zone at each time step in formulating the new index. The MACAQUE model of Watson *et al.* (1999) addresses the same problem in a different way by introducing a 'lateral redistribution factor' that limits the redistribution towards the steady-state water table configuration allowed at each time step.

One of the limitations of the topographic index approach is the assumption that there is always downslope flow from a upslope contributing area that is constant for any point in the catchment. Improved predictions might be possible if this area was allowed to vary dynamically. Barling *et al.* (1994) showed that an index based on travel times could improve prediction of saturated areas for a single time step prediction, but did not suggest how this might be extended to a continuous time model. A dynamic TOPMODEL can also be derived by an explicit redistribution of downslope fluxes from one group of hydrologically similar points to another, where the definition of hydrologically similar can be based on more flexible criteria than the original topographic index. In the extreme case of every pixel in a catchment being considered separately, this approach would be similar to the distributed kinematic wave model of Wigmosta *et al.* (1994). Grouping of similar pixels results in computational efficiency that might be advantageous in applications to large catchments or where large numbers of model runs are required to assess predictive uncertainty. This is the basis for a new, more dynamic, version of TOPMODEL (Beven and Freer 2000).

7

Parameter Estimation and Predictive Uncertainty

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

John W. Tukey, 1962

7.1 Parameter Estimation and Predictive Uncertainty

It should be clear from the preceding chapters that 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 (e.g. Beven *et al.* 1984; Parkin *et al.* 1996; Refsgaard and Knudsen 1996; Loague and Kyriakidis 1997); 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 discussed in Section 1.8. 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. Some examples of such measures that have been used in rainfall–runoff modelling are discussed in Section 7.3.

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. This chapter can

give only a summary of the major themes being explored, and 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 values adjusted by the user, to the variety of *automatic optimization* methods discussed in Section 7.4.
- Methods of model calibration that assume an optimum parameter set, but which make certain assumptions about the *response surface* (see Section 7.2) around that optimum to estimate the predictive uncertainty, can be found. These methods will be grouped under the name *reliability analysis* and will be discussed in Section 7.5.
- Methods of model calibration that reject the idea that there is an optimum parameter set in favour of the idea of *equifinality* of models, as discussed in Section 1.8, can be found. Equifinality is the basis of the GLUE methodology discussed in Section 7.6. In this context it is perhaps more appropriate to use model conditioning rather than model calibration since this approach attempts to take account of the many 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 for 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 (Section 7.2).
- 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 used, 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 responses for more and more extreme conditions relative to the data used in calibration.

7.2 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 (specific performance measures will be discussed in the next section). 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. This is one of the options provided in the TOPMODEL software (see Appendix A). However, how do we 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 (try it with the demonstration TOPMODEL software where up to five parameters may be changed).

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, we 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 7.1 (see also the three-dimensional representation of Figure 1.7). In this example, 10 discrete increments would require $10^2 = 100$ runs of the model. For simple models this should not take too long. For example, 100 runs of TOPMODEL with 1000 time steps on a Pentium PC will take about 2 minutes, although complex fully distributed models will take much longer. 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

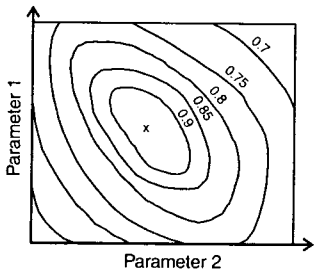

Figure 7.1 *Response surface for two parameter dimensions with goodness of fit represented as contours*

computing for TOPMODEL 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 our simple two-parameter example. The form of the response surface is not always the type of simple hill shown in Figure 1.7. If it was, then finding an optimum parameter set would not be difficult; any of the so-called hill-climbing automatic optimization techniques of Section 7.4 should do a good job in finding the way from any arbitrary starting point to the optimum.

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 7.2(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 7.2(b)) or 'ridges' in the response surface (Figure 7.2(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 (see Blackie and Eeles (1985) for a good two-parameter example and also the discussion in Sorooshian and Gupta (1995)). 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 (see Beven and Kirkby (1979) for an example using the original TOPMODEL). The different local optima may then be in very different parts of the parameter space.

