Efficient Bayesian Model Selection in Hydrological Modelling

L. Marshall[a] **A. Sharma**[a] **and D.J. Nott**[b]

[a]School of Civil and Environmental Engineering, University of New South Wales, Sydney, Australia.

[b]School of Mathematics, University of New South Wales, Sydney, Australia.

Abstract: With the wide range of models available, hydrologic modellers are faced with the choice of which model is best applied to a catchment for a particular modelling exercise. Assessing the relative performance of competing models can be difficult given the limited data that is available and further complicated by difficulties in obtaining a unique set of values for the model parameters. Traditional techniques such as those involving split-sample validation are useful, but suffer from increased uncertainty due to the reduction of the sample used. Bayesian statistical inference, with computations carried out via Markov Chain Monte Carlo (MCMC) methods, offer an efficient alternative allowing for the combination of any pre-existing knowledge about individual models and their respective parameters with the available catchment data to assess the parameter uncertainty. Bayesian inference can also provide a framework to evaluate the evidence in favour of a model, given a group of competing models. The traditional approach requires calculation of the *Bayes factor*, which is the posterior probability ratio of the models (assuming equal prior probabilities). The aim of this study is to present a method by which hydrological models may be compared in a Bayesian framework. The study builds on previous work in which the parameters of the Australian Water Balance Model (AWBM) were estimated using computations carried out via MCMC methods. The study considers the variability of soil moisture within the Bass River catchment, by formulating the AWBM to include a different number of soil moisture stores. A model selection framework is developed by calculating Bayes factors using a method based on direct estimation of a models' marginal likelihood. The framework uses an adaptive Metropolis algorithm to calculate the model's posterior odds. To assess the model selection method in a controlled setting, artificial runoff data were created corresponding to the two storage model. These data were used to check if the method would select the 2 store model convincingly. The method was then applied to real catchment data to determine which model configuration best represents the catchment.

Keywords: *Bayes Factor; Hydrological Modelling; Markov chain Monte Carlo; Australian Water Balance Model.*

1. INTRODUCTION

A challenge that faces the practicing hydrologist is that there isn't a catchment model that will perform accurately over the wide range of conditions that exist. A modeller must choose the most appropriate model from a number of competing models. Typically, models may be compared using classical assessment criteria (such as the mean square model error or AIC). However, difficulties in obtaining unique and accurate parameter values can mean that assessing the performance of models is problematic.

Bayesian inference provides an approach to model comparison that overcomes the difficulties of traditional methods. The uncertainty about the parameter values for a model is ascertained by combining any pre-existing knowledge about the model variables with available data to obtain a distribution on the parameter space (the posterior distribution) summarizing parameter uncertainty. In comparing two models, the traditional approach requires calculation of the Bayes factor, which is the posterior probability ratio of the models (assuming equal prior probabilities).

2. BAYESIAN MODEL COMPARISON THEORY

Traditional comparison of hydrological models has proceeded by splitting available catchment data, and calibration/validation of the model using the split data samples. Models may be assessed according to a range of criterion or via traditional hypothesis testing. Methods based on hypothesis testing generally require models to be nested, which is rarely the case in hydrological modelling. In addition, comparison of models can often be hindered by the uncertainty surrounding model parameters. This uncertainty often arises due to lack of model identifiability, limited data or over-parameterization. This is of particular importance when considering hydrological modelling and the frequent lack of data in comparison to the ever increasing model complexity.

Bayesian theory provides an ideal solution to these limitations. Bayesian methods consider model uncertainty when comparing model performance by describing model and parameter uncertainty probabilistically. Methods based on Bayes factors are flexible and simple. Many models may be compared without a change in the method, and there is no limit on the model structure (models are not required to be nested). Bayesian methods also allow prior knowledge about the models being considered to be taken into account. The results are not solely reliant on information from the available data, but enable specialist and pre-existing knowledge to inform the results. Bayesian model selection methods are also natural Ockham's razors (Berger and Pericchi, 2001), favouring simpler models automatically.

