TRAINING COURSE
ON

HYDROLOGICAL MODELING AND GIS

(MAY 26 TO JUNE 06, 2014)

FOR

UNFAO & Ministry of Energy and Water, Afghanistan

LECTURE NOTE
ON

CALIBRATION AND VALIDATION IN HYDROLOGIC MODELING

By

SHARAD K. JAIN
NIH ROORKEE

NATIONAL INSTITUTE OF HYDROLOGY
AND
INDIAN ASSOCIATION OF HYDROLOGISTS

ROORKEE – 247 667 (UTTARAKHAND)

CALIBRATION AND VALIDATION IN HYDROLOGIC MODELING

INTRODUCTION

Once one or more models have been chosen for consideration in a project, it is necessary to address the problem of parameter calibration. In general, it is not possible to measure the values of model parameters or estimate them a priori. Studies that have attempted this have generally found that, even after intensive of measurements of parameter values, the results have not been entirely satisfactory. Prior estimation of feasible ranges of parameters also often results in ranges of predictions that are wide and may still not encompass the measured responses all of the time.

There are two major reasons for these difficulties in calibration. The first is that the scale of the measurement techniques available is generally much less than the scale at which parameter values are required. For example, hydraulic conductivity is a parameter which is frequently found in hydrologic models. Techniques to measure hydraulic conductivities of the soil generally integrate over areas of less than 1 m^2. However, even the most finely distributed models require values that effectively represent the response of an element with an area of 100 m^2 or, in many cases, a much larger area. For saturated flow, there have been some theoretical developments that suggest how such effective values might change with scale, given some underlying knowledge of the fine-scale structure of the conductivity values. In general, however, carrying out the experimental measurements required to use such a theory at the hillslope or catchment scale would be very time-consuming and expensive, and would result in a large number of holes in the hillslopes. Thus it may be necessary to accept that the small-scale values that it is possible to measure and the effective values required at the model element scale are different quantities, or they are incommensurate – even though the hydrologist has traditionally given them the same name. The effective parameter values for a particular model structure will then still need to be calibrated in some way.

Most past calibration studies have involved some form of optimization of the parameter values by comparing the results of repeated simulations with whatever observations of the catchment response are available. The parameter values are adjusted between each run of the model either manually or by some computerized optimization algorithm until some 'best fit' parameter set has been found. There have been many studies of different optimization algorithms and measures of goodness of fit or objective functions in hydrological modelling. The essence of the problem is to find the highest peak in the response surface in the parameter space defined by one or more objective functions. An example of such a response surface is shown in Figure 1. The two axes are two different parameter values, varied between specified maximum and minimum values. The vertical axis is the value of an objective function, based on the sum of squared differences between observed and predicted discharges that has the value 1 for a perfect fit. It is easy to see from this example why optimization algorithms are sometimes called

'hill climbing' algorithms, since the highest point on the surface will represent the optimum values of the two parameters. Such a response surface is easy to visualize in two-parameter space. It is much more difficult to visualize the response surface in an N-dimensional parameter hyperspace. Such surfaces can often be very complex and much of the research on optimization algorithms has been concerned with finding algorithms that are robust with respect to the complexity of the surface in an N-dimensional space and will find the global optimum set of parameter values.

For most hydrological modeling problems, the optimization problem is ill-posed in that if the optimization is based on the comparison of observed and simulated discharges alone, there may not be enough information in the data to support the robust optimization of the parameter values. Experience suggests that even a single model with only four or five parameter values to be estimated may require at least 15 to 20 hydrographs for a reasonably robust calibration, and if there is strong seasonal variability in the storm responses hydrographs for a longer period will be needed. For more complex parameter sets, many more data and different types of data may be required for a robust optimization unless it might be possible to fix many of the parameters beforehand by independent measurement. This has proven to be very difficult to achieve in practice.

