereas a plot with flow on a log scale normalizes for easy comparison of both high and low flows).

2 Continuous time series plot of the difference between recorded and simulated series. (This may be normalized by, say, recorded flow for nonzero flows, to show relative errors.)

3 Plot of cumulative sum of departures from the mean for the simulated series superposed on the cumulative sum of departures of recorded flow. (The recorded flow mean should be used in both instances.)

4 A scattergram of recorded data plotted against simulated flows as shown on Fig. 11.1 (Bates, 1976). While this method does not preserve the flow sequences contained in the time series plots, differences between a

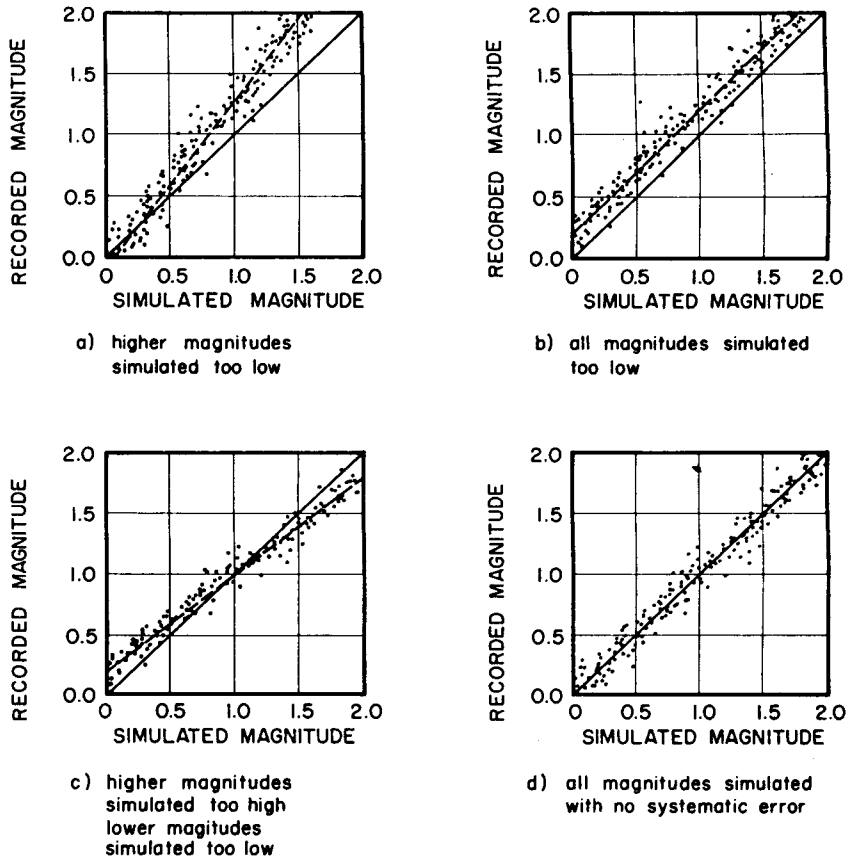

FIG. 11.1 Scattergrams showing presence of systematic errors.

linear regression line through the plotted points and the equality line of the scattergrams help identify errors that cannot be detected as easily from time series plots.

Any desired time aggregation can be used with these four graphical displays, but average daily flow is most common.

If the calibration is only for one watershed at one stream gage location, the graphical displays are easily interpreted. If, however (as is the case in many design situations), numerous basins are being calibrated simultaneously and many of the points of comparison represent the combined runoff from several, careful organization of the displays is required to develop a viewing system for judging the quality of simultaneous simulation at a series of points. Comparison of daily data at many sites along a channel is often physically impractical. Longer periods of aggregation may help the viewer to absorb relevant information. Weekly information for a 3 to 5 yr period gives 150 to 250 observations per site to compare. One suggestion is to plot, one below the other, information at corresponding times at the several sites.

These graphical techniques stimulate very important man-machine interaction in model calibration. Difficulties in objective measurement of goodness of fit have caused some modelers to rely almost entirely on sub-

jective evaluation of agreement between time series. On the other hand, a rigid application of objective measures can replace good hydrologic judgment to the detriment of model performance. Balanced man-machine interaction avoids both extremes.

NUMERICAL CALIBRATION AIDS

Several types of statistics provide useful numerical measures of the degree of agreement between modeled and recorded quantities. Basically, one can compute and display comparisons for each item or develop and use summary statistics for groups of items. Selection requires choices on how to measure differences and on how to aggregate groups of measured differences in a single statistic.

For evaluating the results of hydrologic simulation, one is normally comparing time series. Bates (1976) defined three ways of measuring errors in time series data as

A Absolute errors in magnitudes of flows or other items in the time series as defined by:

$$eA(i) = S(i) - R(i) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad [11.7]$$

where the S(i) are the i items in the simulated and the R(i) are the recorded series.

B Relative errors in time series magnitudes with respect to the raw data origin as defined by:

$$eB(i) = (S(i) - R(i))/R(i) \quad \ldots\ldots\ldots\ldots\ldots \quad [11.8]$$

C Relative errors in time series magnitudes with respect to some other origin as defined by:

$$eC(i) = (S(i) - R(i))/(R(i) - g) \quad \ldots\ldots\ldots\ldots \quad [11.9]$$

where g is a constant equal to the value of the series at the point of the other origin. One example of where a modeler might perfer another origin would be in modeling surface runoff and only caring about runoff above the baseflow.

The i can be defined for any suitable time period such as day, month, or year. The denominators in equations [11.8] and [11.9] must be nonzero.

