Hydrology and Earth System Sciences, 3(4), *491–503* (1999) © EGS

Hydrology & Earth System Sciences

A Bayesian approach to flow record infilling and extension for reservoir design

Jones, D.A. and Sene, K.J.

Institute of Hydrology, Wallingford, Oxfordshire OX10 8BB, United Kingdom
email address of corresponding author: daj@mail.nwl.ac.uk

Abstract

A Bayesian approach is described for dealing with the problem of infilling and generating stochastic flow sequences using rainfall data to guide the flow generation process, and including bounded (censored) observed flow and rainfall data to provide additional information. Solutions are obtained using a Gibbs sampling procedure. Particular problems discussed include developing new procedures for fitting transformations when bounded values are available, coping with additional information in the form of values, or bounds, for totals of flows across several sites, and developing relationships between annual flow and rainfall data. Examples are shown of both infilled values of unknown past river flows, with assessment of uncertainty, and realisations of flows representative of what might occur in the future. Several procedures for validating the model output are described and the central estimates of flows, taken as a surrogate for historical observed flows, are compared with long term regional flow and rainfall data.

Introduction

Stochastic flow generation models are widely used to generate synthetic flow sequences for use in reservoir design. The use of synthetic data allows a wider range of possible future flow scenarios to be investigated than that provided by the observed values and permits a formal assessment of the likely yield and reliability of the scheme under consideration. Typically, for a multi-site reservoir scheme, a multivariate Normal model is developed for the transformed annual or monthly flows which preserves selected key statistical characteristics of the observed flows (Lawrance and Kottegoda, 1977; McMahon and Mein, 1986; Bras and Rodriguez-Iturbe, 1985; Basson *et al.*, 1994). For reservoir design, these usually include the mean, variance, correlation structure and the accumulated volumes during critical, or drought, periods.

Two practical problems which often arise when applying this type of model are that the observed flow records to which the model is to be fitted may contain many gaps, and the records may only cover a short period. The difficulty with short record lengths is that the sample statistics against which the model is to be tested will themselves be subject to uncertainties which should ideally be accounted for in the generation procedure (but usually are not). Also, the periods for which flow data are available may not be representative of the long term regime in the region under consideration; for example, the records may only contain a limited number of flood and drought events, requiring assumptions to be made about the marginal distribution of flows. Similarly, any long term trends or persistence in river flows may not be apparent in short term records. The difficulty with gaps in the data is that most existing methods for flow generation require complete records for use in the calibration procedure. This means that any gaps must first be infilled, typically using physically-based rainfall-runoff models or by univariate or multivariate regression against other flow or rainfall records. Although this may be legitimate for a few short periods of missing data, if a considerable proportion of the flow database is missing, then the imputed values will assume an unjustifiably large role in the calibration procedure, with the uncertainty in these values effectively feeding through to the generated flows but possibly not being taken into account. Depending on the infilling procedure used, the apparent variation in the flow sequences may be either increased or decreased.

These problems have long been recognised in a water resources context and various solutions have been proposed. For example, Bayesian techniques provide one possible way of assessing the impact of parameter uncertainty and sampling errors (Valdes and Rodrigues-Iturbe, 1977; Wood, 1978; Stedinger *et al.*, 1985). Also, observed rainfall records—which are typically of much longer duration than observed flow records—and flow records at nearby stations can be used to bring in additional information on the long-term variability in flows (e.g. Zucchini and Hiemstra, 1983; Grygier *et al.*, 1989; Pegram, 1994). Censored data (i.e values which are only known to lie within certain bounds) can also provide additional information in some situations (Kroll and Stedinger, 1996). Taken together, these are all aspects of a more general inference problem, in which the required flows are to be estimated from the joint distribution of all the unknown values; both the model parameters and the unobserved or bounded flow and rainfall data. In the past few years, various numerical sampling procedures have become available to solve this type of problem directly, rather than through making additional analytical approximations.

One such technique (Geman and Geman, 1984) is called Gibbs sampling, and is one of a family of iterative Monte Carlo based approaches to the calculation of numerical estimates of marginal probability distributions from a complicated multidimensional probability distribution. Gibbs sampling has been widely applied in many fields (e.g. Arnold, 1990; Smith and Roberts, 1993; Besag and Green, 1993; Gilks *et al.*, 1993) but is perhaps new to hydrological applications. This paper describes an example of the application of this procedure to the estimation of river flows for reservoir design, and discusses how conventional procedures for transforming data to normality, and validating the model output, need to be adapted to cope with situations where some of the data are only available in the form of bounded (censored) values. The practical example considered is that of annual flow generation for several potential dam sites in the highlands of Lesotho in southern Africa. One of the attractions of the Bayesian approach for this application was that a single modelling procedure could be used to satisfy the two main project requirements, which were the need to infill and extend historical monthly flow records back in time by reference to observed rainfall records, and the more conventional requirement to generate completely synthetic sequences of flows for use in reservoir design. Also, there has been much discussion about the possibility of long term cycles and trend in rainfall and flow records for southern Africa (see, for example, Tyson, 1991; Sene *et al.*, 1998), so including long term rainfall records in the stochastic flow generation process helped to ensure that any such long term behaviour was automatically reflected in the resulting historical flow sequences.

The stochastic model

MODEL FORMULATION

The stochastic model was required to estimate missing historical flow and rainfall data and to generate synthetic flow sequences from the joint distribution of all the unknown values; both the model parameters and the unobserved or bounded flow and rainfall data. An early decision was taken to formulate the model in terms of annual rainfall and flow values, with disaggregation to monthly values (see later), since this allowed serial correlation in rainfall values to be neglected, although serial correlation in flows was allowed for in the model. The resulting model had the following three main components:

(a) a linear relationship predicting (transformed) flow from (transformed) rainfalls;
(b) a multivariate Normal model for the residuals from the linear relation;
(c) a multivariate Normal model for the (transformed) rainfalls.

There are well known results for Bayesian analysis of multivariate Normal distributions, specifically for the posterior distributions (e.g. Box and Tiao, 1973), which make the use of this family substantially easier than any other class. By comparison, most conventional stochastic flow generation models would use simply a single multivariate Normal model for the (transformed) observed or infilled flows.

In order to reduce the number of parameters within the model, and to include certain realistic physical assumptions about the relation between rainfall and flow (see later), the model was given the following sub-structure. Define the following quantities:

r : the vector of transformed rainfalls in a given year;
f : the vector of transformed flows in a given year;
f_{-1} : the vector of transformed flows in the previous year;
z : a vector of averaged transformed rainfall for groups of raingauge sites for the given year.

The group-averages z are defined by $z = Wr$, where W is a known matrix of weights defining both the selection of raingauges and the averaging within groups. The assumption was made that, in terms of explaining the variation of f, knowing z is as good as knowing r; that is, the residuals in a multivariate regression of f on z are uncorrelated with r. The transformations used are described later but were variants on the logarithmic transformation. The resulting stochastic relationship for flows, including serial correlation, can then be expressed as:

$$f = \mu_f + D(f_{-1}-\mu_f) + B\,(z-\mu_z) + e \qquad (1)$$

where D is a diagonal matrix of serial correlation coefficients, B is a matrix of regression parameters, μ is the mean of parameter values and e is a vector of residuals which are assumed to be uncorrelated with other random variables on the right-hand side of the equation and with r. The term μ_z is the vector given in terms of the mean rainfall by $\mu_z = W\mu_r$. To provide some control over which raingauges are used to generate flows at each site, elements of the regression matrix B may be fixed at zero so that flows are only related to rainfall at nearby raingauges or groups of gauges. This allows flows to be generated by the rainfalls recorded at only a few neighbouring sites, as specified by the structures imposed on matrices B and W but, in years where these raingauge data are incomplete, inferred values for these will be used based on values at other raingauges. This transfer of information is controlled by the covariance matrix Σ_{rr} of the transformed rainfall values r.