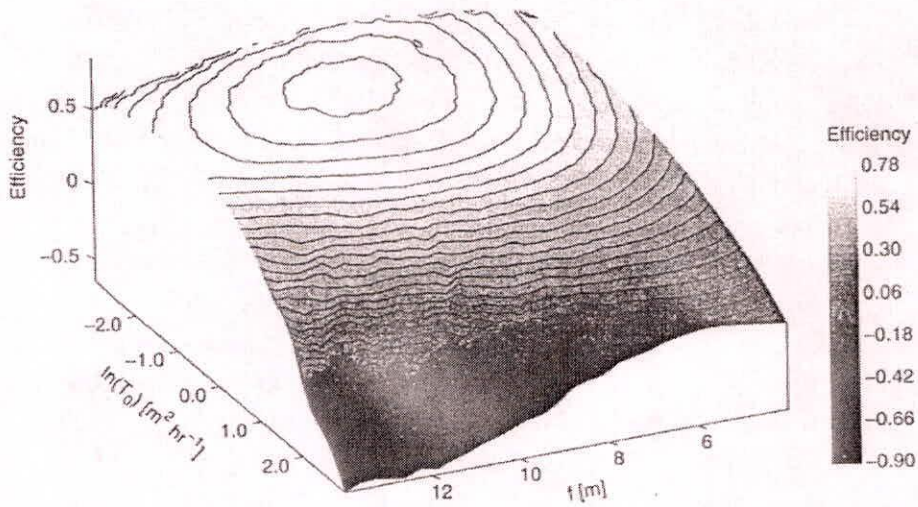

Fig. 1. Response surface for two parameters of model in an application to modeling the stream discharge of a small catchment. The objective function is the Nash-Sutcliffe efficiency which has a value of 1 for a perfect fit of the observed discharges.

These are not the only problems with finding an optimum parameter set. Optimization generally assumes that the observations with which the simulations are compared are error-free and that the model is a true representation of that data. We know, however, at least for hydrological models, that both the model structure and the observations are not error-free. Thus the optimum parameter set found for a particular model structure may be sensitive both to small changes in the observations, or the period

of observations considered in the calibration, and possibly to changes in the model structure such as a change in the element discretization for a distributed model.

A number of important implications follow from these considerations:

- The parameter values determined by calibration are effectively valid only inside the model structure used in the calibration. It may not be appropriate to use those values in different models (even though the parameter may have the same name) or in different catchments.
- The concept of an optimum parameter set may be ill-founded in hydrological modeling. While one optimum parameter set can often be found there will usually be many other parameter sets that are very nearly as good, perhaps from very different parts of the parameter space. It is most unlikely that, given a number of parameter sets that give reasonable fits to the data, the ranking of those sets in terms of the objective function will be the same for different periods of calibration data. Thus to decide that one set of parameter values is the optimum is a somewhat arbitrary choice.
- If the concept of an optimum parameter set must be superseded by the idea that many possible parameter sets (and perhaps models) may provide acceptable simulations of the response of a particular catchment, then if follows that validation of those models may be equally difficult. In fact, rejection of some of the acceptable models given additional data may be a much more practical methodology than suggesting that models might be validated.

EQUIFINALITY OF PARAMETERS

The concept of equifinality of parameters suggests that, given the limitations of both our model structures and observed data, there may be many representations of a catchment that may be equally valid in terms of their ability to produce acceptable simulations of the available data. In essence then, different model structures and sets used within a model structure are competing to be considered acceptable as simulators. Some may be rejected in the process of different model structures, but even if only one model is retained then the evaluation of the performance of different parameter sets against the observed data will usually result in many parameter sets that produce acceptable simulations.

The results with different parameter sets will not, of course, be identical either in simulation or in the predictions required by the modeling project. An optimum parameter set will give only a single prediction. Multiple acceptable parameter sets will give a range of predictions. This may actually be an advantage since it allows the possibility of assessing the uncertainty in predictions, as conditions on the calibration data, and then using that uncertainty as part of the decision – making process arising from a modeling project.