If a model user is governing his calibration on minimizing departures between recorded and synthesized flow (equation [11.7]), the results will be weighted so heavily toward accurate reproduction of the larger flows that the determined optimum calibration may miss the low flows by an order of magnitude. One approach to overcoming this problem is to normalize by dividing each synthesized flow by the corresponding recorded flow (equation [11.8]). Two difficulties surface with this method. Extremely low flows may be associated with large ratios even though the magnitude of the difference may be small. Second, a flow synthesized at one tenth the recorded value will have a departure of 0.9 and hence be given much less weight than a flow synthesized at ten times the recorded value and having a departure of 9. To overcome these difficulties, Liou (1970, pp. 99-102) proposed the calibration statistic:

D

$$eD(i) = \frac{S(i) + a}{R(i) + a} - 1 \quad S(i) > R(i)$$

$$eD(i) = 1 - \frac{R(i) + a}{S(i) + a} \quad S(i) \leqslant R(i) \quad \ldots [11.10]$$

where a is a constant set according to the desired emphasis in calibration, a larger value gives greater weight to matching higher flows.

Bates (1976) defined two other ways for measuring errors based on the pattern of change within a time series.

E Absolute errors in the magnitude of differences between consecutive items in a time series as defined by:

$$eE(i) = \Delta S(i) - \Delta R(i) \quad \ldots [11.11]$$

where $\Delta S(i) = S(i) - S(i\text{-}1)$; and $\Delta R(i) = R(i) - R(i\text{-}1)$

F Relative errors in time series differences as defined by:

$$eF(i) = (\Delta S(i) - \Delta R(i))/\Delta R(i) \quad \ldots [11.12]$$

The above error series may be computed either directly from numerical data or from some transform such as the logarithm of the data. After they are computed, the error series may be used directly or transferred by squaring or raising to some other power. The log series generally provide a better comparison of relative error.

If errors at higher magnitudes (e.g., flood hydrographs) are more important, absolute errors (A) are of interest. If errors near some other level are more important, then relative errors with respect to an origin at that level (B or C) are of interest. For symmetrical treatment of relative errors around the origin, equation [11.10D] provides an option. For errors in differences between consecutive time series items, if higher rates of change are important, the absolute errors (E) are of interest. If all rates of change are equally important, then the relative errors (F) are of interest.

Selection among the expressions must take into account such computational difficulties as zero occurring in the denominator of a relative error or when the mean or other statistic is extracted (equation [11.9]) by substraction from the recorded value. If zeros are absent, no problem will be encountered with the relative errors; extraction of a value of a size within the range of the process may, however, cause a zero denominator. Rejection limits should be set on the term R(i)-g such that no zero values are obtained. The rejection limit is arbitrary; a logical value would be some multiple (typically 2 or 3) of the standard error of the value R(i). If negative data are encountered and relative errors are of interest, all data should be transferred to positive values by addition of a constant; this gives a new arbitrary origin.

A number of alternatives can be used to condense these error series into more compact effectiveness measures. Clarke (1973) gave a thorough discussion of the alternatives. As an illustration of the variety of opinion on the subject and the danger of choosing a single promising effectiveness measure, Aitken (1973) illustrated a number of alternatives and favored a residual mass curve coefficient (similar to equation [11.33]). Later, Wallis and Todini (1975) showed how the residual mass curve coefficient is functionally analogous to the adjusted range (Rodriguez-Iturbe et al., 1972) and

indicated that the presence of a weak seasonal cycle in the cumulative sum of departures series makes the residual mass curve coefficient inconclusive. Their comments serve to emphasize the elusiveness of a single effectiveness measure to judge model performance.

Some of the problems that can result from using a single coefficient as a measure of performance can be illustrated by considering a simple linear regression equation. If the linear relationship $y = mx$ has a coefficient of determination, D, of unity, all data fall on a straight line. If m equals unity, simulated flows equal recorded flows rather than some constant multiple of recorded flows. Neither number alone tells how good a calibration is because departures of m from unity indicate systematic error and departures of D from unity indicate random error.

McCuen and Snyder (1975) proposed weighting the Pearson product-moment correlation coefficient by the ratio of the standard deviations as a simple index for comparing hydrographs of specific events. If R(i) are values read off the recorded hydrograph at some selected time increment and S(i) are corresponding simulated values, the Pearson moment is

$$r = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{R(i) - \bar{R}}{S_R} \right) \left(\frac{S(i) - \bar{S}}{S_S} \right) \quad \ldots\ldots [11.13]$$

where $\bar{R}$ and $\bar{S}$ are the means and S_R and S_S are the standard deviations of the N values in the recorded and simulated hydrographs respectively. The weighted moment is cr where c equals S_S/S_R if $S_S < S_R$ and c equals S_R/S_S if $S_S \geq S_R$. The weighted moment is thus always smaller than the Pearson moment and approaches unity with perfect agreement between $R_{(i)}$ and $S_{(i)}$.

The error series (eA(i) to eF(i)) defined by equations [11.7] to [11.12] can be represented by summary statistics, mean (me), standard deviation (se), and serial correlation coefficient (re) as follows:

$$me = \sum_{i=1}^{N} e(i)/N \quad \ldots\ldots [11.14]$$

$$se = \left\{ \sum_{i=1}^{N} (e(i) - me)^2/(N-1) \right\}^{1/2} \quad \ldots\ldots [11.15]$$

$$re = \left\{ \sum_{i=1}^{N} (e(i) - me) * (e(i-1) - me) \right\} /se^2 \quad \ldots\ldots [11.16]$$

Equation [11.16] is only defined for continuous series. The summary statistics are measures of aggregate agreement of the simulated with the recorded time series and can be used to evaluate the performance of a model. For summary statistics calculated using the magnitude errors (equations [11.7] to [11.10]), the mean indicates if the model is under- or over-simulating; a non-zero value suggests a probable systematic error with respect to magnitude. The standard deviation gives a measure of the range of errors in the probability distribution of the error occurrence. The serial correlation gives a measure of the impact that an error in previous periods has on subsequent periods (systematic errors with respect to time). For the difference errors (equations [11.11] and [11.12]), the statistics indicate the ability of the

model to predict rates of process change instead of the magnitudes of the process. The serial correlation gives an indication of the effect of error persistence since a low correlation means errors are damped out rapidly.