The parameters of the model with this assumed structure are:

B	: the regression parameters of f on z;
D	: the diagonal matrix of serial correlation coefficients;
Σ_{rr}	: the covariance matrix of r;
Σ_{ee}	: the covariance matrix of the residuals of the regression of f on z and on f_{-1};
μ_f, μ_r	: the mean vectors of f and r.

The more general model, without the structural restrictions imposed above, may be defined in terms of the mean and covariance matrix of an extended set of quantities. Specifically, let μ and Σ (without subscripts) be the mean and the covariance matrix of the vector consisting of (r, f_{-1}, f, f_{+1}) which contains

r	: rainfall in the given year
f_{-1}	: flow in the previous year
f	: flow in the given year
f_{+1}	: flow in the next year.

The parameters μ and Σ can be defined in terms of the above variables in an obvious way so details will not be given here.

CHOICE OF SOLUTION PROCEDURE

Several possible solution procedures were considered of which Gibbs sampling and the EM (Expectation-Maximization) algorithm were the two main candidates. The general principles of Gibbs sampling are now well known and are discussed in the references cited in the Introduction and in outline later in this section. A description of the more widely known EM algorithm is presented by Tanner (1993, Ch. 4), who puts what is often thought of as a non-Bayesian technique into a Bayesian context. For the present application, the difference between these approaches, and the reason for our preference for Gibbs sampling, lies in two areas. Firstly, the principal outcomes of these two approaches differ. Gibbs sampling provides a set of random outcomes which, taken together, summarise the uncertainty about all quantities described in the model, including model parameters and observed, incompletely observed (censored) and missing data-values. By contrast, the EM algorithm supplies as its principal outcome a set of 'best estimates' of the model parameters, with additional steps being required to extract information about uncertainty in the model parameters in approximate form and further steps then needed to deal with uncertainty in censored and missing data-values.

A second difference in the approaches is their complexity, particularly in a multivariate situation, as here. Gibbs sampling proceeds by finding certain conditional distributions, where the events being conditioned-on are particularly simple. The EM algorithm involves finding conditional expectations where the conditioning events, where data-values are censored, are rather more complicated: in a multivariate situation the best hope of evaluating these conditional expectations may be via random simulations. While Gibbs sampling also involves random simulations, these may be considered integral to the final outcome while, for the EM algorithm, they are principally concerned with evaluating the conditional expectations: they are essentially discarded at each step of an iterative loop and even the 'final' set are of little use in describing overall uncertainty except in a very approximate way. Given these advantages, Gibbs sampling was selected as the best solution procedure to use. This selection also took into account the need to be able to deal with the several non-standard practical considerations which arose for this hydrological application. These included the need to allow for serial correlation, to deal with bounded data-values and with null values in the regression matrices, and to ensure that incremental flow values sum to the measured total values at flow gauging sites, when available.

IMPLEMENTATION OF THE GIBBS SAMPLING PROCEDURE

In brief, the Gibbs sampling procedure can be viewed as solving an overall statistical inference problem by treating within an iterative loop a number of simpler inference problems, where each of these simple 'inference' steps is treated by generating random values for a particular set of unknown quantities from their known posterior distribution conditioned on observed data and on the current values for other unknown quantities. For descriptive purposes, it is convenient to separate the steps within the overall estimation scheme into a number of sub-problems, each of which is then further subdivided although, in the actual solution procedure, all steps are performed concurrently. Thus the overall problem can be reduced in the first instance to two rather simpler problems:

(i) inference about the parameters of the model given a set of data in which every item is assumed to be observed;
(ii) inference about missing and bounded data for fixed (*i.e.* assumed known) values of the model parameters.

The structure of the model outlined above means that it is convenient to subdivide the first of these problems into several further sub-problems each of which is of one of two relatively simple types:

i(a) inference about regression parameters, with known regression covariance matrix;

i(b) inference about a covariance matrix with known mean vector.

In particular, steps of type i(b) are performed after the corresponding steps of type i(a). Here a 'regression parameter' is, at each sub-step, either a mean vector, or a regression matrix. In the case of step (ii), it is again convenient to subdivide the problem into two parts:

ii(a) inference about unknown rainfalls conditional on a complete set of flows and the observed rainfalls;

ii(b) inference about unknown flows conditional on a complete set of rainfalls and the observed flows.

For the first of these steps, step ii(a), the model structure implies that individual years can be considered separately. The conditional mean and covariance matrix of rainfalls given the flows in the same year are readily computed from the model parameters, μ and Σ, and the problem is then reduced to infilling missing and bounded values of a multivariate Normal vector with known mean and covariance. Step ii(b) can be handled similarly, except that the dependence across years means that it is necessary to employ a further Gibbs sampling step here in that new flow values generated for one year are used as conditioning values for the next year to be treated. The required parameters are computed by finding the conditional mean and covariance matrix of flows conditioned on rainfalls in the same year and on flows in the immediately previous and succeeding years.

The remaining question of infilling values in a multivariate Normal vector is straightforward except in the case of the flow variates where the information available is in the form of bounds on totals of transformed versions of the Normal variates. In order to cope with this case the following procedure was used. First the problem is reduced by conditioning on any variates for which exact observations are available. Then a further set of Gibbs sampling steps is undertaken in which every possible pair of the remaining variates for which there is any information are replaced by new values from their joint distribution conditional on the remaining set of variables which are either exactly observed or have information about bounds. It is not possible to treat a single variate at a time because of the possibility that the information about a total is in the form of an exact value for the total, which would mean that the value for any single component of the total is fixed, given values for the others. Following this, values for any variate for which there is no information at all are generated conditional on the values for the variates already treated.

A SIMPLE EXAMPLE

These last steps are perhaps best illustrated by an example. Suppose that one is considering four variates A, B, C and D for a given year, for which the observed information about these quantities is:

$$A = a, \quad b_1 \le B \le b_2, \quad c_1 \le C \le c_2, \quad d_1 \le D \le d_2,$$
$$t_1 \le y_A(A) + y_B(B) + y_C(C) + y_D(D) \le t_2.$$

Here the functions $y_A()$, etc., denote the inverse of the transformations required for bringing the observed flows close to a Normal distribution and the final constraint derives from a censored observation of the total of the untransformed versions of A, B, C and D. Results from earlier steps within the overall Gibbs sampling procedure provide the mean vector and covariance matrix for the joint multivariate Normal distribution from which values for A, B, C and D are to be generated, subject to the above observational constraints. The earlier steps provide values A_0, B_0, C_0 and D_0 for these variates which are to be replaced by values A_2, B_2, C_2 and D_2 as the outcome of the present step, where the subscript 1 will be reserved for intermediate quantities. In the first of the sub-steps indicated above, one recognises that the variate A does not need to be treated here since one needs only to set $A_2=a$: note that $A_0=a$ is provided by previous iterations. One can therefore deal only with B, C and D having evaluated the mean and covariance matrix of this triplet conditional on $A=a$. The remaining problem can then be treated by the following set of three steps, each of which involves a pair from the variables B, C and D.