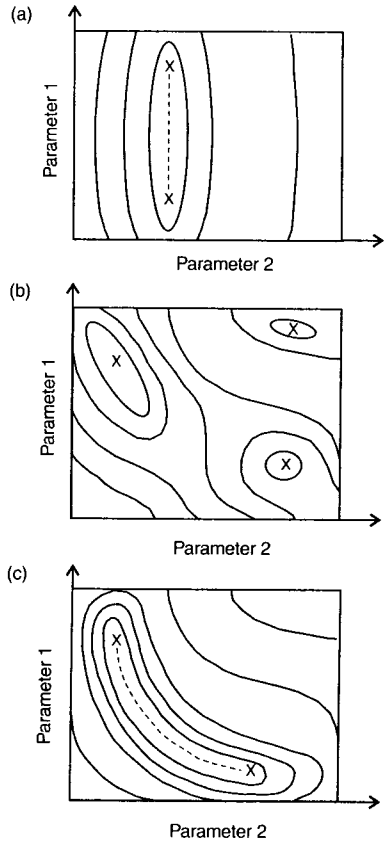

Figure 7.2 *More complex response surfaces in two parameter dimensions. (a) Flat areas of the response surface revealing insensitivity of fit to variations in parameter values. (b) Multiple peaks in the response surface indicating multiple local optima. (c) Ridges in the response surface revealing parameter interactions*

The types of behaviour shown in Figure 7.2 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

modellers). The modern 'hill-climbing' algorithms described in Section 7.4 are designed to be robust with respect to such complexities of the response surface.

There is, however, another way of approaching the problem, i.e. by designing hydrological models to avoid such calibration problems. A model could be structured, for example, to avoid the type of threshold maximum storage capacity parameter that gets activated only for a small number of time steps. Early work on this type of approach in rainfall–runoff modelling was carried out by Richard Ibbitt (see Ibbitt and O'Donnell 1971, 1974) using conceptual ESMA-type models, while, as noted in Section 6.2, the PDM model was originally formulated by Moore and Clarke (1981) with this in mind. Normally, of course, hydrological models are not designed in this way. The hydrological concepts are given priority rather than the problems of parameter calibration, particularly in physics-based models. However, for any model that is subject to calibration in this way, these considerations will be relevant.

There are particular problems in assessing the response surface and sensitivity of parameters in distributed models, not least because of the very large number of parameter values involved and the possibilities for parameter interaction in specifying distributed fields of parameters. This will remain a difficulty for the foreseeable future and the only sensible strategy in calibrating distributed models would appear to be to insist that most, if not all, of the parameters are either fixed (perhaps within some feasible range, as in Parkin *et al.* (1996)) or calibrated with respect to some distributed observations and not catchment discharge alone (such as in Franks *et al.* (1998) and Lamb *et al.* (1998b)). The special problems of calibrating distributed models were discussed in earlier Sections 5.1.1 and 5.7.

7.2.1 Assessing Parameter Sensitivity

The efficiency of parameter calibration would clearly be enhanced if it was possible to concentrate the effort on those parameters to which the model simulation results are most sensitive. This requires an approach to assessing parameter sensitivity within a complex model structure. Sensitivity can be assessed with respect to both predicted variables (such as peak discharges, discharge volume, water table levels, snowmelt rates, etc.) or with respect to some performance measure (see next section). Both can be thought of in terms of their respective response surfaces in the parameter space. One definition of the sensitivity of the model simulation results to a particular parameter is the local gradient of the response surface in the direction of the chosen parameter axis. This can be used to define a normalized sensitivity index of the following form:

$$S_i = \frac{\mathrm{d}Z/\mathrm{d}x_i}{x_i} \tag{7.1}$$

where S_i is the sensitivity index with respect to parameter i with value x_i, and Z is the value of the variable or performance measure at that point in the parameter space (see McCuen 1973). The gradient will be evaluated locally, given values of the other parameters, either analytically for simple models, or numerically by a finite difference, i.e. by evaluating the change in Z as x_i is changed by a small amount (say 1 percent). Thus, since the simulation results depend on all the parameters, the sensitivity S_i for any particular parameter i will tend to vary through the parameter space (as illustrated by the changing gradients for the simple cases in Figure 7.2). 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–Young (HSY) method (see Hornberger and Spear 1981; Young 1983; Beck 1987), which was a precursor of the GLUE methodology described in Section 7.6. 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.