All these statistics are useful in evaluating the adequacy of a given calibration. If, however, one needs to choose between models instead of just between calibrations of a given model, it is important to explore the distribution of results or the variation within the process. To allow for the process variation, the statistics of the error magnitudes may be standardized to a zero-mean-unit-variance process with the error series denoted as

$$e(i) = (S(i) - R(i))/S_R \quad \ldots\ldots\ldots\ldots\ldots\ldots [11.17]$$

where s_R is the standard deviation of the recorded process, R(i).

The lag serial correlation can be used for processes which are periodic (monthly streamflows are a good example). A periodic process has less variation from one time period to the next than a process which is not periodic, even if both have the same overall variance. The effect of periodicity on the statistics can be removed if the ability of the model to account for differences in the process, normalized by the variance of the differences of the process, is considered. The error series is:

$$e(i) = (\Delta S(i) - \Delta R(i))/s\Delta R \quad \ldots\ldots\ldots\ldots [11.18]$$

where $s\Delta R$ is the standard deviation of the series $\Delta R(i)$.

All the statistics and summary statistics of the above error series give an indication of disagreement between two time series. All parameters approach zero (except for the weighted Pearson moment which approaches unity) as the two time series approach complete agreement. For example, a model is judged to be well calibrated when eF(i) tends toward zero, and the calibration is judged unbiased when the values of eF(i) follow a pattern of white noise around the origin.

A common method for evaluation of time series agreement is by examination of the sum of the squared differences. Limitations of this approach have been discussed by Clarke(1973) and Wallis and Todini (1975) among others. The respective equations for the six error measurements presented in equations [11.7] to [11.12] are

$$CP_A = \sum_{i=1}^{N} eA(i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots [11.19]$$

$$CP_B = \sum_{i=1}^{N} eB(i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots [11.20]$$

$$CP_C = \sum_{i=1}^{N} eC(i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots [11.21]$$

$$CP_D = \sum_{i=1}^{N} eD(i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots [11.22]$$

$$CP_E = \sum_{i=1}^{N} eE(i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots [11.23]$$

$$CP_F = \sum_{i=1}^{N} eF(i)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots [11.24]$$

CP_A is used to designate "coefficient of performance, error series A." If desired, any of the above terms may be divided by the length of the series to obtain a measure of the error in individual values within the series. To compare models of processes having different degrees of variation, equations [11.19] to [11.21], [11.23] to [11.24] are divided by the variation of the process:

$$CP'_A = \frac{CP_A}{\sum_{i=1}^{N} (R(i) - mR)^2} \quad \ldots\ldots\ldots\ldots [11.25]$$

$$CP'_B = \frac{CP_B}{\sum_{i=1}^{N} (R(i)/mR - 1)^2} \quad \ldots\ldots\ldots\ldots [11.26]$$

$$CP'_C = \frac{CP_C}{\sum_{i=1}^{N} [R(i)/(R(i) - g) - g/(R(i) - g)]^2} \quad [11.27]$$

$$CP'_E \frac{CP_E}{\sum_{i=1}^{N} (\Delta R(i) - m\Delta R)^2} \quad \ldots\ldots\ldots\ldots [11.28]$$

$$CP'_F = \frac{CP_F}{\sum_{i=1}^{N} (\Delta R(i)/m\Delta R - 1)^2} \quad \ldots\ldots\ldots [11.29]$$

All of the above indices of error approach zero as two time series reach complete agreement. The indices may be arranged to approach any value p by adding or subtracting the index from p. The most commonly used constant (p) is unity, as shown below the error type A (equation [11.19]):

$$CP'_A{}^* = 1 - CP'_A, \text{ or}$$

$$CP'_A{}^* = \frac{\sum_{i=1}^{N} (R(i) - mR)^2 - \sum_{i=1}^{N} (S(i) - R(i))^2}{\sum_{i=1}^{N} (R(i) - mR)^2}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [11.30]$$

Equation [11.30] is similar to the coefficient of efficiency presented by Nash and Sutcliffe (1971).

If the simulated values are replaced by values, SR(i), estimated from the regression of R(i) on S(i), and the simulated differences by values, ΔSR(i), estimated from the regression of ΔR(i) on ΔS(i), the preceding equations may be used to determine the degree of systematic errors in simulated magnitudes and rates given by the model procedure. If model errors of the type exposed in a coefficient (eA to eF) are completely random in magnitude, that type of error will plot randomly about a 45 deg line on a scattergram, and the values of equations [11.19] to [11.29] will be equal to those with regressed values substituted for the simulated values. Equation [11.30] with the regressed values substituted is similar to the coefficient of determination given by Aitken (1973).

Differences in the magnitudes of the six coefficients (equations [11.7] to [11.12]) indicate the amount of systematic error in the model. Depending on which coefficients are larger or smaller, it is possible to determine if relative or absolute error dominates. Likewise, in the statistical evaluation of the error, if differences (ΔS(i) or ΔR(i)) rather than magnitudes (S(i) or R(i)) are examined, the effect of periodic processes is removed to give equal evaluation of process models.