(i) generate B_1, C_1 from the conditional distribution given $D = D_0$ and $A = a$;

(ii) generate B_2, D_1 from the conditional distribution given $C = C_1$ and $A = a$;

(iii) generate C_2, D_2 from the conditional distribution given $B = B_2$ and $A = a$.

In each case the conditional distribution is a bivariate Normal, but subject to the additional constraints derived from the observational constraints. For example in case (i), the constraints are

$$b_1 \le B \le b_2, \quad c_1 \le C \le c_2, \quad t_1 - y_A(A_0) - y_D(D_0)$$
$$\le y_B(B) + y_C(C) \le t_2 - y_A(A_0) - y_D(D_0).$$

The required bivariate generation can then be achieved by generating B_1 from the marginal distribution given these constraints and C_1 from the conditional given that $B = B_1$. Here the marginal density $\phi(b)$ for B is obtained by integrating the joint density function $\phi(b,c)$ over the values of c satisfying the above constraints.

DISAGGREGATION PROCEDURES

The main stochastic model is formulated in terms of annual flow and rainfall values at measurement sites such as river gauging stations and raingauges, so one further aspect of the overall problem was the need to transpose these annual flow estimates to the sites of interest (in this case, dam sites) and to disaggregate these values to monthly values. Some possible approaches to this aspect of the problem are reviewed by McMahon and Mein (1986), Grygier and Stedinger (1991) and Maheepala and Perera

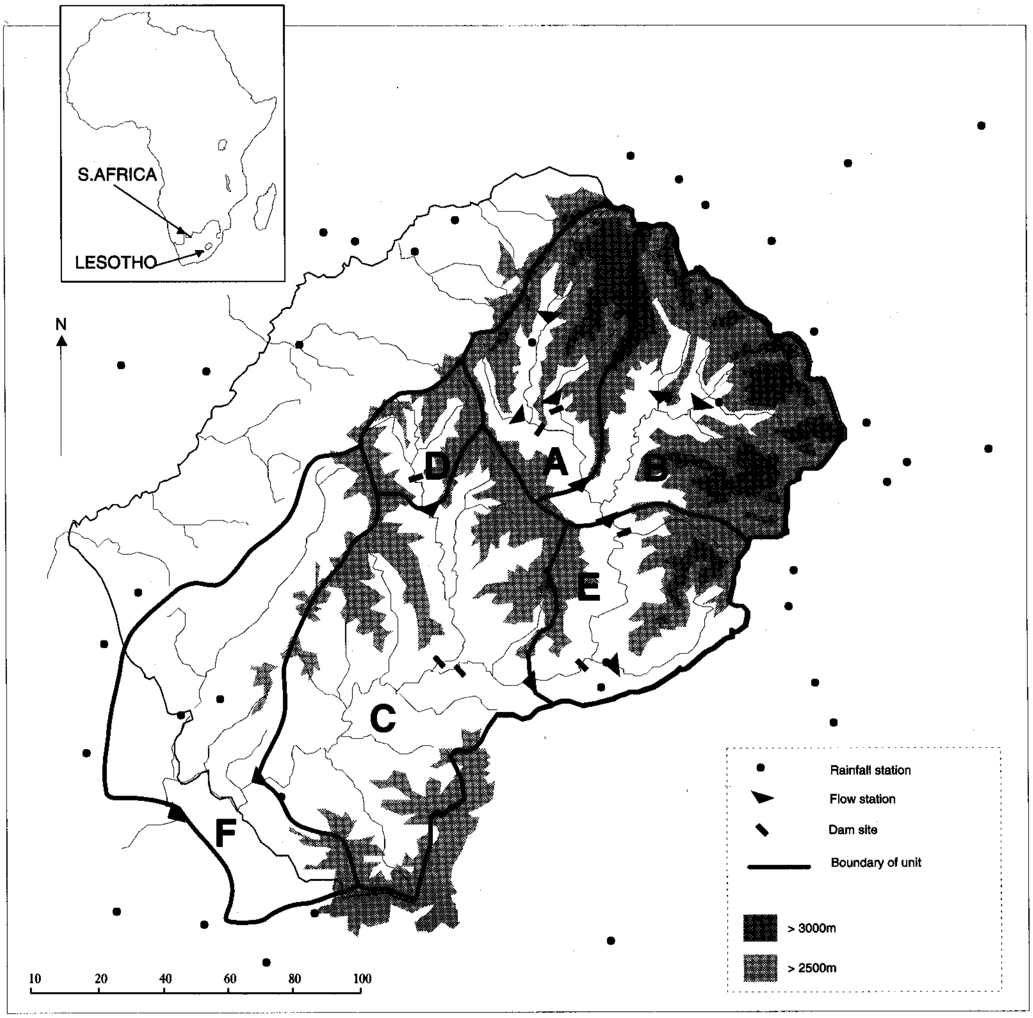

Fig. 1. *Location map showing the long-term rainfall and flow measurement sites considered in the model.*

(1996), amongst others. For the present example, the transposition was performed simply by a linear regression model with a stochastic component for the residuals: this was done separately for each of the groups of dam sites. The parameters of these regressions were fixed and could, in most cases, be estimated with reasonable accuracy due to the existence of several short-term flow records for measurement sites near to the proposed dam sites. This approach was felt to be consistent with the quality and availability of the observed flow and rainfall data and a more sophisticated approach, such as nesting additional rainfall-flow regression sub-models within the main sampling procedure, was not felt to be justified. The disaggregation to monthly values was performed using a simple yet robust procedure, called the Method of Fragments (e.g. McMahon and Mein, 1986). This essentially searches the observed flow record for the year which most closely matches the generated flows across all sites. The flow at each dam site is then distributed according to the monthly flow pattern observed at the closest flow measurement site. This approach has the advantage of preserving the natural variability and cross correlation between monthly flow values across the region and ensures that monthly values sum to the correct annual totals.

Calibration of the model

THE STUDY AREA

The suggested model provides one possible solution to the problem of estimating monthly flow and rainfall data (both historical and synthetic) for the situation of a catchment, or region, where some historical values in the observational period are missing, and some may only be known within known limits (i.e. bounded or censored values). The practical situation for which the model was originally developed, and which is described here, was flow estimation as part of the design studies related to the Lesotho Highlands Water Project (Fig. 1). This is a major scheme to construct a network of reservoirs and transfer tunnels to supply water from the highlands of Lesotho to the heavily populated and relatively dry regions to the north in South Africa. Flow estimates were required both for design of future phases of the scheme and for input to an operational study to assess the likely quantity of water available for delivery to South Africa. Sene *et al.* (1998) give further information on the scheme and on the hydrology of the Lesotho Highlands.