HSY sensitivity analysis then looks for differences between the behavioural and non-behavioural sets for each parameter. It does so by comparing the cumulative distribution of that parameter in each set (e.g. Figure 7.3). Where there is a strong difference between the two distributions for a parameter, it may be concluded that the simulations are sensitive to that parameter (Figure 7.3(b)). Where the two distributions are very similar, it may be concluded that the simulations are not very sensitive to that parameter (Figure 7.3(c)). A quantitative measure of the difference between the distributions can be calculated using the nonparametric Kolmagorov–Smirnov d statistic, although for large numbers of simulations this test is not robust and will suggest that small differences are statistically significant. The d statistic can, however, be used as an index of relative difference. This approach may be extended, given enough Monte Carlo simulation samples, to more than two sets of parameters (the GLUE software, for example, uses 10 different classes in assessing sensitivity). Other examples of the use of the HSY approach in rainfall–runoff modelling include Hornberger *et al.* (1985) using TOPMODEL, and Harlin and Kung (1992) using the HBV model. The HSY approach is essentially a *nonparametric method* of sensitivity analysis in that it makes no prior assumptions about the variation or covariation of different parameter values, but only evaluates sets of parameter values in terms of their performance.

7.3 Performance Measures and Likelihood Measures

The definition of a parameter response surface as outlined above and shown in Figures 7.1 and 7.2 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

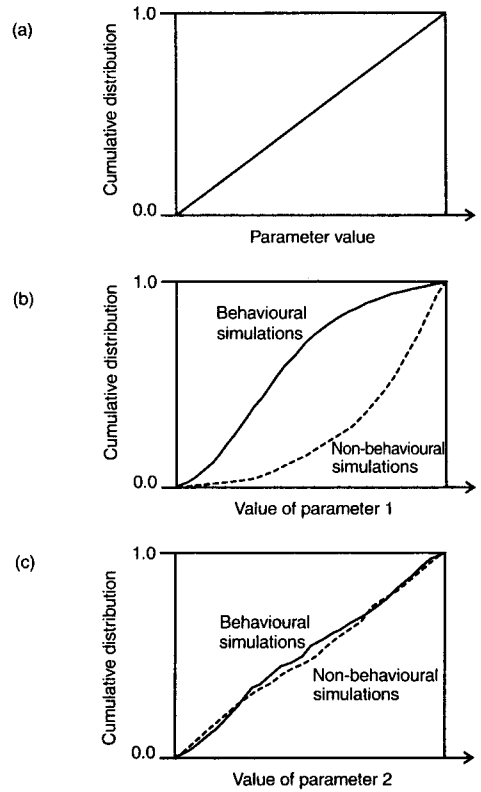

Figure 7.3 *Generalized (Hornberger–Spear–Young) sensitivity analysis. (a) Initial cumulative distributions of parameter values for uniform sampling of prior parameter values across a specified range. (b) Cumulative distributions of parameter values for behavioural and non-behavioural simulations for a sensitive parameter. (c) Cumulative distributions of parameter values for behavioural and non-behavioural simulations for an insensitive parameter*

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

$$\sigma_\varepsilon^2 = \frac{1}{T-1}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2 \tag{7.2}$$

where $\hat{y}_t$ is the predicted value of variable y at time step $t = 1, 2, \ldots, T$. 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 modelling efficiency of Nash and Sutcliffe (1970), defined as

$$E = \left[1 - \frac{\sigma_\varepsilon^2}{\sigma_o^2}\right] \tag{7.3}$$

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 is $\sigma_\varepsilon^2 = 0$; it has the value of 0 when $\sigma_\varepsilon^2 = \sigma_o^2$, 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 modelling efficiency are not ideal measures of goodness of fit for rainfall–runoff modelling 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 7.4 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 7.4 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

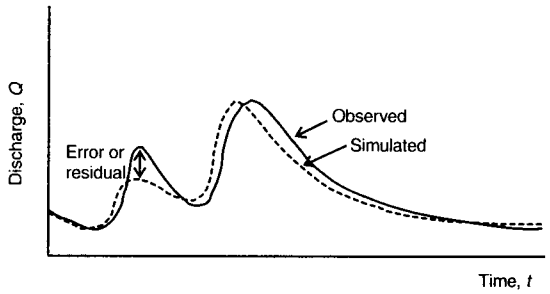

Figure 7.4 *Comparing observed and simulated hydrographs*

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*, e.g. Sorooshian *et al.* 1983; Hornberger *et al.* 1985). Maximum likelihood aims to maximize the probabilty 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 modelling errors.