Accumulated errors may be considered by examining differences in the residual mass curves. The equations are:

$$CP''_A = \sum_{j=1}^{N} \left\{ \sum_{i=1}^{j} S(i) - \sum_{i=1}^{j} R(i) \right\}^2 \quad \ldots\ldots [11.31]$$

$$CP''_B = \sum_{j=1}^{N} \left\{ \frac{\sum_{i=1}^{j} S(i) - \sum_{i=1}^{j} R(i)}{\sum_{i=1}^{j} R(i)} \right\}^2 \quad \ldots\ldots [11.32]$$

$$CP''_C = \sum_{j=1}^{N} \left\{ \frac{\sum_{i=1}^{j} S(i) - \sum_{i=1}^{j} R(i)}{\sum_{i=1}^{j} (R(i) - mR)} \right\}^2 \quad \ldots\ldots [11.33]$$

Equation [11.33] is similar to that Aitken (1973) termed the residual mass curve coefficient. Similarly as for the previous coefficients, regressed values of S(i) with the mean error removed, can be substituted into the coefficients to determine the degree of systematic error in the residual mass of the simulated process.

SYSTEMATIC APPROACH TO CALIBRATION AND TESTING

A systematic approach to model calibration and testing is outlined in Fig. 11.2. Three stages are involved: preliminary calibration, refined calibration, and model testing and evaluation. The first stage uses a preliminary inspection followed by a more refined comparison using the model to obtain the best possible data for the model. The second stage uses the resulting meteorological inputs and streamflows to achieve the best possible calibration. In the third stage, data measured at the same stations, but for a different time period than used for calibration, are used to test the calibrated model and determine its accuracy.

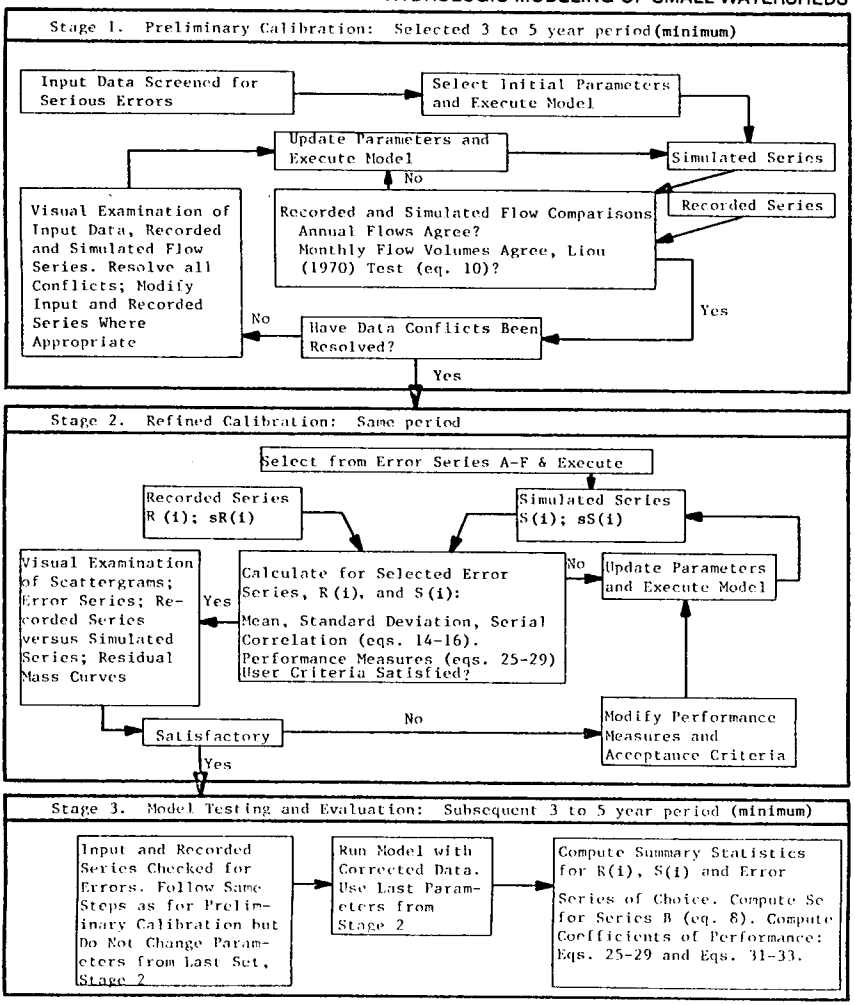

FIG. 11.2 Steps involved in model calibration and testing.

At the start of the preliminary calibration stage, input data (precipitation, evapotranspiration, etc.) and the recorded streamflow series (at all locations of interest) need to be screened for obvious errors. The two major checks are for (a) recorded hydrographs whose volume exceeds recorded rainfall, or is such a high fraction of recorded rainfall as to be highly improbable, a condition usually associated with major storms over the watershed but largely missing the precipitation gages, and (b) large recorded rainfalls simultaneous with little runoff, a condition usually associated with intense rainfalls registered at the gage not being representative of precipitation over the watershed. If storms with these properties produce a major portion of the annual runoff in a given year, it is normally best to use other, better years for calibration. When such storms occur during years used for calibration, adjustment of the precipitation data to values the model suggests as commensurate with recorded streamflow is recommended in order to avoid a major upset in antecedent moisture conditions for the next storm, but the time period covered by the adjusted data should not be used in calculating test statistics for parameter optimization.

Actual modeling begins as initial conditions and initial parameter values are assigned, and streamflows are simulated for the input data. The recorded and simulated series are compared for agreement between annual and seasonal flow volumes. Usually it is desirable for monthly and perhaps weekly flow volumes to agree at this stage too. Liou's (1970) test statistic (equation [11.10]) is useful here. The mean and standard deviation of daily flow are computed and compared to ensure that the simulated and recorded flows have comparable low order moments. Model parameters are updated and calibration continued until the preliminary calibration criteria have been satisfied. Parameters may be changed manually or by programmed rules.