In common with many other parts of Africa, flow measurements in the highlands only began comparatively recently (in the 1960s) and, due to the remoteness of the area and occasional damage by floods, the records contain many gaps. There are few raingauges within the catchments of interest in northern Lesotho, although there are several good quality long term rainfall records further afield in Lesotho and in neighbouring regions of South Africa. A major review of the quality of the flow and rainfall data was undertaken as part of this study from which the records for 12 flow measurement sites and 35 raingauges in and around Lesotho were selected for inclusion in the model. A subset of 6 key flow measurement sites was chosen for use in the main flow generation component, based on physical considerations (e.g. proximity of raingauges to the catchments), the quality and completeness of the records and the results of cross-checking between sites. Throughout, annual total flow and rainfall values were computed for a hydrological year (August to July) chosen to minimise the serial correlation between successive annual values.

Figure 2 shows the schematic representation adopted for most of the modelling studies, and indicates the basic approach which was to model the incremental flows from the separate contributing areas, or 'flow units', between the flow gauging sites. Thus the model is posed in terms of flows from Units A to F, but can also take into account observations on totals of flows from these units, denoted by 'T_AB', etc. The advantage of expressing the flow data in this form is that it highlights the relationships between rainfall and flow records in each part of the region.

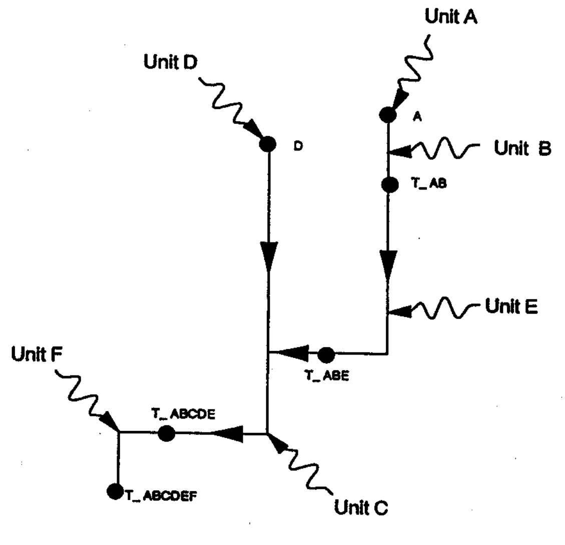

Fig. 2. *Schematic indicating the division into flow contributing areas between the flow measurement sites (areas defined on Fig. 1).*

PROCEDURES FOR ESTIMATING BOUNDS ON DATA

For years with incomplete or missing data, two main procedures were used for estimating upper and lower bounds on the annual values. Figure 3 shows typical examples of the resulting bounded rainfall and flow series. The overall aim was to define bounds which were wide enough to definitely contain the true value, yet not so wide that they contributed little additional information to the model. In practice, for many of the years considered, the estimated bounds were often within a few percent of each other.

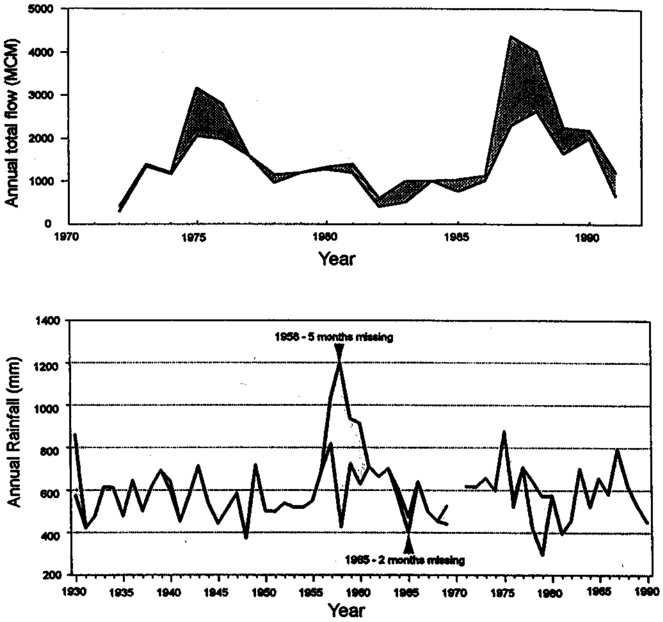

Fig. 3. *Typical examples of bounded annual flow and rainfall records.*

For the flow data, use was made of the fact that all of the measurement sites lie on the same river network. The daily mean flow at a given station is therefore unlikely to be less than the sum of the flows from all contributing stations further upstream or to exceed the flow at stations further downstream (less any contributions from tributary inflows). In the case that the nearest stations upstream or downstream also have missing data, then the search for measured values was extended further upstream or downstream as appropriate. The data used in this procedure therefore included all 12 of the basic flow measurement records. The resulting bounded and actual daily values were then summed to give the required bounded annual totals for the 6 main flow units. One consequence of expressing the annual flow data in the form of incremental values from each of the separate contributing catchment areas is that the above procedure produces bounds not only on these individual component flows but on certain totals of them. For example, the record at the most downstream site, when available, provides, in a given year, a value for the total of the flows across all 6 of the major flow units. The stochastic modelling procedure was designed to be capable of including both the information provided by bounds on the individual flow-unit values and the extra information given by the bounds and exact values for various sub-totals amongst them.

The second procedure used for estimating bounds was to prepare scatter plots of annual values at a given site against values at neighbouring sites. Envelope curves encompassing all values were drawn where justified and these allowed plausible bounds to be estimated in years with missing data. This method was used primarily for rainfall data but also served as a cross-check on the bounds estimated for the flow values and, in some cases, provided a basis for specifying slightly tighter bounds than those estimated using the first of these methods. For years where data for only a few months were missing, these values were used in combination with the scaled monthly maximum and minimum records for the region to derive plausible bounds on the annual totals for the year. In some instances, when the bounds produced by these procedures were very wide compared to the range of the exact observations, a missing value was substituted instead.

NORMAL SCORE PLOTS

In stochastic modelling of river flows, the usual way of judging the adequacy of the assumption of marginal Normality is to employ a Normal-score plot. To maximise the amount of data used in these assessments, it was desirable to use bounded values as well as observed values, and a new procedure was developed for this situation. This was based on a revised form of the empirical distribution function F_n defined, for a sample of size n, as:

$$F_n(x) = n^{-1}\Sigma\ G_i(x),$$

where the function $G_i(x)$ is the contribution associated with the i'th ranked data-item. For the case of an exact observation, the obvious choice is to use $G_i(x) = \mathrm{I}(x-x_i)$, where I is an indicator function defined here by

$$\begin{aligned}\mathrm{I}(x) &= 0, & x &< 0;\\ &= ½, & x &= 0;\\ &= 1, & x &> 0.\end{aligned}$$