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 developed in Box 7.1. More complex error models will result in more complex likelihood functions (e.g. Cox and Hinkley 1974). 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. Some examples, for the prediction of single variables such as discharges in hydrograph simulation, are given in Box 7.1. 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, and some examples are given in Box 7.2. Some of the more interesting recent developments are based on a set theoretical approach to model calibration (see Section 7.6 below).

Remember that all these measures are aimed at providing a relative measure of model performance. That measure should reflect the 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. The next section will examine the techniques for finding optimal parameter sets, after which a more flexible approach to model calibration will be discussed.

7.4 Automatic Optimization Techniques

A full description of all the available techniques for automatic optimization is well beyond the scope of this book, particularly since we have already noted that the concept

of the optimum parameter set may not be a particularly useful one in hydrological modelling. In this section, we will give just a brief outline of the algorithms that are available. For more specifics, descriptions of different algorithms are available in Press *et al.* (1992) and Sen and Stoffa (1995), and a discussion of techniques in respect of hydrological models is given in Sorooshian and Gupta (1995).

7.4.1 Hill-Climbing Techniques

Hill-climbing techniques for parameter calibration have been an important area of research since the start of computer modelling in the 1960s. Hill climbing from any point on the response surface requires knowledge of the gradient of that surface so that the algorithm knows in which direction to climb. The available techniques may be classified into two basic types. The first are algorithms that require the gradient of the response surface to be defined analytically for every point in the parameter space. Mathematically, this requires that an analytical expression be available for the differential of the model output with respect to each parameter value. These gradient methods are not generally used with hydrological models since it is often impossible to define such differentials analytically for complex model structures. Much more commonly used are direct search algorithms, which search along trial directions from the current point with the aim of finding improved objective function values. Different algorithms vary in the search strategies used. Algorithms that have been widely used in rainfall–runoff modelling include the Rosenbrock method (Rosenbrock 1960) and the Simplex method (Nelder and Mead 1965). The latter is explained in Sorooshian and Gupta (1995).

Hill climbing is, of course, much easier on smooth response surfaces than on flat or jagged surfaces. Many hydrological models will not give smooth response surfaces but, as noted above, with three or more parameter values it may be difficult to evaluate or visualize the full shape of the surface. If a hill-climbing technique is used for parameter calibration, a *minimal* check on the performance of the algorithm in finding a global optimum is to start the algorithm from a number of very different (or randomly chosen) starting points in the parameter space and check the consistency of the final sets of parameter values found. If the final sets are close, then it may be implied that there is a single optimum. If not, then consider one of the algorithms in the sections that follow, which have all been developed to be robust with respect to complexities in the response surface.

7.4.2 Simulated Annealing

Another way of using random starting points to find a global optimum is simulated annealing. The name arises from an analogy between the model parameters included in the optimization and particles in a cooling liquid, which is the basis of the algorithm. If the particles are initially all in the liquid state they will be randomly distributed through the space occupied by the fluid. As the liquid is cooled to a lower temperature, annealing will take place in a way that minimizes the energy of the system. If the cooling is too fast, this energy minimization will occur locally; if very slow then eventually a global minimum energy state will result. The idea of simulated annealing is to mimic this cooling process, starting from randomly distributed sets of parameters in the parameter

space to find a global optimum state with respect to the performance measure of the optimization problem.