When acceptable preliminary agreement is achieved (e.g. recorded and simulated annual flow values agree to within 10 percent), visual examination of input data and recorded and simulated streamflow should be made to detect remaining conflicts, e.g. recorded data being in error, incompatibility between precipitation and runoff events, etc. Conflicts must be resolved by correcting the input and recorded series; even a relatively crudely calibrated model is very effective in identifying errors in recorded data.

Once the preliminary calibration has developed compatible data series, the calibration needs to be refined to obtain the best possible matching of daily hydrographs; shorter time scales are of interest for certain flood events. The quantification of best is based on selection by the user of one or more of the error series and performance measures discussed above. After a calibration achieves the user's criteria with respect to the numerical performance measures, it is important to use graphical aids to assess visually the quality of the calibration.

A scattergram of recorded and simulated flow is the simplest method to view the calibration quality in one convenient graph. Stochastic information is lost but the user can determine immediately if high and low flows are being modeled equally well or systematic errors are present (Fig. 11.1). It is often useful to plot the scattergram of the logarithms of flow to look at poor model performance at low flows. Overall stochastic behavior is exhibited best by plotting one or more of the error series as well as the residual mass curve. Most users prefer to plot recorded and simulated flow against time. Additionally, it is valuable to plot series B (equation [11.8]) for a visual display of relative error. The residual mass curves of recorded and simulated flow provide an excellent means for viewing error persistence.

The user must specify error series and performance measures with desired values for each as acceptance criteria which indicate a satisfactory calibration. If, after examining a time series of simulated flows that achieves these criteria, he desires added refinement, the acceptance criteria must be made more demanding; perhaps error series and performance measures should be changed. The parameters are modified and the model is rerun; the procedure is continued until the new criteria are satisfied.

Some thought needs to be given to selecting appropriate criteria to use at the beginning of the calibration. Error series A and B (equations [11.7] and [11.8] are of most general interest. Series D (equation [11.10]) provides a means for added refinement when results seem too strongly influenced by a few points. While ideally the mean, standard deviation, and lag one correlation coefficients of series B are all zero, practical criteria are to try to attain a mean level to within 5 percent, standard deviation 5 to 10 percent, and satisfy the lag one correlation (p_1) tendency to zero. As a good calibration is approached, p_1 is normally distributed with mean zero and variance $(N-2)/(N-1)^2$ (Anderson, 1942).

For the calibration period (3 to 5 yr), there are between 1095 and 1827 daily values of recorded and simulated flow to compare, so the samples are statistically large. It should be remembered, however, that the individual series (recorded and simulated) are highly auto-correlated so their information content is associated with a much smaller number of independent events. Bayley and Hammersley (1946) give details for computing the reduced information content. For example, the effective sample size to be used when computing the standard error of the mean is N_d^* as defined by:

$$\frac{1}{N_d^*} = \frac{1}{N} + \frac{2}{N^2} \sum_{j=1}^{N-1} (N-j)\rho_j \quad \ldots\ldots\ldots\ldots [11.34]$$

with ρ_j the jth autocorrelation coefficient. Many streamflow series, after nonstationary components have been removed, follow a lag-one Markov process. For such cases N_d^* is

$$\frac{1}{N_d^*} = \frac{1}{N} + \frac{2\rho_1}{N^2\,(1-\rho_1)} \left[N + \frac{(\rho_1^N - 1)}{(1-\rho_1)} \right] \quad \ldots [11.35]$$

Two other criteria are useful to set. These are values for the coefficient of efficiency (equation [11.30]) and the coefficient of residual mass (equation [11.33]). Excellent calibrations are achieved if both coefficients exceed 0.97. Generally, criteria that apply to error series B are more restrictive because the model must replicate both high and low flows satisfactorily. Minimization of the standard deviation of error series B, subject to errors in model structure, input data, and recorded flow data is desirable, but it is unrealistic to expect to achieve values lower than 5 to 10 percent.

The modeler should beware of the possibility of achieving a calibration that reduces prediction error, sSe, below the recorded series relative error, sRe. For such a calibration, the errors in modeling the process are less than the errors in measuring the process. Another danger is for the modeler to match his results too closely to the particular flow series that occurred during the calibrating period and lose generality so as actually to make results worse when the model is tested later on data for other time periods. This is one of the dangers in calibrating using a period that is either too short or that contains non-representative flow sequences.

After the model has been calibrated, a different time period is used to evaluate its performance. Use of the 3 to 5 yr immediately following the calibration period is recommended for verification and evaluation. A major reason for using data immediately following the calibration period is to make use of end conditions from the calibration stage as initial conditions for the evaluation stage. The effects of initial conditions are observable for a period of six to twelve months when using models similar to the Stanford Watershed model. Avoiding uncertainties in initialization by following the above procedure ensures maximum use of scarce data for model evaluation.

An overall statement of the quality of the calibration model can be made by determining the mean and standard deviation of series B (equation [11.8]), the series of relative errors, and the coefficients of efficiency (equation [11.30]) and the residual mass (equation [11.33]). When these coefficients exceed 0.97, very little systematic error remains. Other coefficients may be

used; the above measures suffice for most situations. The standard deviation of error series B gives the most useful general measure of prediction confidence.

It is usually not possible to have a model perform uniformly well over the entire low and high flow range. Therefore, several coefficients of performance might be used. One can subdivide the total time series into several parts, e.g., wet periods, low-flow periods, etc. for each year, and compute performance coefficients for each part. For example, the mean and standard deviation (error series B), could be computed for the wet period in each year, and the computation repeated for all years used for verification to develop average summary statistics.

CALIBRATION PROCEDURE ILLUSTRATION

An example calibration with summary statistics and recorded and simulated streamflow is given in this section to illustrate measures and visual aids helpful in the refined calibration stage (Fig. 11.2). A 407-square kilometer watershed draining to USGS gage 12-1252 located on the Sammamish River, Washington, was subdivided into four sub areas and modeled via the Hydrocomp (1969) HSP rainfall-runoff digital computer model. This model did not contain pattern search parameter optimization features; parameters were changed manually after evaluating each calibration run. An initial calibration had been effected by a user who thought that a good calibration had been achieved.