An equivalent version of the conventional Normal-score plot is obtained by plotting the values of the Normal score $\Phi^{-1}\{F_n(x)\}$ as a function of the observed values of x, where Φ is the standard Normal distribution function. For bounded data it then remains to define a suitable function to take account of the lower and upper bounds, L_i and U_i. To retain the character of the standard definition of the empirical distribution function and to recover this function when the bounds become very close together, it is clear that $G_i(x)$ should take the values zero below L_i and one above U_i. In the case that the bounds are far apart, this requires a formulation that will take into account the intended use as part of the definition of the Normal-score plot. For this reason, the following procedure was adopted. First, using a maximum likelihood approach, a Normal distribution is fitted to all the data available for a given record (after any initial transformations), including both exact and bounded observations. This yields estimates for the mean m and standard deviation s of the distribution. Then the contribution of a bounded observation to the empirical distribution function is defined to be the distribution function of the Normal distribution with parameters m and s, truncated to the range L_i to U_i. Specifically, for $L_i < x < U_i$:

$$G_i(x) = [\Phi\{(x\text{-}m)/s\} - \Phi\{(L_i\text{-}m)/s\}]/ \\ [\Phi\{(U_i\text{-}m)/s\} - \Phi\{(L_i\text{-}m)/s\}].$$

This definition has been chosen because it should mean that the contribution to the Normal score plot arising from an incompletely observed data-item should tend to be a portion of straight line, rather than a curve. Figure 4 shows an example of a Normal-score plot for the Lesotho Highlands data constructed in this way. Here, the line shown is the Normal-score function, including all the information for the variate, whilst the crosses mark the function at the "exact" data points, which always lie on a horizontal segment of the plot. A more conventional plot, using only observed data points, would of course only show the crosses, and would omit the extra information provided by the bounded values.

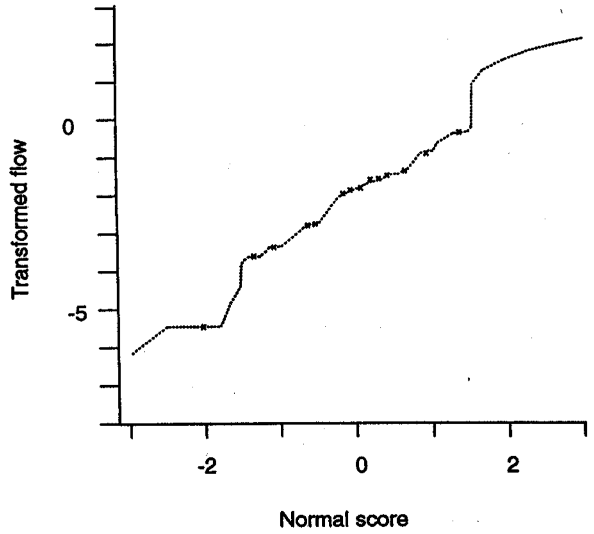

Fig. 4. *Example of a Normal Score plot for transformed annual flow data.*

TRANSFORMATIONS OF FLOW AND RAINFALL DATA

The Normal Score plots were used to evaluate the transformations by which annual data were transformed to Normality. Since the model uses rainfall data as well as flow data, the choice of transformations needs to be based on a compromise between the following three considerations:

(a) marginal Normality of the transformed variables,
(b) constancy of variance in the conditional distribution of one variable given another (particularly of transformed flow given transformed rainfall), and
(c) physically sensible results from the regression of flow on rainfall when expressed in the original units.

Most conventional stochastic flow generation models do not consider rainfall data and so usually concentrate only on satisfying the first of these factors. The main factor determining the choice of transformation was the visual appearance of the Normal-score plots, although point (b) was also checked using scatter plots of transformed flow against transformed rainfall, and inspection of the full set of scatter plots of pairs of variates. A computer graphics program, showing up to 12 plots per 'screen', was written to allow a large number of plots to be inspected in a short time by 'paging' through all possible combinations of data. The final transformation chosen for most of the simulations was of the form:

$$t(y) = y \qquad for\ y \geq h$$
$$t(y) = h + (h-L)\ log\ \{(y-L)/(h-L)\} \qquad for\ L < y \leq h$$

where y is the observed value, t is the transformed value, L is a lower bound on the generated values and h is a threshold parameter above which the logarithmic transformation does not apply. Values for the parameters h and L were selected individually for each flow and rainfall site by iterative adjustment based on the visual judgement of Normal-scores plots as described above.

The lower limit was introduced to improve the fit of the transformations at the lower end of the flow range, and was constrained to be below any of the observed flow values available (which included data from the worst known drought on record). The threshold parameter, h, was also introduced to improve the relationships between rainfall and flow data. For the Lesotho Highlands, it was anticipated that, when considering annual data, it should be possible to develop linear relationships between transformed rainfall and transformed flow since, at the time of these design studies (i.e. before dam construction), there was little storage in the basin (either surface or sub-surface) to provide any carry over of flows between hydrological years. This was subsequently justified by the data (see later). Ordinary logarithmic transformations resulted in a regression equation of the form:

$$flow \sim constant.\ (rainfall)^b$$

with b typically close to 2. However, a relationship such as this would mean that the model might predict unreasonably large values for flows in years with high rainfall (i.e. more flow than rainfall), and the threshold parameter was introduced as a means of avoiding this unrealistic behaviour.

Several other types of transformation were also evaluated during this study including other variants of the logarithmic transformation to Normality and included a 'logistic' type of transformation which allows for the model to represent data which are bounded both above and below: this particular transformation has, for flow data alone, been found suitable in previous simulation studies in southern Africa (Basson *et al.*, 1994) and elsewhere. These comparisons produced broadly similar results, suggesting that the remaining uncertainty in the estimates derives more from the availability and variability of the data than from the choice of transformation. Figure 5 shows an example of a plot of the transformed annual rainfall and annual flow values plotted together as a scatter plot. Bounded data are shown as dashed lines or boxes covering the range within which the data are known to lie; however these are omitted if the range covered is large since otherwise the plots are obscured. Scatter plots such as these were also used to help assess the constancy of variance of the regressions between pairs of variables.

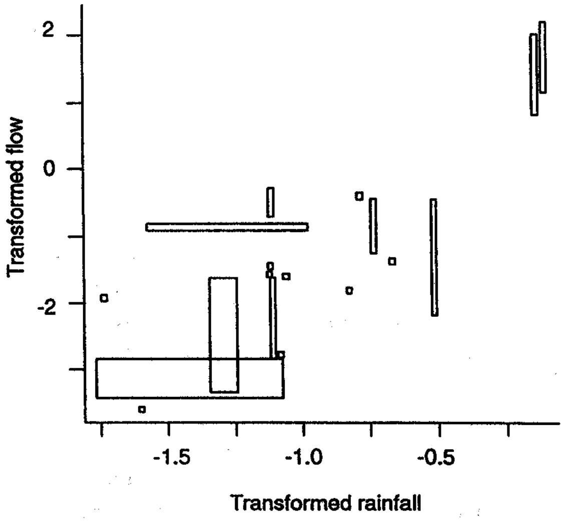

Fig. 5. *Example of a scatter plot of transformed annual flow and rainfall data.*

Results

GENERAL APPROACH TO MODEL VALIDATION

The main output from the Gibbs sampling procedure was a set of individual realisations of the annual flow sequences for each of the incremental flow-units and, by summing these, for the originally selected flow measurement sites. The three main questions to consider when assessing the performance of the model were:

a) In periods with flow data, can the model simulate the key statistical and time varying features of the observed flows by using rainfall information alone ?
b) In periods with rainfall data but no flow data, do the generated flows correspond with other indicators of

flow variability, such as regional rainfall records, flows generated by other types of model and flow records from more distant sites ?

c) When generating completely synthetic values, does the model generate flow values with the same statistical characteristics as the observed flow data ?