There are a number of different variants on simulated annealing including very fast simulated reannealing and mean field annealing (see Tarantola 1987; Ingber 1993; Sen and Stoffa 1995). The essence of all of the methods is a rule for the acceptance of new parameter sets. Given a starting parameter set, a perturbation of one or more parameter values is generated and the new performance measure is calculated. If it is better than the previous one, the new model is accepted. If it is not better, it may still be accepted with a probability based on an exponential function of the difference in the performance measure value scaled by a factor that is equivalent to the temperature in the annealing analogy. As the temperature is gradually reduced over a number of iterations, this probability is reduced. This way of allowing parameter sets with worse performance to be accepted ensures that the algorithm does not get trapped by a local optimum, at least if the rate of cooling is slow enough. The choice of the cooling schedule is therefore important and will vary from problem to problem. The various simulated annealing methods differ in the ways that they attempt to increase the number of accepted models relative to those rejected and therefore increase the efficiency of the search. In hydrology, a recent application of simulated annealing may be found in Thyer *et al.* (1999).

There are similarities between simulated annealing and some of the Monte Carlo Markov Chain (MC^2) methods for parameter estimation that have seen a rapid recent development in statistics. Sen and Stoffa (1995) note that the Metropolis MC^2 algorithm is directly analogous to a simulated annealing method. This has been used in rainfall–runoff model parameter estimation by Kuczera and Parent (1998) and Overney (1998).

7.4.3 Genetic Algorithms

Genetic algorithm (GA) methods are another way of trying to ensure that a global optimum is always found, but are based on a very different analogy, that of biological evolution. A random population of 'individuals' (different parameter sets) is chosen as a starting point and then allowed to 'evolve' over successive generations or iterations in a way that improves the 'fitness' (performance measure) at each iteration until a global optimum fitness is reached. The algorithms differ in the operations used to evolve the population at each iteration, which include selection, cross-over and mutation. A popular description has been given by Forrest (1993), and more detailed descriptions are given by Davis (1991). Sen and Soffa (1995) show how some elements of simulated annealing can be included in a genetic algorithm approach. GA optimization has been used by Wang (1991) in calibrating the Xinanjiang model, Kuczera (1997) with a five-parameter conceptual rainfall–runoff model and Franchini and Galeati (1997) with an 11-parameter model.

One form of algorithm that has been developed for use in rainfall–runoff modelling, and which combines hill-climbing techniques with GA ideas, is the shuffled complex evolution (SCE) algorithm developed by Duan *et al.* (1992). In this algorithm, different simplex searches are carried out in parallel from each random starting point. After each iteration of the multiple searches, the current parameter values are shuffled to form new simplexes which then form new starting points for a further search iteration.

This shuffling allows global information about the response surface to be shared and means that the algorithm will generally be robust to the presence of multiple local optima. Kuczera (1997) concluded that the SCE algorithm was more successful in finding the global optimum in a five-parameter space than a classical crossover GA algorithm.

7.5 Recognizing Uncertainty in Models and Data: Reliability Analysis

The techniques of the last section are designed to find an optimum parameter set as efficiently as possible. A run of the model using that optimum parameter set will give the best fit to the observations used for the calibration, *as defined by the performance measure used*. It has long been recognized that different performance measures will generally result in different optimum parameter sets. Thus, as far as is possible, the performance measure should reflect the purpose of the modelling. The optimum parameter set alone, however, will reveal little about the possible uncertainty associated with the model predictions. There are many causes of uncertainty in a modelling study. Errors in initial and boundary conditions, errors in the calibration data and errors in the model itself, will all tend to induce uncertainty in the model predictions that should be assessed. As noted earlier, a review of sources of uncertainty in rainfall–runoff modelling and methods for uncertainty estimation is provided by Melching (1995). He includes methods based on Monte Carlo simulation; Latin Hypercube simulation; mean-value first-order second-moment estimation (MFOSM); an advanced first-order second-moment (AFOSM) method; Rosenblueth's point estimation method; and Harr's point estimation method. These are essentially all ways of sampling the response surface for the performance measure in the parameter space. Where enough runs of the model can be made, the Monte Carlo simulation technique will generally produce the most accurate results; the others are approximations to save computer time. However, a high dimensional parameter space will require many, many Monte Carlo samples, as explained in Section 7.2 above, so that the approximate methods still have value in practical applications.