A refined calibration was based on water years 1966 to 1968. Complete details of the model, parameters used, and the flow simulation (including problems of obtaining representative precipitation, etc.) are given in Bates (1976). The parameters were modified systematically to attempt the best possible calibration. The watershed experienced marked temporal and spatial precipitation variability, particularly during smaller precipitation events; reliable rain gages were not available within the watershed boundaries. Snowmelt runoff was negligible.

For the pollution control application the user had in mind, it was necessary to model high and low flow with comparable relative precision. The unfortunate location of the rain gage (outside the watershed) was a clear warning that the response of the watershed to light to moderate rain would not be predicted with much accuracy. Based on the precision needed in the required application and the constraints the precipitation data posed as to what was possible, selected model acceptance criteria were coefficient of determination, coefficient of efficiency and coefficient of residual mass should all exceed 0.97; mean daily flow in each year (series B, equation [11.8]) $\leq$ 5 percent; standard deviations of daily flows (series B, equation [11.8]) $\leq$ 15 percent; individual monthly volumes $\leq$ 10 percent.

Inspection of Table 11.1, which summarizes results from the preliminary calibration and three refined calibrations, indicates the criteria were too demanding. The particular situation did not lend itself to further parameter adjustment to achieve calibration better than the second refined calibration (Run 2). The statistics for the standard deviation of error series B were approximately the same in each water year for each time increment investigated; therefore, to save unnecessary computations and deluging the viewer with summary information, summaries were computed for the middle water year only for Runs 2 and 3. Added summary statistics are given in Table 11.2. The performance coefficients all exceeded the criteria, but a better calibration could have been achieved had better precipitation data been available.

TABLE 11.1. SUMMARY STATISTICS FOR SIMULATING MODELING, USGS STREAM GAGE 12-1252

Water year	1966	1967				1968			
Flow period	Year	Day	Week	Month	Year	Day	Week	Month	Year
Preliminary	-16.7*	25.0†	21.0†	17.0†	-12.0*	27.0†	22.0†	16.0†	-15.2*
Refined calibration									
Run 1	-3.1	22.0	22.0	15.0	+0.0	23.0	22.0	14.0	+ 4.5
Run 2	5.6	17.0	16.0	12.0	4.1				- 1.8
Run 3	3.9	20.0	17.0	13.0	-1.7				- 6.4

Relative error series (equation [8]) in percent
*Mean Value (daily time increment)
†Standard Deviation (day, week, and month time increments)

Visual evaluation of the simulated flow from Run 2 can be made from Figs. 11.3 to 11.5 which contain daily flow information for water year 1967. Fig. 11.3 indicates that high flows were well modeled. Fig. 11.4, which plots error series A against recorded flow, shows that simulated low flows overestimate recorded low flows. Dawdy (personal communication, 1978) pointed out that the information could have been conveniently displayed had Fig. 11.3 been ploted to logarithmic scales. The display is a matter of personal choice and the type of error one needs to emphasize for a given application.

Most hydrologist draw comfort from examining the recorded and simulated time series on a single plot. These daily flow series are shown in Fig. 11.5 on a linear scale; use of a semi-log scale might be preferred by some users. Inspection of Fig. 11.5 clearly shows the hydrograph recessions were not well modeled and that there is substantial error in the low flow simulations. This graph clearly explains the large standard deviation statistics given in Table 11.1. The cumulative mass curves shown in Fig. 11.5 indicate that little systematic error remains, a fact supported by the three performance coefficients in Table 11.2. Use of the cumulative mass curves is not widespread; the authors find them generally useful to supplement the other summaries.

TABLE 11.2. STATISTICAL SUMMARY (DAILY) FOR CALIBRATED MODEL AT USGS GAGE 12-1252 (SAMMANISH RIVER) FOR RUN 2, WATER YEAR 1967

Statistics of Series	Mean	Standard deviation
Recorded	9.30*	8.71*
Simulated	9.70*	8.37*
Difference (absolute error series A equation [4])	0.4*	1.14*
Percent difference (relative error series B equation [8])	4.1	17.0
Coefficients		
Coefficient of determination (square of correlation coefficient between R and S)		0.984
Coefficient of efficiency, equation [30]		0.981
Coefficient of residual mass, CP''_A, equation [33]		0.980

*Cubic meters per second.

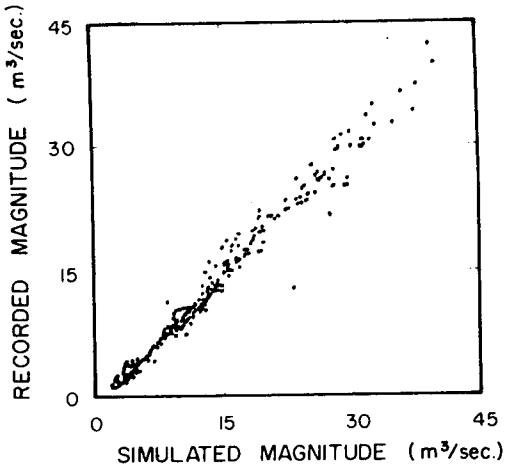

FIG. 11.3 Scattergram of recorded and simulated daily flows at USGS gage 12-1252 for water year 1967.

This illustration provides several lessons. The most important is to show that coefficients of performance are not complete measures of calibration quality. High values of performance coefficients can be attained while low flows are poorly modeled. The utility of the standard deviation of error series B, the relative errors, is clear. This statistic was not close to zero (at its optimum it should represent the accuracy of stream gaging, i.e. be about 2.5 percent for an excellent record) while the coefficients of performance were almost unity. Joint performance measures, as well as final graphical comparison, are essential to achieve physically sensible and accurate calibrations.