To assist in these validation tests, the model was configured to produce automatically three types of output in which flows were generated using (I) all of the available flow and rainfall data, (II) rainfall data alone and (III) 'no data'.

The differences between these three modes of operation can best be seen from an example of the probability distributions for each year for the generated flows computed across all realisations. Figure 6 shows one such example for the 5,25,50,75 and 95 percentage points calculated in each mode for one of the sites in the Lesotho Highlands study. In Mode (I) any missing flow values are generated taking account of all of the available information, including bounds on values for the given flow-unit, flow values or bounds for other flow-units, and all the available rainfall information in the form of exact values or bounds. This mode corresponds directly to the Gibbs sampling procedure, while the other two modes were added to create other versions of the flow-series which take no part in the estimation procedure: however the 'current' versions of the model parameters were used at each iteration of the procedure to create these other realisations. In Mode (II), a revised procedure was implemented which ignores all the flow information so that the model effectively operates as a stochastic rainfall runoff model. In this mode, infilled values for the rainfall data are generated based on the observations and bounds on rainfall, and these are then used via the regression equations, with the regression noise included, to produce a realisation of the flows. The Mode (III) 'no data' case corresponds most closely to a conventional stochastic flow generation model, in which the flow sequences produced are completely synthetic, but hopefully retain the key statistical characteristics of the observed flows. In this mode of operation, the present model effectively generates artificial sequences of rainfall and flow values concurrently, although it does this in a somewhat more efficient way by generating values for the regression variables z only.

In addition to producing these different types of output, several standard computational checks on the procedure were performed: for example, using different seeds in the random number generators and rearranging the internal orderings of the sets of flow and rainfall variates. Between them, these tests provided a check not only on the programming, but on the effect of the initial values set for the 'unknown' parameters and data-values. In the case of the flow-variates, test runs were made in which the variates were re-ordered to ensure that the initial values obeyed the constraints imposed by information about totals of flows from groups of flow-units. The results from all these runs of the model were in substantial agreement. To check the independence of successive realisations, simple correlation analyses were performed across realisations of various sample statistics, such as the means and standard deviations. These suggested that a gap of only 4 or 5 between realisations was enough for them to be treated as statistically independent due to the inherently weak correlation between successive realisations. Since the validation analyses reported here do not rely on the realisations being statistically independent of one another, it seemed appropriate to use all the realisations in a sequence rather than using only every tenth, say. The results shown in the following sections were based on using 800 realisations obtained after discarding the first 20 as a warm-up period.

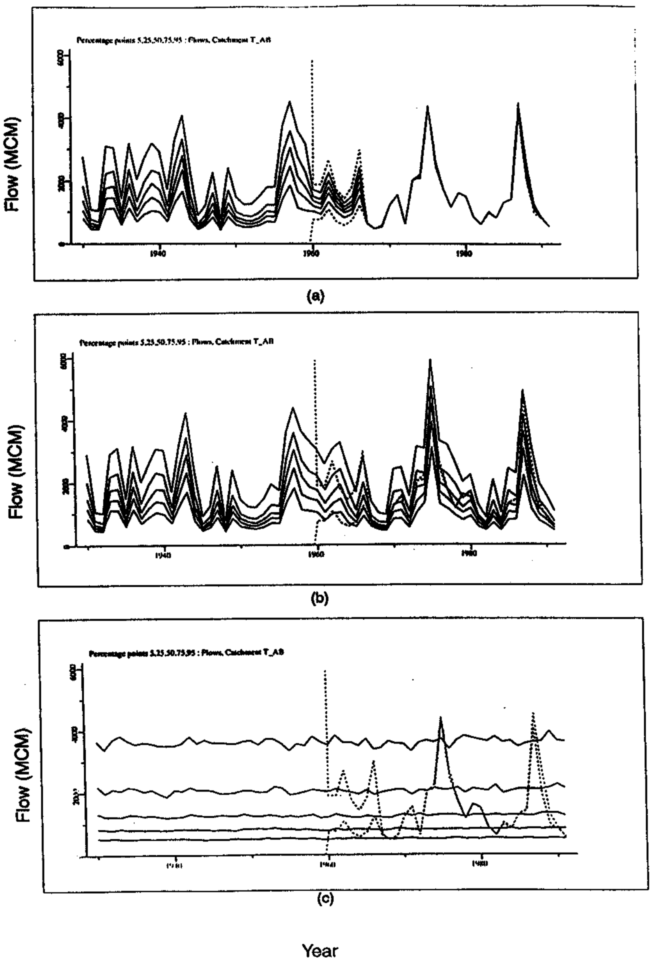

Fig. 6. *Examples of model output. Percentage points of modelled flows, together with the known flow data or bounds (a) Conditioned on all flow and rainfall data (Mode I); (b) Conditioned on rainfall data only (Mode II); (c) Unconditional (Mode III).*

EXAMPLE RESULTS

The usual approach to evaluating the output from conventional stochastic flow generation models is to compare selected statistical characteristics of the generated flows, such as the mean, variance and skewness, with those of the observed flows. This type of comparison can be attempted using both the Modes(II) and (III) of operation described above i.e. conditionally and unconditionally on the observed rainfall information. However, here the observed flow records contain many gaps and instances of information in the form of bounds, and hence the task is more complicated than usual in that a single 'observed' value for each of the sample statistics cannot be calculated. Instead, there is a probability distribution associated with each statistical measure, for example reflecting how much is known about what the actual sample mean annual flow is in a given period and one must ask whether the observed statistic was likely to have arisen from the distribution. These distributions are provided by the Mode (I) results of the model, which take full account of all relevant data. Figure 7 shows an example of these types of comparison in the form of a box plot for one site in the Lesotho Highlands for a choice of averaging periods for two of the sample statistics selected; the mean and the standard deviation. If the model is performing well, the median values should lie within a few percent of each other, at least for these particular statistics. The three averaging periods selected here overlap: the 'site-data' period was selected individually for each flow-site to represent the period having observations for that site; the 'flow-data' period represents the period when there were data for any flow-site; the 'all data' period is the overall period when either rainfall or flow information is available. Note that the variability of the results increases across these three periods for the Mode I case, because a larger fraction of the information is derived from rainfall: the variability decreases for the Mode III case because the distribution is identical from year to year but the statistic is calculated for a longer period.