2.1. Bayes Factors

The most widely used Bayesian method of model comparison is the Bayes factor. The Bayes factor is the posterior probability ratio of the models, assuming equal prior probabilities. Say we wish to choose a model from the set of models $M=\{M_1,\dots M_n\}$, given data y for implementing the model. Let $p(M_i)$ be the a priori probability set for model M_i, and θ_i be the set of uncertain model parameters corresponding to model M_i. The traditional approach to Bayesian model selection proceeds by pairwise comparison of the models through their posterior probability ratio:

$$\frac{P(M_1 \mid y)}{P(M_2 \mid y)} = \frac{P(M_1)}{P(M_2)} \times \frac{m(y \mid M_1)}{m(y \mid M_2)} \qquad (1)$$

where $P(M_i \mid y)$ is the posterior probability of model M_i, $P(M_i)$ is the prior probability of model M_i, and $m(y \mid M_i)$ is the model's marginal likelihood. The second term on the right hand side of Equation (1) is known as the Bayes factor. The difficulty in calculating Bayes factors largely lies in estimation of the marginal likelihood of the model. The marginal likelihood can rarely be obtained analytically. The densities $m(y \mid M_i)$ are defined through an integral over the parameter space, which must usually be estimated numerically, typically using specialized numerical methods such as Markov chain Monte Carlo (MCMC).

2.2. Problems in Computing Bayes Factors

The direct calculation of Bayes factors can be difficult. MCMC methods have been successfully implemented to sample model and parameter space, but in high dimensional models the tuning of such MCMC schemes to get appropriate mixing can be difficult. Alternatively, methods based on separate MCMC runs for each model have been developed (Chib and Jeliazkov, 2001).

The marginal probability is estimated using Bayes' theorem as a function of the parameter posterior distribution evaluated at optimal parameter values. Estimating the Bayes factor using a traditional Bayesian approach would involve first formulating the parameter's posterior distribution, then estimating the optimal parameter values and then evaluating the marginal probability $m(y|M)$. If MCMC methods are used, this would involve generating traces twice – first to estimate the parameter posterior distribution and associated optimal parameter values, and then to estimate the marginal probability given the identified optimal parameters. The algorithm proposed in the next section reduces some of the difficulties associated with this calculation.

Bayes factors require specification of prior distributions for all parameters in the models considered. This can be an advantage, as it provides a way in which other information may be taken into account in the model. However, for many models it may be hard to set meaningful priors for all model parameters. As model dimension grows, the task becomes more demanding. Estimation of the Bayes factor tends to be sensitive to the choice of prior (Kass and Raftery, 1995) and difficulties can be encountered when using improper priors.

3. METHODOLOGY

3.1. Estimation of Models' Marginal Likelihood

To overcome the problems associated with prerun tuning and computational effort in computing Bayes factors, Chib and Jeliazkov (2001) provide a MCMC framework for estimating the marginal likelihood of a model by integrating over the parameter space. The marginal likelihood is the normalizing constant of the model's posterior density, and (using the notation defined previously) can be written as:

$$m(y \mid M) = \frac{f(y \mid M,\theta)\pi(\theta \mid M)}{\pi(\theta \mid y,M)} \qquad (2)$$

where $f(y \mid M,\theta)$ is the likelihood or error function, $\pi(\theta \mid M)$ is the prior density and $\pi(\theta \mid y,M)$ is the posterior density of the model parameters. The problem with using MCMC methods for calculating the marginal likelihood is that the method obtains draws from the posterior distribution, whereas the marginal likelihood is obtained by integrating the likelihood function with respect to the prior. Chib and Jeliazkov have devised a method in the context of MCMC methods by relating the marginal likelihood to the posterior density calculated at a single point. By taking logs, the above identity at some value θ* is obtained as:

$$\log m(y \mid M) = \log f(y \mid M, \theta^*) + \log \pi(\theta^* \mid M) - \log \pi(\theta^* \mid y, M) \quad (3)$$

From this, the marginal likelihood can be found by estimating the log posterior ordinate, $\log \pi(\theta^* \mid y, M)$. Chib and Jeliazkov's method for estimating the posterior ordinate uses the sampled parameter values from the MCMC process. Thus little further computation is required than that to obtain samples from the model parameters' posterior distributions.