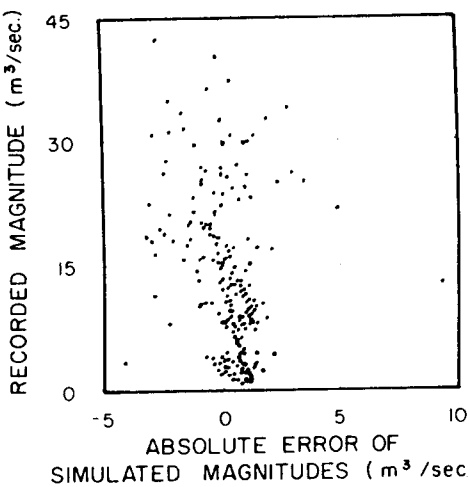

FIG. 11.4 Absolute errors of simulated flows at USGS gage 12-1252 for water year 1967.

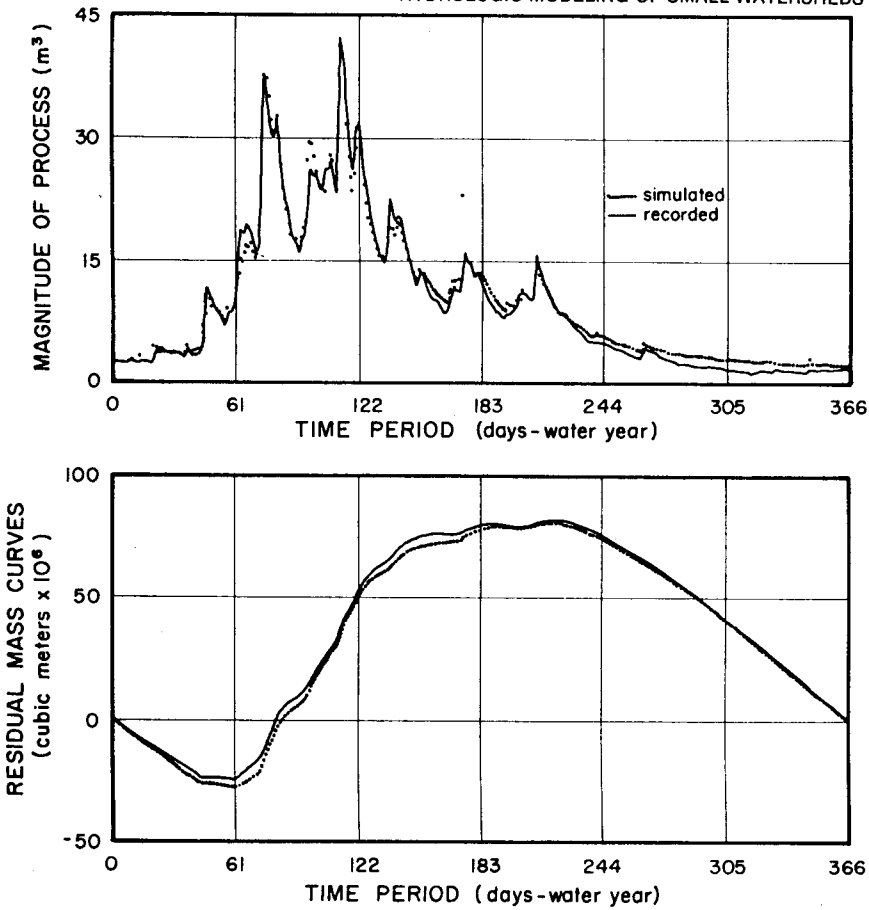

FIG. 11.5 Recorded and simulated time series plots for USGS gage 12-1252 for water year 1967.

SUMMARY

This chapter has reviewed basic issues in selection, calibration, and testing of hydrologic models. The first half describes selection of modeling strategy, conceptualization of the components to include in the selected model type, development of a model structure combining these components into a working whole, calibration to estimate parameter values in the selected model, testing the adequacy of the calibrated results, communicating modeling advances to other scientists, and displaying model results to the user public. The second half provides specific descriptions of graphical and numerical aids that can be used to obtain the best possible calibration for a given watershed and illustrates their power through an example of how they were used to achieve significant improvement over a preliminary calibration.

References

1 Aitken, A. P. 1973. Assessing systematic errors in rainfall-runoff models. Journal of Hydrology 20:121-136.

2 Anderson, R. L. 1942. Distribution of the serial correlation coefficient. Annals Math. Statistics 13:1-13.

3 Bates, C. L. 1976. Analysis of time series modeling errors with application to the Lake Sammamish hydrologic system. MSCE Thesis, Dept. of Civil Engineering, University of Washington, Seattle.

4 Bayley, G. U., and J. M. Hammersley. 1946. The effective number of independent observations in an autocorrelated time series. Journal Royal Statistical Society 8(1B):184-197.

5 Beard, L. R. 1962. Statistical methods in hydrology. Civil Works Investigations Project CW-151, U.S. Army Engineers, Sacramento, CA.

6 Bowles, D. S., and W. J. Grenney. 1978. Steady state river quality modeling by sequential extended Kalman filters. Water Resources Research, 14:84-96.

7 Clarke, R. T. 1973. A review of some mathematical models used in hydrology with observations on their calibration and use. Journal of Hydrology 19:1-20.

8 Coleman, G., and D. G. DeCoursey. 1976. Sensitivity and model variance analysis applied to some evaporation and evapotranspiration models. Water Resources Research 12(5):873-879.

9 Crawford, N. H., and R. K. Linsley. 1966. Digital simulation in hydrology: Stanford Watershed Model IV. Technical Report 39, Dept. of Civil Engineering, Stanford University, Stanford, CA.