The starting point in the modeling process is to assume, a priori, all the available modeling strategies and all feasible parameter sets within those modeling strategies are

potential models of a catchment for a particular project. The aims of the given project, the budget available for the project, and the data available for calibrating the different models will all limit the potential range of simulators. The important point is that choices between models and between parameter sets must be made in a logical and scientifically defensible way. At the end of this process, there will not be a single model of the catchment but a number of acceptable models (even if only different parameter sets within one chosen model structure) to provide predictions.

There are clearly implications for other studies that depend on models of rainfall-runoff processes. Predictions of catchment hydrogeochemistry, sediment production and transport, the dispersion of contaminants, hydroecology, and, in general, integrated catchment decision support systems depend crucially on good predictions of water flow processes. Each additional component that is added to a modeling system will add additional choices in terms of the conceptual representation of the processes and the values of the parameters required. In that all these components will depend on the prediction of water flows, they will be subject to the types of uncertainties in predictive capability. This is not only a research issue. Uncertainties in model predictions have already played a major role in decisions made at public inquiries into proposed developments.

PARAMETER ESTIMATION AND PREDICTIVE UNCERTAINTY

Limitations of both model structures and the data available on parameter values, initial conditions and boundary conditions, will generally make it difficult to apply a hydrological model (of whatever type) without some form of calibration. In very few cases reported in the literature have models been applied using only parameter values measured or estimated a priori. In the vast majority of cases the parameter values are adjusted to get a better fit to some observed data. This is the model calibration problem. The question of how to assess whether one model or set of parameter values is better than another is open to a variety of approaches, from a visual inspection of plots of observed and predicted variables, to a number of different quantitative measures of goodness of fit, known variously as objective functions, performance measures, fitness (or misfit) measures, likelihood measures or possibility measures.

All model calibrations and subsequent predictions will be subject to uncertainty. This uncertainty arises in that no rainfall-runoff model is a true reflection of the processes involved, that it is impossible to specify the initial and boundary conditions required by the model with complete accuracy, and that the observational data available for model calibration are not error-free. A good discussion of these sources of uncertainty may be found in Melching (1995). There is a rapidly growing literature on model calibration and the estimation of predictive uncertainty for hydrological models. For the purposes of this discussion, we will differentiate three major themes as follows:

- Methods of model calibration that assume an optimum parameter set and that ignore the estimation of predictive uncertainty can be found. These methods range

from simple trial and error, with parameter value adjusted by the user, to the variety of automatic optimization methods.

- Methods of model calibration that assume an optimum parameter set, but which make certain assumptions about the response surface around that optimum to estimate the predictive uncertainty, can be found. These methods are grouped under the name reliability analysis.
- Methods of model calibration that reject the idea that there is an optimum parameter set in favour of the idea of equifinality of models. Equifinality is the basis of the GLUE methodlogy. In this context it is perhaps more appropriate to use model conditioning rather than model parameter sets that give acceptable simulations. As a result, the predictions will be necessarily associated with some uncertainty.

In approaching the problem of model calibration or conditioning, there are a number of very basic points to keep in mind. These may be summarized as follows:

- It is most unlikely that there will be one right answer. Many different models and parameter sets may give good fits to the data and it may be very difficult to decide whether one is better than another. In particular, having chosen a model structure, the optimum parameter set from one period of observations may not be the optimum set for another period.
- Calibrated parameter values may only be valid inside the particular model structure used. It may not be appropriate to use those values on different models (even though the parameters may have the same name) or in different catchments.
- The model results will be much more sensitive to changes in the values of some parameters than to changes in others. A basic sensitivity analysis should be carried out early on in a study.
- Different performance measures will usually give different results in terms of both the 'optimum' values of parameters and the relative sensitivity of different parameters.
- Sensitivity may also depend on the period of data use, and especially whether a particular component of the model is being 'exercised' in a particular period. If it is not (e.g., if an infiltration excess runoff production component only gets to be used under extreme rainfalls), then the parameters associated with these components will generally appear insensitive.
- Model calibration has many of the features of a simple regression analysis in that an optimum parameter set will be one that, in some sense, minimizes the overall error or residuals. There are still residuals, however, and this implies uncertainty in the predictions of a calibrated model. As in regression, these uncertainties will normally get larger as the model predicts the response for more and more extreme conditions relative to the data used in calibration.