Evaluation of the models' marginal likelihood first requires an estimate of the value θ*, where the posterior density of the model parameters is maximized. Then, the aim of the exercise is to estimate the posterior density at θ*. By integrating over the parameter space, an estimate of this posterior ordinate is obtainable as:

$$\pi(\theta^* \mid y) = \frac{K^{-1} \sum_{g=1}^{K} \alpha(\theta^{(g)}, \theta^* \mid y) q(\theta^{(g)}, \theta^* \mid y)}{J^{-1} \sum_{j=1}^{J} \alpha(\theta^*, \theta^{(j)})} \quad (4)$$

where $\theta^{(g)}$ (for g = 1…K) are draws from the posterior distribution generated via the adaptive Metropolis algorithm (outlined below); $\theta^{(j)}$ are draws from $q(\theta^*, \theta \mid y)$, the proposal density for the transition from the fixed parameter value θ* to θ; and $\alpha(\theta, \theta^* \mid y)$ is the acceptance probability for a transition from θ to θ*.The benefits of the method lie in its flexibility and relative computational ease. The method can be applied to high dimensional models, of any structure for which a suitable MCMC sampling algorithm can be devised.

3.2. Adaptive Metropolis Algorithm

The adaptive Metropolis algorithm (Haario et al. 2001) is a variation on the conventional Metropolis algorithm. The adaptive algorithm is characterised by a multivariate normal proposal distribution, with mean at the current parameter value. The proposal covariance matrix is calculated at each iteration based on the covariance matrix of the parameter values in the parameter chain to that point. In this way, the proposal distribution is updated using the knowledge learnt so far about the posterior distribution. At iteration t, Haario et al. (2001) consider a multivariate normal proposal $N(\theta_t, C_t)$ where C_t is the proposal covariance.

For an initial period t_0, $C_t = C_0$ (where C_0 is some arbitrary initial covariance). After this initial period the proposal covariance is based on the estimated posterior covariance of the parameters:

$$C_t = \begin{cases} C_0 & t \le t_0 \\ s_d Cov(\theta_0 \ldots\ldots \theta_{t-1}) + s_d \varepsilon I_d & t > t_0 \end{cases} \quad (5)$$

where ε is a small value simply to ensure C_t does not become singular; and s_d is a scaling parameter to ensure reasonable acceptance rates of the proposed states.

The adaptive Metropolis algorithm is flexible and computationally straightforward. It is easily implemented in models of high dimension. The use of the parameters' covariance matrix ensures reasonable acceptance rates between highly correlated and interdependent parameters (which are often prevalent in conceptual rainfall-runoff models).

3.3. Algorithm for Sampling the Marginal Likelihood

The adaptive Metropolis algorithm was used to obtain the posterior distributions for parameters in all the models considered. Initially, a full run of 100,000 iterations was performed to obtain the approximate parameter values with maximum posterior density. Using the method provided by Gelman and Rubin (1992) it was determined that convergence was obtained after 20,000 iterations. Based on this, after a warm up run of 30,000 iterations the covariance was kept fixed (rather than updated at each iteration) and this fixed covariance was used to obtain 70,000 further draws from the posterior distributions. Estimation of the model's marginal likelihood then proceeded as:

1. For g = 1,…, K (where K = 50,000 for our study)
   - Sample $\theta^{(g)}$ from the posterior distribution (using the adaptive Metropolis algorithm).
   - Calculate $\alpha(\theta^{(g)}, \theta^* \mid y)$, the probability of accepting the move from $\theta^{(g)}$ to θ*.
   - Calculate $q(\theta^{(g)}, \theta^* \mid y)$, the proposal density at $\theta^{(g)}$.
2. For j = 1,…, J (where J=50,000 for our study)
   - Sample $\theta^{(j)}$ from q(θ*,.θ|y) (given the fixed θ*).
   - Calculate $\alpha(\theta^*, \theta^{(j)} \mid y)$, the probability of accepting the move from θ* to $\theta^{(j)}$.
3. Calculate the posterior ordinate using (4), and substitute into the marginal likelihood identity (3)
4. To compare models, estimate the model's posterior odds ratio using (1).