10 Dawdy, D. R., and J. M. Bergman. 1969. Effect of rainfall variability on streamflow simulation. Water Resources Research 5(5):958-966.

11 Diskin, M. H. 1970. Objectives and techniques of watershed modeling. Proceedings ARS and SCS Watershed Modeling Workshop, Tucson, AZ, Chapter 3.

12 Fleming, G. 1975. Computer simulation techniques in hydrology. American Elsevier Pub. Co., Inc., New York.

13 Fletcher, J. E. et al. 1976. Runoff estimates for small rural watersheds and development of sound design method. Research Report and Manual, 2 volumes, Utah Water Research Laboratory, Logan.

14 Gelb, A. (Ed.). Applied optimal estimation. The MIT Press, Cambridge, MA.

15 Haan, C. T. 1977. Statistical methods in hydrology. Iowa State University Press, Ames.

16 Haan, C. T., and D. M. Allen. 1972. Comparison of multiple regression and principal component regression for predicting water yields in Kentucky. Water Resources Research 8(6):1593-1596.

17 Hooke, R., and T. A. Jeeves. 1961. Direct search solution of numerical and statistical problems. Journal of the Association of Computing Machines 8(2):212-229.

18 Huang, Y. H., and R. K. Gayner. 1977. Effects of stream channel improvements on downstream floods. Research Report No. 102, Water Resources Research Institute, University of Kentucky, Lexington.

19 Hydrocomp. 1969. Hydrocomp simulation programming operations manual. Hydrocomp, Inc., Palo Alto, CA.

20 James, L. D. 1969. Computers in flood control planning. Proc. of ASCE 95(HY6): 1859-1870.

21-28 Appendix

29 Liou, E. Y. 1970. OPSET: Program for computerized selection of watershed parameter values for the Stanford Watershed Model. Research Report No. 34, Water Resources Institute, University of Kentucky, Lexington.

30 Linsley, R. K., Jr. et al., 1975. Hydrology for engineers. Second Edition. McGraw-Hill Book Co., New York.

31 Lumb, A. M., and L. D. James. 1976. Runoff files for flood hydrograph simulation. Proc. of ASCE 102(HY10):1515-1531.

32 Lumb, A. M. et al. 1975. GTWS: Georgia Tech watershed simulation model. Report ERC-0175, Environmental Resources Center, Georgia Institute of Technology, Atlanta.

33 Maass, A. et al. 1962. Design of water-resource systems. Harvard University Press, Cambridge, MA.

34 McCuen, R. H. 1973. The role of sensitivity analysis in hydrologic modeling. J. of Hydrology 18:37-53.

35 McCuen, R. H., and W. M. Snyder. 1975. A proposed index for comparing hydrographs. Water Resources Research 11(6):1021-1024.

36 Munro, J. C. 1971. Direct search optimization in mathematical modeling and a watershed model application. NOAA Technical Memorandum NWS HYDRO-12.

37 Nash, J. E., and J. V. Sutcliffe. 1971. River flow forecasting through conceptual models. Journal of Hydrology 13:297-324.

38 Overton, D. E., and M. E. Meadows. 1976. Stormwater modeling. Academic Press, New York.

39 Rodriguez-Iturbe, I., J. M. Mejia, and D. Dawdy. 1972. Streamflow simulation, 1. A new look at Markovian models, fractional Gaussian noise and crossing theory. Water Resources Research 8(4):921-930.

40 Saah, A. D. et al. 1976. Development of regional regression equations for solution of certain hydrologic problems in and adjacent to Santa Clara County. Santa Clara Valley Water District, San Jose.

41 Todini, E. and J. R. Wallis. 1978. A real time rainfall runoff model for an on-line flood warning system. Proceedings, Chapman Conference on Application of Kalman Filter to Hydrology, Hydraulics, and Water Resources, Pittsburgh, PA.

42 Wallis, J. R. and E. Todini. 1975. Comment on the residual mass curve coefficient. Journal of Hydrology 24:201-205.

43 Wolf, P. O. 1966. Comparison of methods of flood estimation. Proceedings of the River Flood Hydrology Symposium, the Institution of Civil Engineers, London.

44 Woodward, D. E. 1973. Hydrologic and watershed modeling for watershed planning. TRANSACTIONS of the ASAE 99(3):582-584.

21 James, L. D. 1970. An evaluation of relationships between streamflow patterns and watershed characteristics through the use of OPSET, a self calibrating version of the Stanford Watershed Model. Research Report No. 36, University of Kentucky Water Resources Institute, Lexington, Kentucky.

22 James, L. D. 1972. Hydrologic modeling, parameter estimation, and watershed characteristics. J. of Hydrology, 17:283-307.

23 James, L. D. 1973. Surveys required to design nonstructural measures. Proc. of ASCE, 99(HY10):1823-1836.

24 James, L. D., and R. R. Lee. 1971. Economics of water resources planning. McGraw-Hill Book Co., New York.

25 James, L. D., et al. 1975. Integration of hydrologic, economic, ecologic, social, and well-being factors in planning flood control measures for urban streams. Report ERC-0375, Environmental Resources Center, Georgia Institute of Technology, Atlanta, Georgia.

26 Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. Transactions, ASME, Journal of Basic Engineering, 82(2):35-45.

27 Kitanidis, P. K., and R. L. Bras. 1978. Error identification in conceptual hydrologic models. Proceedings, Chapman Conference on Applications of Kalman Filter to Hydrology, Hydraulics, and Water Resources, Pittsburgh, Pennsylvania.

28 Lettenmaier, D. P., and S. J. Burges. 1976. Use of state estimation techniques in water resource system modeling. Water Resources Bulletin, 12:83-99.