The aim of uncertainty estimation is to assess the probability of a certain quantity, such as the peak discharge of an event, being within a certain interval but it is worth noting that different types of interval might be required. Haan and Meeker (1991), for example, distinguish three different types of interval. A confidence interval will contain the estimate of an unknown characteristic of the quantity of interest, e.g. the mean peak discharge of the event. Since we cannot estimate the peak discharge precisely from the sample of model runs available, then even the estimate of the mean will be uncertain. The confidence interval can then be used to define the mean estimate with specified probability. Most often, 5 and 95 percent limits are used to define a confidence interval (i.e. a 90 percent probability that the value lies within the interval). Confidence limits can also be calculated for other summary quantities for the distribution of peak discharge, such as the variance or even a quantile value.

A second type of interval is the tolerance interval. This is defined so as to contain a certain proportion of the uncertain model estimates of an observation used in model calibration. For the peak discharge example, tolerance intervals could be defined for

the model predictions of a particular observed peak used in model calibration. Finally, the third type of interval is the prediction interval. In the rainfall–runoff context this could be defined as the interval containing a certain proportion of the uncertain model estimates of peak discharge (or any other predicted variable) for a future event. In rainfall–runoff modelling we are mostly interested in prediction intervals after calibration or conditioning of a model.

Uncertainty limits are related to the changes in the predicted variable in the parameter space or, more precisely, if a predicted variable (rather than the performance measure) is represented as a surface in the parameter space, to the *gradient* or slope of the surface with respect to changes in the different parameter values. If the slope is steep, then methods such as MFOSM will predict that the uncertainty in the predictions will be large. If the slope is quite small, however, then the methods will predict a small uncertainty since the predicted variable will change little if the parameter is considered to be uncertain. Recalling equation (7.1), the slopes are an indication of the local sensitivity of the predictions to errors in the estimation of the parameter values.

One question that then arises is where to calculate the values of the slopes in order to get a good estimate of the uncertainty limits. This is where the approximate methods must make certain assumptions. The classical assumption is to assume that the response surface is locally multivariate normal around the prediction of the optimum parameter set. The variance of the estimate of a variable Q will then be given by:

$$\mathrm{Var}(Q) = \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial Q}{\partial x_i} \frac{\partial Q}{\partial x_j} E\left[(x_i - \hat{x}_i)(x_j - \hat{x}_j)\right] \tag{7.4}$$

where the slopes (the differential terms) are evaluated close to the optimum, $E[\cdot]$ represents an expected value, the x values are the parameter values, and the $\hat{x}$ values are the optimum parameter set. The term $E\left[(x_i - \hat{x}_i)(x_j - \hat{x}_j)\right]$ reflects the covariation of the parameters. If the parameters can be considered to be statistically independent, then

$$\mathrm{Var}(Q) = \sum_{i=1}^{p} \left[\frac{\partial Q}{\partial x_i} \sigma_i\right]^2 \tag{7.5}$$

where σ_i is an estimate of the variance of parameter x_i.

If the model response is linear then this may be an adequate approximation, but many rainfall–runoff models are highly nonlinear. Thus linearization around the optimum will not then provide accurate estimates of uncertainty in the predictions. Melching (1995) notes that this may be a particular problem in reliability analysis of engineering designs where the interest is often in the risk of a particular design failing under extreme conditions (such as a reservoir overflow channel or flood protection scheme in the case of rainfall–runoff modelling). Uncertainty then becomes important in evaluating risk, but for a nonlinear model it will be important to investigate the behaviour away from the best estimate or model response in the region of the more extreme responses. This is the purpose, for example, of the AFOSM method, which still uses linearization but does so around an estimate of a failure point or confidence limit rather than around the mean prediction. More details and references may be found in the Melching (1995) review.

7.6 Model Calibration Using Set Theoretic Methods

There is another approach to model calibration that relies much less on the idea that there is an optimum parameter set. It was noted in Section 1.8 that detailed examination of response surfaces reveals many different combinations of parameter values that give good fits to the data, even for relatively simple models. The concept of the optimum parameter set may then be ill-founded in hydrological modelling, carried over from concepts of statistical inference. A basic foundation of the theory of statistical inference is that there is a correct model; the problem is to estimate the parameters of that model given some uncertainty in the data available. In hydrology it is much more difficult to make such an assumption. There is no correct model, and the data available to evaluate different models may have large uncertainties associated with them, especially for extreme events, which are often those of greatest interest.