PARAMETER RESPONSE SURFACES AND SENSITIVITY ANALYSIS

Consider, for simplicity, a model with only two parameters. Some initial values of the parameters are chosen and the model is run with a calibration data set. The resulting predictions are compared with some observed variables and a measure of goodness of fit is calculated and scaled so that if the model was a perfect fit the goodness of fit would have a value of 1.0, and if the fit was very poor it would have a value of 0. Assume that the first run resulted in a goodness of fit of 0.72, i.e. we would hope that the model could do better (get closer to a value of 1). It is a relatively simple matter to set up the model to change the values of the parameters, make another run, and recalculate the goodness of fit. However, how to decide which parameter values to change in order to improve the fit?

One way is by simple trial and error, plotting the results on screen, thinking about the role of each parameter in the model, and changing the values to make the hydrograph peaks higher, or the recessions longer, or whatever is needed. This can be very instructive, but as the number of parameters gets larger it becomes more and more difficult to sort out all the different interactions of different parameters in the model and decide what to change next.

Another way is to make enough model runs to evaluate the model performance in the whole of the parameter space. In the simple two-parameter example, one could decide on a range of values for each parameter, use 10 discrete increments on each parameter range, and run the model for every combination of parameter values. The ranges of the parameters define the parameter space. Plotting the resulting values of goodness of fit defines a parameter response surface such as that shown as contours in Figure 2. In this example, 10 discrete increments would require $10^2 = 100$ runs of the model. For simple models this should not take too long. The same strategy for three parameters is a bit more demanding: 10^3 runs would be required. For six parameters, 10^6 or a million runs (about two weeks of computing for a simple model on a PC, and very much more for more complex models) would be required, and 10 increments per parameter is not a very fine discretization of the parameter space. Not all those runs, of course, would result in models giving good fits to the data. A lot of computer time could therefore be saved by avoiding model runs that give poor fits. This is a major reason why there has been so much research into automatic optimization techniques, which aim to minimize the number of runs necessary to find an optimum parameter set.

The form of the response surface may also become more and more complex as the number of parameters increases, and it is also more and more difficult to visualize the response surface in three or more parameter dimensions. Some of the problems likely to be encountered, however, can be illustrated with simple two-parameter example. The form of the response surface is not always the type of simple hill shown in Figure 1. If it was, then finding an optimum parameter set would not be difficult; any of the so-called hill-climbing automatic optimization techniques should do a good job in finding the way from any arbitrary starting point to the optimum.

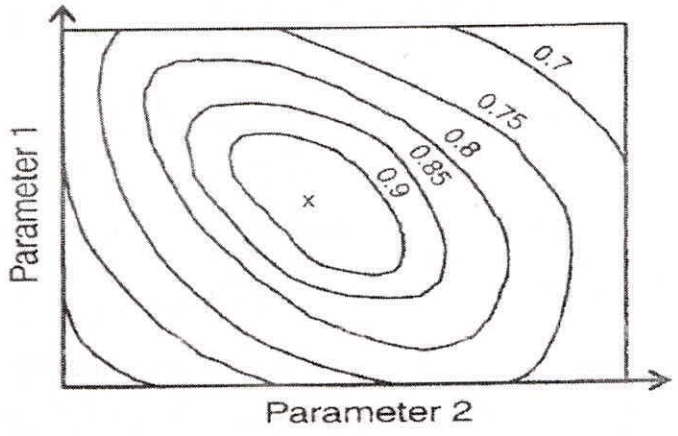

Figure 2 Response surface for two parameter dimensions with goodness of fit represented as contours