4. CASE STUDY: AUSTRALIAN WATER BALANCE MODEL AND THE BASS RIVER WATERSHED

The selected study area was the Bass River watershed, a 52 km^2 catchment located at Loch in the South Gippsland Basin on the western slopes of the Strzelecki Ranges. Data was available in the form of daily rainfall, evapotranspiration and runoff data, over a period of eleven years.

4.1. Model Selected for Comparison

The study builds on previous work in which the parameters of the Australian Water Balance Model (AWBM) were estimated using computations carried out via Markov chain Monte Carlo methods (Bates and Campbell 2001). The AWBM (Figure 1) is an eight parameter saturation overland flow model that was developed by Boughton (1993) to compute daily runoff from daily rainfall and evapotranspiration records. The model consists of three surface storages that are associated with three fractional areas to represent the variability of moisture capacity over the catchment. In a catchment, the surface storage capacity can vary considerably over the catchment area. The use of three surface storages to represent the catchment variability may not be appropriate for a specific catchment. In this study, a method is used to consider the variability of soil moisture over the Bass River catchment by formulating the model to vary the number of surface storages. In addition to the traditional model (Figure 1) in which three surface stores represent the variability of soil moisture capacity, the model was formulated to include a total of two or four surface stores.

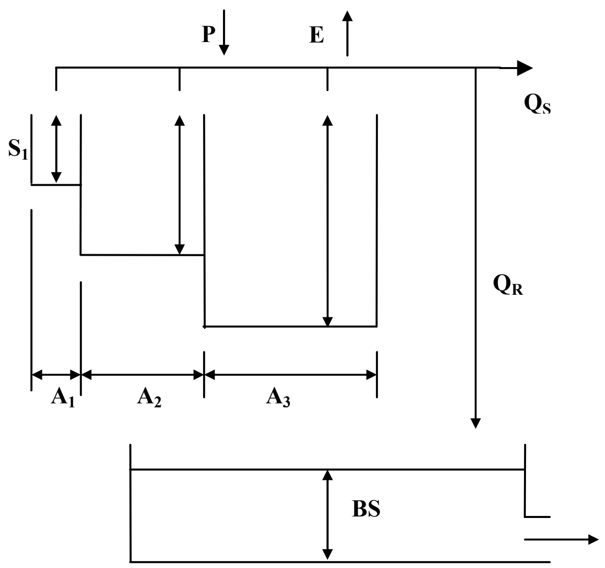

Figure 1. Australian Water Balance Model

Two different forms of the likelihood function were used to estimate the posterior distributions. The first assumed normally, independently and identically distributed data. The second assumed heteroscedastic, correlated error terms by applying a Box Cox transformation and fitting an autoregressive (order 4) error model. For each AWBM structure, both likelihood functions were considered, giving a total of 6 different models to be considered. It would be expected for hydrological models of this type that the assumption of normally, independently and identically distributed data would not be appropriate, considering the dominance of low or zero flows. The soil moisture accounting method by which the daily flow is calculated indicates that the assumption of correlated error terms would be more appropriate. The extent of the likely superiority of the likelihood with correlated errors is not known. This study yields a measure of this superiority.

4.2. Prior Distributions

At this stage, priors are specified based on our knowledge and experience in working with the model variables. Where there is little prior knowledge, priors should be flat in the region where the likelihood is large, so their impact on the Bayes factor is small (Kass and Raftery 1995). Priors were chosen in reference to those used by Bates and Campbell (2001). Further work will examine the results of the study if non-informative or conditional priors are assumed.

4.3. Approach Taken

Two sets of data were used in the study. To assess the model selection method in a controlled setting, artificial runoff data were created corresponding to the two storage model and then used to check if the method would select the 2 store model. The method was then applied to real catchment runoff data to determine which model configuration best represents the catchment.