An alternative approach to model calibration is to try to determine a set of acceptable models. Set theoretic methods of calibration are generally based on Monte Carlo simulation. A large number of runs of the model are made with different randomly chosen parameter sets. Those that meet some performance criterion or criteria are retained; those that do not are rejected. The result is a set of acceptable models, rather than a single optimum model. Using all the acceptable models for prediction results in a range of predictions for each variable of interest, allowing an estimation of prediction intervals. This type of method has not been used widely in rainfall–runoff modelling (with the exception of the GLUE variant of the next section) but there have been a number of studies in water quality modelling (e.g. Klepper *et al.* 1991; Rose *et al.* 1991; van Straten and Keesman 1991).

A recent development in set theoretic approaches has been the multi-criteria calibration strategy of Yapo *et al.* (1998) and Gupta *et al.* (1998). Their approach is based on the concept of the Pareto Optimal Set, a set of models with different parameter sets that all have values of the various performance criteria that are not inferior to any models outside the optimal set on any of the multiple criteria. In the terminology of the method, the models in the optimal set are dominant over those outside the set. Yapo *et al.* (1998) have produced an interesting method to define the Pareto Optimal Set, related to the SCE optimization of Section 7.3. Rather than a pure Monte Carlo experiment, they start with N randomly chosen points in the parameter space and then use a search technique to modify the parameter values and find N sets within the optimal set (Figure 7.5). They suggest that this will be a much more efficient means of defining the Pareto Optimal Set.

They demonstrate the use of the model and the resulting prediction limits with the Sacramento ESMA-type rainfall–runoff model, used in the US National Weather Service River Forecasting System, in an application to the Leaf River catchment, Mississippi. The model has 13 parameters to be calibrated. Two objective functions were used in the calibration: a sum of squared errors and a heteroscedastic maximum likelihood criterion. To find the Pareto Optimal Set, 500 parameter sets were evolved, requiring 68 890 runs of the model. The results are shown in Figure 7.6, in terms of the grouping of the 500 final parameter sets on the plane of the two objective functions (from Yapo *et al.* 1998) and the associated ranges of discharges predicted by the original randomly chosen parameter sets and the final Pareto Optimal Set (from Gupta *et al.* 1998). A major advantage of the Pareto Optimal Set methodology is that it does not