One of the problems commonly encountered is parameter insensitivity. This will occur if a parameter has very little effect on the model result in part of the range. This may result from the component of the model associated with that parameter not being activated during a run (perhaps the parameter is the maximum capacity of a store in the model and the store never gets filled). In this case part of the parameter response space will be 'flat' with respect to changes in one or more parameters (e.g. Parameter 1 in (Figure 3(a)). Changes in that parameter in that area have very little effect on the results. Hill-climbing techniques may find it difficult to find a way off the plateau and towards higher goodness of fit functions if they get onto such a plateau in the response surface. Different starting points may then lead to different final sets of parameter values.

Another problem is parameter interactions. This can lead to multiple optima (Figure 3(b)) or 'ridges' in the response surface (Figure 3(c)), with different pairs of parameter values giving a very similar goodness of fit. In these latter cases a hill-climbing technique may find the ridge very easily but may find it difficult to converge on a single set of values giving the best fit. Again, different starting values may give different final sets of parameter values.

The problem of multiple local optima can make hill-climbing optimization particularly difficult. One of these local peaks will be the global optimum, but there may be a number of local optima that give a similar goodness of fit. The response surface may also be very irregular or jagged for a good two-parameter example. Again, different starting points for a hill-climbing algorithm might lead to very different final values. Most such algorithms will find the nearest local optimum, which may not be the global optimum.

This is not just an example of mathematical complexity; there may be good physical reasons why this might be so. If a model has components for infiltration excess runoff production, saturation excess runoff production or subsurface stormflow (we might expect more than two parameters in this case), then there will likely be sets of parameters

that give a good fit to the hydrograph using the infiltration excess mechanism; sets giving good fits using a saturation excess mechanism; sets giving good fits by a subsurface stormflow mechanism; and even more sets giving good fits by a mixture of all three processes. The different local optima may then be in very different parts of the parameter space.