Similar plots were prepared for other sites and other sample statistics which, taking into account the application to reservoir design, included the sample maximum and minimum, the serial and cross correlation matrices and various storage related statistics computed assuming a range of assumed values for reservoir yield, such as the maximum volumetric deficit, the duration of maximum deficit and the minimum cumulative (run-sum) totals for various assumed critical periods. Most attention was placed on statistics defined in terms of the flow data on the original, untransformed scale, but some consideration was made of statistics in the transformed space. Clearly, for this many sites and statistical parameters, the number of comparisons required can be very large so, to assist in this work, a formal 'measure of fit' was defined on a sliding scale of 0.0 to 1.0, where a score of 1.0 indicated a very good fit (agreement of medians perfect given the sample size) and a score of 0.0 indicated the predicted median value fell outside the observed range estimated for the statistic under consideration. These scores were averaged by site and/or by statistic and were useful in assessing rapidly different configurations of the model, including the effect of different transformations. There was some difficulty in assessing how close a match between the distributions to expect. Some statistics, such as the sample means in the transformed space, should be perfectly preserved because of the structure of the model: conversely, if there were statistics which are completely unrelated to the model structure one might treat these as if the 'observed' statistic was supposed to be a random sample from the distribution derived from the model. However, for the present model, all statistics calculated from the untransformed values seem to fall somewhere between these two cases. This

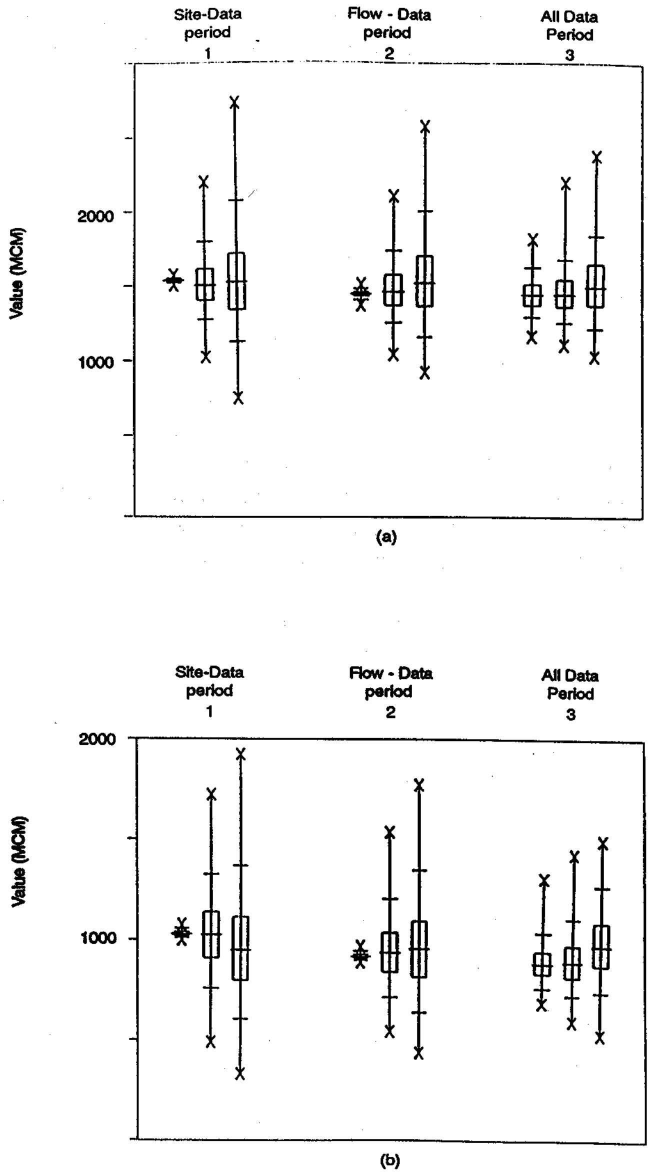

Fig. 7. *Examples of box plots comparing the estimated mean and standard deviations in flows for three different averaging periods. Within each group of three box plots, values are for Modes I, II and III of operation (a) Mean (b) Standard Deviation.*

measure of fit was therefore used primarily as a way of comparing the performance of different configurations of the model, without attaching too much importance to the actual scores obtained in each configuration.

In Mode (II) of operation, the model performance might also be evaluated on a time series basis if the central estimates of flows in each year are considered to be representative of the true flows. Figure 8 shows an example of this type of comparison for one of the sites in the Lesotho Highlands, where the time series values are shown as a scatter plot. Note that, in this plot, the observed data are indicated as a range of values between the estimated bounds. A more truly independent test is to perform the same type of comparison with various other indicators of flow variability which were not included in the model. For the Lesotho Highlands example, such indicators included flow and rainfall records from more distant stations in south Africa and also 'non-stochastic' flow estimates derived in earlier unpublished studies using physically-based conceptual rainfall-runoff models driven by monthly rainfall data. Figure 9 shows an example of this type of comparison for one of the sites included in the stochastic model. The overall conclusion from these various tests was that the model was capable both of providing realistic estimates for the historic flows, and of generating plausible synthetic flow sequences for use in reservoir design.

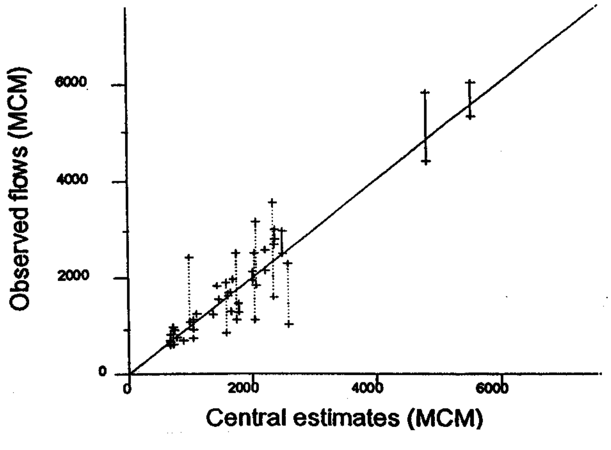

Fig. 8. *Scatter plot of observed and predicted flows in Mode I.*

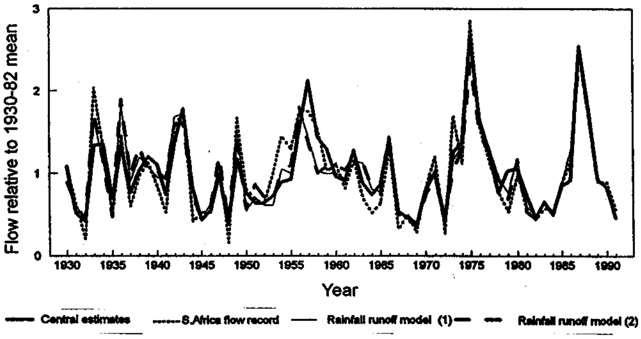

Fig. 9. *Time series comparison of central estimates (Mode I) against other models and records from more distant stations.*

SENSITIVITY TESTS

Various tests were also performed to determine the sensitivity of the model results to various assumptions and different configurations of the model. Although many statistical measures were considered as indicated above (run-sums etc), the emphasis was on the sample mean and standard deviations of the flow series since, for reservoir design, these are the key indicators of the average amount and variability of the water available for storage. In particular, the posterior means and standard deviations of these two sample statistics were used to compare various model configurations.

These comparisons showed that typically the posterior means remained fairly constant whilst the standard deviations varied slightly for different model configurations, depending on the choice of raingauges included within the rainfall-flow model and on the number of raingauges included in the overall model. These differences reflected the amount of information about the flows in particular years being extracted from the rainfall data. As well as using the full set of 35 raingauges, models based on only 25 and 13 raingauges were tried. Some results for this comparison are presented in Table 1 for the period 1930 to 1982, which was one of the periods for which reservoir yield assessments were required in this study. In these runs, the same selection of raingauges was used in the rainfall-flow regressions throughout. This table also includes the results for several independent runs using 25 raingauges overall in order to illustrate the remaining uncertainty in deriving estimates of the posterior means and standard deviations from only 800 realisations. Similar comparisons were also performed for the full set of model validations runs, which were designed to assess the sensitivity of the model output to factors such as:

(i) use of separate selections of raingauges for the different flow-units, compared with using a single combined rainfall sequence for all units and with using no rainfall information;
(ii) the effect of using a different family of transformations;
(iii) the effect of initial values for the unknown quantities; and
(iv) the effect of the parameters of the prior distributions used for the covariance matrices of rainfalls and flow residuals.