The Bayesian information criterion (BIC) is an asymptotic approximation to the marginal likelihood. To confirm the results obtained from our MCMC scheme, the BIC was calculated for each model.

5. RESULTS: SIMULATED DATA STUDY

To obtain the 2 storage synthetic data, the Bass River rainfall data was used to simulate runoff with the AWBM. A Box-Cox transform was applied and a normally distributed random error term added. Considering the transformed data, the 2, 3 and 4 storage models were compared using a likelihood function that assumed heteroscedastic, correlated error terms. The resulting estimated marginal likelihood and BIC are given in Table 1.

Table 1: Results Using Simulated Data

Model	Log Marginal Likelihood	BIC
2 Storage Model	4730	4710
3 Storage Model	4053	3992
4 Storage Model	3131	4233

6. RESULTS: BASS RIVER CATCHMENT DATA

Using the real runoff data from the Bass River catchment, the 6 main models were compared to determine the relative benefits of modeling the data errors as heteroscedastic and correlated, and to examine the effect of assuming a different number of partial areas for runoff. The resulting estimated marginal likelihood and BIC of each model are given in Table 2.

Table 2: Results Using Bass River Data

Model	Marginal Log-likelihood	BIC
2 Storage Model, Independent Errors	-6821	-6774
3 Storage Model, Independent Errors	-6649	-6590
4 Storage Model, Independent Errors	-6633	-6566
2 Storage Model, Correlated Errors	-602	-535
3 Storage Model, Correlated Errors	-491	-410
4 Storage Model, Correlated Errors	-413	-403

7. DISCUSSION

7.1. Comparison of Results from Simulations

The above marginal log likelihoods show that the models with correlated error functions and transformed flow overwhelmingly outperform the models that assume independent errors. This reflects that the assumption of normally distributed, independent errors is not appropriate. The data shows a dominance of low and zero flows with fewer high runoff events, meaning that normally distributed errors are not likely. The results show the benefit of applying a transformation to the data. Similarly, the soil moisture accounting method within the model means that errors are likely to be correlated in time.

Within each likelihood group, the 4 storage models perform better than the 3 and 2 storage models. (The extent of difference in the marginal likelihood between the models means that calculating the pairwise Bayes factors is redundant). It would be likely that introduction of further storages could improve the model results. Over-parameterization is a continual problem in model selection and model fitting. Bayesian methods naturally penalize model complexity (Berger and Pericchi, 2001), which helps to avoid this problem.

7.2. Bayes Factors and the Benefit of Bayesian Model Selection

Bayesian versus non-Bayesian Selection Methods

A frequent issue in applying hydrological models is the paucity of data for many catchments. This lack of current, error-free and complete data means that parameters in models can be poorly identified. A single, unique set of values for models can be difficult, if not impossible, to obtain. One way in which to handle these issues is to account for the uncertainty about the model parameters when comparing models. Bayesian methods automatically account for both model and parameter uncertainty.

As mentioned, Bayesian methods favour simpler models, by naturally penalizing model complexity. With limited data available for many catchments, models with simpler parameterization should yield accurate results, and will be less complicated to implement. However, with growing computing power, there is a trend to develop increasingly complex models that allow the modeller to address the numerous dynamics within a catchment. The problem that lies in using complex models is that of poor parameter identifiability. Bayesian methods naturally account for these problems by favouring less complex models.

Bayesian methods also provide a way of incorporating other information. Criterion based methods don't allow prior choice to be taken into account when comparing models. By specifying the prior model probability, Bayes factors provide a way of including any specialist knowledge.

Finally, the method of comparing models via their posterior probabilities is flexible and simple. It allows many models to be compared without a change in method. There are no limits on the types of models being selected, and it is not necessary to use standard distributions.

Evaluation of Chib/Jeliazkov Method

The widespread use of Bayes factors has been hindered by the computational difficulties in accurately calculating the marginal likelihood of a model. In rare cases it is possible to determine the marginal likelihood analytically. In lower dimensional models, it may be possible to use ordinary Monte Carlo simulation. However, methods can become complicated in higher dimensional models. One of the main benefits of the methodology proposed by Chib and Jeliazkov (2001) lies in its flexibility and wide applicability. It can be used for high dimensional models and for comparing models of different dimension or structure. Little extra computational effort and programming is necessary than that required for a full MCMC run for sampling from the posterior.