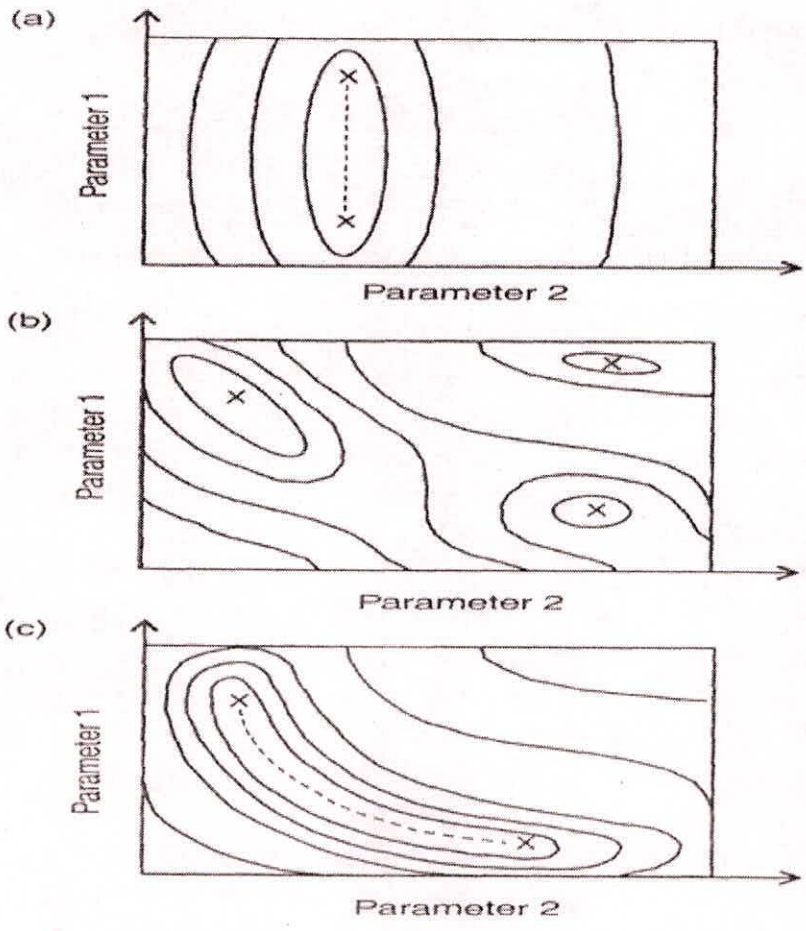

Figure 3 More complex response surfaces in two parameter dimensions. (a) Flat areas of the response surface revealing insensitivity of fit to variations in parameter values. (b) Multiple peaks in the response surface indicating multiple local optima. (c) Ridges in the response surface revealing parameter interactions

The types of behavior shown in Figure 3 can make finding the global optimum difficult, to say the least. Most parameter optimization problems involve more than two parameters. To get an impression of the difficulties faced, try to imagine what a number of local optima would look like on a three-parameter response surface; then on a four – parameter response surface, and so on. Some advances have been made in computer visualization of higher dimensional response surfaces but trying to picture such a surface soon becomes rather taxing for bears of very little brain (or even expert hydrological the

changing gradients for the simple cases in Figure 3). Because of this, sensitivities are normally evaluated in the immediate region of a best estimate parameter set or an identified optimum parameter set after a model calibration exercise.

This is, however, a very local estimate of sensitivity in the parameter space. A more global estimate might give a more generally useful estimate of the importance of a parameter within the model structure. There are a number of global sensitivity analysis techniques available, but one that makes minimal assumptions about the shapes of the response surface is variously known as generalized sensitivity analysis (GSA), regionalized sensitivity analysis (RSA) or the Hornberger – Spear – Yong (HSY) method (Beven 2001). The HSY method is based on Monte Carlo simulation. Monte Carlo simulation makes use of many different runs of a model, with each run using a randomly chosen parameter set. In the HSY method the parameter values are chosen from uniform distributions spanning specified ranges of each parameter. The ranges should reflect the feasible parameter values in a particular application. The idea is to obtain a sample of model simulations from throughout the feasible parameter space. Those simulations are classified in some way into those that are considered behavioural and those that are considered non-behavioural in respect of the system being studied. Behavioural simulations might be those with a high value of a certain variable or performance measure; non-behavioural simulations might be those with a low value. The HSY approach is essentially a nonparametric method of sensitivity analysis in that it makes no prior assumptions about the various or covariation of different parameter values, but only evaluates sets of parameter values in terms of their performance.

PERFORMANCE MEASURES AND LIKELIHOOD MEASURES

The definition of a parameter response surface as outlined above and shown in Figures 2 and 3 requires a quantitative measure of performance or goodness of fit. It is not too difficult to define the requirements of a rainfall-runoff model in words: we want a model to predict the hydrograph peaks correctly (at least to within the magnitude of the errors associated with the observations), to predict the timing of the hydrograph peaks correctly, and to give a good representation of the form of the recession curve to set up the initial conditions prior to the next event. We may also require that, over a long simulation period, the relative magnitudes of the different elements of the water balance should be predicted accurately. The requirements might be somewhat different for different projects, so there may not be any universal measure of performance that will serve all purposes.