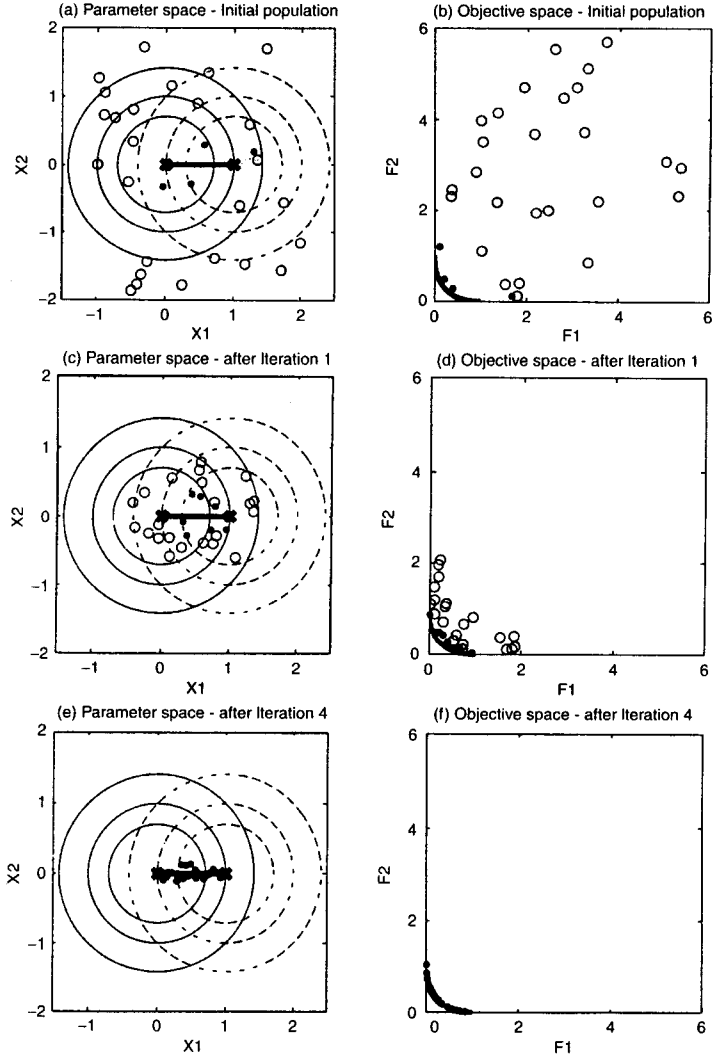

Figure 7.5 *Iterative definition of the Pareto Optimal Set using a population of parameter sets initially chosen randomly. (a) Initial parameter sets in a two-dimensional parameter space (parameters X1, X2). (b) Initial parameter sets in a two-dimensional objective function space (functions F1, F2) (c),(d) Grouping of parameter sets after one iteration. (e),(f) Grouping of parameter sets after iteration 4. After the final iteration, no model with parameter values outside the Pareto Optimal Set has higher values of the objective functions than the models in the Pareto Set (after Yapo et al. 1998). Reprinted from Journal of Hydrology 204: 83–97, copyright (1998), with permission from Elsevier Science*

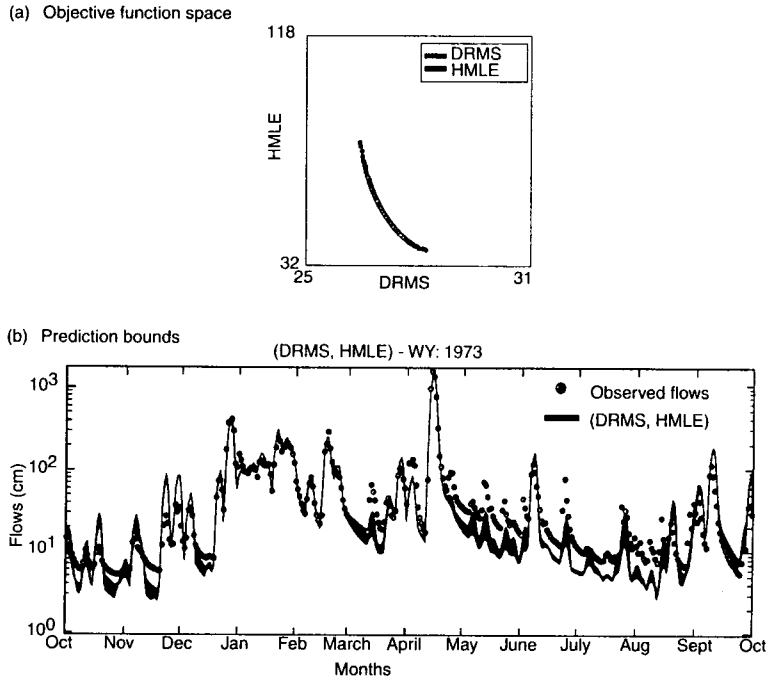

Figure 7.6 *Pareto Optimal Set calibration of the Sacramento ESMA rainfall–runoff model to the Leaf River catchment, Mississippi (after Yapo et al. 1998). (a) Grouping of Pareto Optimal Set of 500 model parameter sets in the plane of two of the model parameters. (b) Prediction limits for the 500 Pareto Optimal parameter sets. Reprinted from Journal of Hydrology 204: 83–97, copyright (1998), with permission from Elsevier Science*

require different performance measures to be combined into one overall measure. Gupta *et al.* (1999) suggest that this method is now competitive with interactive methods carried out by a modelling expert in achieving a calibration that satisfies the competing requirements on the model in fitting the data.

As shown in Figure 7.6(a), the set of models that is found to be Pareto Optimal will reflect the sometimes conflicting requirements of satisfying more than one performance measure. Figure 7.6(b), however, shows that this does not guarantee that the predictions from the sample of Pareto optimal models will bracket the observations since it cannot compensate completely for model structural error or discharge observations that are non-error-free. The original randomly chosen sets do bracket the observations, but with limits that are considerably wider (note the log discharge scale in Figure 7.6(b)). It must be remembered that the method is not intended to estimate prediction limits in any statistical sense, but one feature of this approach is that it does seem to result in an over-constrained set of predictions in comparison with the observations. The user would then have difficulty in relating the range of predictions to any degree of confidence that they would match any particular observation.