Any concerns with the method are due to the issues generally associated with applying any MCMC techniques. It can be difficult to implement MCMC methods for complex or high dimensional models. Of particular importance is the rate of mixing and efficiency of the method in sampling the posterior. The convergence to the posterior distribution should also be considered. How to determine when to stop sampling from the posterior is a critical issue to those implementing MCMC schemes, and should be considered in this method of calculation of the marginal likelihood.

Why Not Bayes Factors? Problems in Successful Implementation of Bayes Factors

The sensitivity of results to the choice of prior for parameters is an important issue in calculating Bayes factors. It can often be difficult to specify meaningful priors for all model parameters. When estimating posterior distributions, vague priors are chosen to limit the effect on resulting posterior distributions given enough data. In testing models, the Bayes factor tends to be more sensitive to the choice of priors than in estimating posteriors (Kass and Raftery, 1995). Non-informative priors are also problematic in comparing posterior probabilities. In addition, Bayesian methods require specification of priors on all parameters for each model and this can be difficult when considering many high dimensional models.

Bayesian methods rely on the use of probability models for prior distributions and in specification of a likelihood function. This can affect results when estimating Bayes factors, and all assumptions should be checked.

8. CONCLUSIONS

When using a hydrological model it can be difficult to choose from the variety of models that exist. The range available to the practicing hydrologist means that choosing the best model, depending on the aim of the modelling exercise, is a complex task. This difficulty can be amplified by the uncertainty surrounding the parameters of the model.

Bayesian methods can provide an ideal means to compare competing models whilst allowing for model uncertainty. In comparing two models, the *Bayes factor* may be calculated, which is the posterior probability ratio of the models (assuming equal prior probabilities). Calculation of Bayes factors is complicated by the computational effort required, particularly for high dimensional models. Chib and Jeliazkov (2001) have provided an ideal solution that uses the sampled values from an MCMC chain to calculate the marginal likelihood.

Assessment of hydrological models in a Bayesian setting has been hindered by model complexity and parameter interaction. This has led to difficulty in developing MCMC regimes that may be applied to models of different structure, size and constraints. The method provided here is easily implemented. Calculation of the marginal likelihood is straightforward and computationally undemanding when a MCMC scheme is available. When combined with the flexible adaptive Metropolis algorithm the method described offers a powerful framework for Bayesian model comparison.

In this study, the method was applied to investigate the effects of using a different number of surface storages to represent partial area runoff. The AWBM was reformulated to consider 2, 3 or 4 surface storages. Two different likelihood models were considered for the data. The results showed that data from the Bass River catchment was better modelled as having heteroscedastic, time correlated errors. The best model assumed 4 surface storages. The methodology was confirmed by simulating data from the 2 storage model and ensuring that the corresponding model was selected.

9. REFERENCES

Bates, B. C., and E. Campbell. A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resources Research 37*(4), 937-947, 2001.

Berger, J. O., and L. R. Pericchi. Objective Bayesian methods for model selection: introduction and comparison, in *Model Selection,* P. Lahirini (ed.), Ohio, pp. 135-207, 2001.

Boughton, W. C. A hydrograph based model for estimating the water yield of ungauged catchments, *Nat. Conf. Publ. Inst. Eng. 93*(14), 317-324,1993.

Chib, S., and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output, *Journal of the American Statistical Association 96*, 270-281, 2001.

Gelman, A., and D. B. Rubin. Inference from iterative simulation using multiple sequences, *Statistical Science 7*(4), 457-511, 1992.

Haario, H., E. Saksmen, and J. Tamminen. An adaptive Metropolis algorithm, *Bernoulli 7*, 223-242, 2001.

Kass, R. E., and A. E. Raftery. Bayes Factors, *Journal of the American Statistical Association, 90*, 773-795, 1995.