Most measures of goodness of fit used in hydrograph simulation in the past have been based on the sum of squared errors, or error variance. Taking the squares of the residuals results in a positive contribution of both overpredictions and underpredictions to the final sum over all the time steps. The error variance, σ_ϵ^2, is defined as

$$\sigma_\epsilon^2 = \frac{1}{T-1}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2 \qquad (1)$$

where $\hat{y}_t$ is the predicted value of variable y at time step t = 1.2 ..... T. Usually the predicted variable is discharge, Q (as shown in Figure 7.4), but it may be possible to evaluate the model performance with respect to other predicted variables so we will use the general variable y in what follows. A widely used goodness of fit measure based on the error variance is the modeling efficiency of Nash and Sutcliffe (1970), defined as

$$\mathrm{E} = E = \left[1 - \frac{\sigma_\epsilon^2}{\sigma_o^2}\right] \tag{2}$$

where σ_o^2 is the variance of the observations. The efficiency is like a statistical coefficient of determination. It has the value of 1 for a perfect fit when $\sigma_\epsilon^2 = 0$; it has the value of 0 when $\sigma_\epsilon^2 = \sigma_o^2$ which is equivalent to saying that the hydrological model is no better than a one-parameter 'no-knowledge' model that gives a prediction of the mean of the observations for all time steps. Negative values of efficiency are indicating that the model is performing worse than this 'no-knowledge' model.

The sum of squared errors and modeling efficiency are not ideal measures of goodness of fit for rainfall-runoff modeling for three main reasons. The first is that the largest residuals will tend to be found near the hydrograph peaks. Since the errors are squared this can result in the predictions of peak discharge being given greater weight than the prediction of low flows (although this may clearly be a desirable characteristic for some flood forecasting purposes). Secondly, even if the peak magnitudes were to be predicted perfectly, this measure may be sensitive to timing errors in the predictions. This is illustrated for the second hydrograph in Figure 4 which is well predicted in shape and peak magnitude but the slight difference in time results in significant residuals on both rising and falling limbs.

Figure 4 also illustrates the third effect, i.e. that the residuals at successive time steps may not be independent but may be autocorrelated in time. The use of the simple sum of squared errors as a goodness of fit measure has a strong theoretical basis in statistical inference, but for cases where the samples (here the predictions at each time step) can be considered as independent and of constant variance. In many hydrograph simulations there is also a suggestion that the variance of the residuals may change in a consistent way over time, with a tendency to be higher for higher flows. This has led to the use of measures borrowed from the theory of maximum likelihood in statistics, which attempt to take account of the correlation and changing variance of the errors (heteroscedastic errors).

Maximum likelihood aims to maximize the probability of predicting an observation, given the model. These probabilities are specified on the basis of a likelihood function, which is a goodness of fit measure that has the advantage that it can be interpreted directly in terms of such prediction probabilities. However, the likelihood function that is appropriate will depend on defining an appropriate structure for the modeling errors.

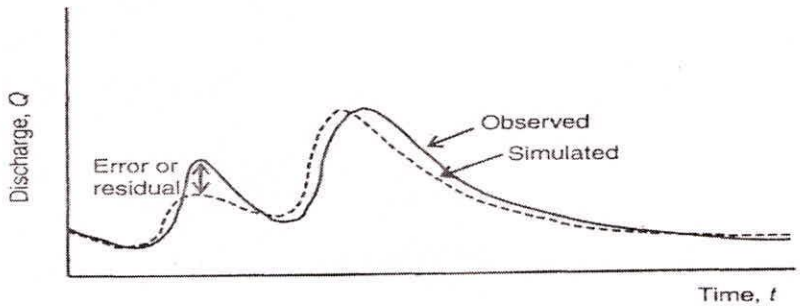

Fig. 4 Comparing observed and simulated hydrographs

Underlying the development of the likelihood functions used in maximum likelihood approaches is the idea that there is a correct model, focusing attention on the nature of the errors associated with that model. Ideally, we would hope to find a model with zero bias, and purely random errors with minimum variance and no autocorrelation. For the relatively simple case of an additive error with a Gaussian distribution and single time step autocorrelation, the likelihood function is easy to develop. More complex error models will result in more complex likelihood functions. In principle, the structure of the errors should be checked to ensure that an appropriate error model is being used. In practice, this must be an iterative process since, under the assumption that there is a correct model, it is the structure of the errors of that optimum model that must be checked, but finding the optimum depends on defining a likelihood function for an error structure.