These various tests generally produced results which fell within the range of values indicated in Table 1, showing that the main factors influencing the accuracy of the model were the number of raingauges used in the model,

Table 1. Comparison of the posterior distributions for models based on differing numbers of raingauges overall: estimated posterior means and standard deviations for statistics of yearly flows at site *T_AB*

Model	*Sample mean for years 1930–1982 ($Mm^3 yr^{-1}$)*		*Sample between year standard deviation for years 1930–1982 ($Mm^3 yr^{-1}$)*	
	Mean	*St.Dev.*	*Mean*	*St.Dev.*
No raingauge information	1485	160	917	144
13 raingauges	1474	86	841	81
25 raingauges	1464	78	843	75
	1461	84	838	77
	1447	81	845	80
35 raingauges	1453	80	832	70

and the completeness (and closeness of the bounds) of the observed data.

Discussion and conclusions

The Gibbs Sampling approach, when incorporated within a Bayesian multivariate regression framework, provides an attractive way of maximising the amount of observed information used when generating river flow sequences for reservoir design. The method allows both rainfall data, and flow and rainfall values for which only bounds are available, to be used to help guide the flow generation procedure. It also allows considerable flexibility in the way that the model is configured, so that it is easy to experiment with different combinations of flow measurement sites and raingauges, and to build physically realistic assumptions into the regression relationships between flow and rainfall. For the Lesotho Highlands application described in this paper, the method provided a way of generating plausible long-term flow sequences, both historic and synthetic, from only a limited set of observed flow data. The modelling procedure adopted met both of the original main objectives of the study, which were:

(i) estimation of the flows that actually occurred in past years when flows were not measured;
(ii) generation of stochastic flow sequences, either conditioned on the actual past rainfall or unconditionally.

Although the model used was tailored specifically for the Lesotho Highlands, a similar approach could be used in other regions with limited, patchy flow data, but with longer, more reliable regional rainfall records. In the Lesotho Highlands application, one particular advantage of using rainfall data in the stochastic model was that this provided a long-term view on the climatic variability in the region. This was considered important given that, as mentioned earlier, several studies have shown strong, but not conclusive, evidence of cyclical behaviour in rainfall records for southern Africa. With the structure adopted, any cyclical behaviour in the rainfall data will be automatically transferred to the generated historical flow values. Given that the presence, or otherwise, of cycles is a contentious issue, this side-stepped the problem of formally identifying the cycles from the flow data alone and of including some representation of this behaviour in the flow generation process.

Acknowledgements

This work was performed under contract to the Lesotho Highlands Development Authority, whom we thank for their assistance in preparing the rainfall and flow data-sets. Data were also provided through the Lesotho Meteorological Service and the Department of Water Affairs and Forestry in South Africa. Several people from the Institute of Hydrology also made valuable contributions to this study, including Jonathon Moores, Flora Cecil, Matthew McCartney, Jeremy Meigh and Val Bronsdon.

References

Arnold, S.F., 1993. Gibbs Sampling. In *Computational Statistics (Handbook of Statistics, Vol 9)*, ed. C.R. Rao, 599–625. North-Holland, London.

Basson, M.S., Allen, R.B., Pegram, G.G.S. and van Rooyen, J.A., 1994. *Probabilistic Management of Water Resources and Hydropower Systems.* Water Resources Publications, Colorado, USA.

Besag, J. and Green, P.J., 1993 Spatial Statistics and Bayesian Computation, *J. Roy. Statist. Soc. B*, **55**, 25–37.

Box, G.E.P. and Tiao, G.C., 1973. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, London.

Bras, R.L. and Rodriguez-Iturbe, I., 1985. *Random functions and Hydrology.* Addison-Wesley, London.

Geman, S. and Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intelligence,* **6**, 721–741.

Gilks, W.R., Clayton, D.G., Spiegelhalter, D.J., Best, N.G., McNeil, A.J., Sharples, L.D. and Kirby, A.J., 1993. Modelling complexity: applications of Gibbs sampling in medicine. *J. Roy. Statist. Soc. B*, **55**, 39–52.

Grygier, J.C., Stedinger, J.R. and Yin, H.B., 1989. A generalised maintenance of variance procedure for extending correlated series. *Water Resourc. Res.,* **25**, 345–349.

Grygier, J.C. and Stedinger, J.R., 1991. Condensed disaggregation procedures and conservation corrections for stochastic hydrology. *Wat. Resour. Res.,* **24**), 1574–1584.

Kroll, C.N. and Stedinger, J.R., 1996 Estimation of moments and quantiles using censored data. *Water Resourc. Res.,* **32(4)**, 1005–1012.

Lawrance, A.J. and Kottegoda, N.T., 1977. Stochastic modelling of riverflow time series. *J. Roy. Statist.Soc.A,* **140**, 1–47.

Maheepala, S. and Perera, B.J.C., 1996 Monthly hydrologic data generation by disaggregation. *J. Hydrol.,* **178**, 277–291.

McMahon, T.A. and Mein, R.G., 1986. *River and Reservoir yield,* Water Resources Publications, Colorado, USA

Pegram, G.G.S., 1994. Patching monthly streamflow data—a case study using the EM algorithm and Kalman filtering. In *Stochastic and Statistical Methods in Hydrology and Environmental Engineering, Vol. 3, 449–457.* K.W. Hipel *et al.* (eds.). Kluwer Academic Publishers, Netherlands.

Sene, K.J., Jones, D.A., Meigh, J.R. and Farquharson, F.A.K., 1998. A review of the climatology and hydrology of the Lesotho Highlands. *Int. J. Climatol.,***18**, 329–345.

Smith, A.F.M. and Roberts, G.O., 1993. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *J. Roy. Statist. Soc. B*, **55**, 3–23.

Stedinger, J.R., Pei, D. and Cohn, T.A., 1985. A condensed disaggregation model for incorporating parameter uncertainty into monthly reservoir simulations. *Wat. Resour. Res.,* **21(5)**, 665–675.

Tanner, M.A., 1993. *Tools for Statistical Inference.* Springer-Verlag, New York.

Tyson, P.D., 1991. Climatic Change in Southern Africa: Past and Present Conditions and Possible Future Scenarios. *Climatic Change,* **18**, 241–258.

Valdes, J.B. and Rodriguez-Iturbe, I., 1977. Bayesian generation of synthetic streamflows 2. The Multivariate case. *Wat. Resour. Res.,* 13, 291–295.

Wood, E.F., 1978. Analyzing hydrological uncertainty and its impact upon decision making in water resources. *Adv. Water Resour.,* 1, 299–305.

Zucchini, W. and Hiemstra, L.A.V., 1983. A method which preserves the statistics of importance in storage applications using related records to augment the available record. In *Proc. S.African Nat. Hydrol. Symp.,* H.Maaren (ed.), Dept. Water Affairs and Forestry, Pretoria.