Experience suggests that hydrological models do not, in general, conform well to the requirements of the classical techniques of statistical inference and that a more flexible and application oriented approach to model calibration is required. There are certainly many other performance measures that could be used. It may also be necessary to combine goodness of fit measures for more than one variable, e.g. both discharge and one or more predictions of observed water table level. Again, a number of different ways of combining information are available. Some of the more interesting recent developments are based on a set theoretical approach to model calibration.

All these measures are aimed at providing a relative measure of model performance. That measure should reflect that aims of a particular application in an appropriate way. There is no universal performance measure and whatever choice is made, there will be an effect on the relative goodness of fit estimates for different models and parameter sets, particularly if an optimum parameter set is sought.

CALIBRATION AND VALIDATION OF DISTRIBUTED MODELS

Validation of distributed models has received a great deal of attention. Some of the discussion has, in fact, suggested that validation is not an appropriate term to use in this context, since no model approximation can be expected to be a valid representation of a complex reality. Model evaluation has been suggested as a better term. Because

distributed models make distributed predictions, there is a lot of potential for evaluating not only the predictions of discharge at a catchment outlet, but also the internal state variables such as water table levels, soil moisture levels, channel flows at different points on the network, etc. It appears that there have still been relatively few studies of distributed models that have attempted such an elevation. Most models are evaluated only on the basis of predicted discharge, which leaves plenty of scope for the runoff to be simulated by a variety of different mechanisms.

The lack of evaluation with respect to internal state variables is clearly partly due to the expense of collecting widespread measurements of such internal state variables. There are also some difficulties in measuring quantities that can truly be compared with model predictions since the scale of the measurements may be significantly different from the model element scale at which the predictions of the model are made. In addition, even in the very first application of this type of model to a field site, because of uncertainties in the boundary conditions, initial conditions and parameter values of a distributed model, it is unlikely that a true model validation will ever be possible since the errors in representing the system and specifying the inputs will surely induce unavoidable errors in the simulations, however well a model appears to have been calibrated.

There is a further interesting interaction with the problem of model calibration for the case of distributed models. Suppose that the parameters of a distributed model have initially been calibrated only on the basis of prior information about soil and vegetation type, with some adjustment of values being made to improve the simulation of measured discharges. The model might well do a very good job of simulating the catchment discharge but we would have little idea of how well it was doing in predicting the internal state variables such as water table levels. In fact, because of the lack of information about the internal responses of the catchment, the model user would probably use effective values of model parameters such as hydraulic conductivity over wide regions of the flow domain.

Assume that after this initial calibration, a decision is made to collect more spatially distributed information about the catchment response. Measurements might be made of water table heights and soil moisture storage, and some internal stream gauging sites might be installed. We would expect that the predictions of the calibrated distributed model turn out to be wrong in many places, since the calibration has taken little account of local heterogeneities in the catchment characteristics (other than the broad classification of soil and vegetation types). There is now the potential to use the new internal measurements not to evaluate the model, but to improve the local calibration, a process that will not necessarily improve the prediction of catchment discharge which was the subject of the original calibration. It will generally mean a much greater improvement in the prediction of the internal state variables for which measurements have now been available. But if the new data are being used to improve the local calibration, more data will be needed for model evaluation.

REFERENCES

Beven, K.J. (2001). Rainfall-runoff Modelling: The Primer. Wiley, New York.

Melching, C.S. (1995). Reliability Estimation, in Computer Models of Watershed Hydrology,. Edited by V P Singh, Water Resources Publications, Colorado, USA.

Singh, V.P. (1995). Watershed Hydrology, in Computer Models of Watershed Hydrology,. Edited by V P Singh, Water Resources Publications, Colorado, USA.

Sorooshian, S., and V.K. Gupta (1995). Model Calibration, in Computer Models of Watershed Hydrology,. Edited by V P Singh, Water Resources Publications, Colorado, USA.

****