Mixtures of experts approaches in rainfall-runoff modelling

Lucy Marshall
Department of Land Resources and Environmental Sciences
Montana State University

Ashish Sharma
School of Civil and Environmental Engineering
University of New South Wales

David Nott
Department of Statistics and Applied Probability
National University of Singapore

Outline

1. Quantifying predictive uncertainty in rainfall-runoff modelling.
2. Bayesian inference and computation.
3. Adaptive Monte Carlo methods.
4. Model uncertainty and model choice.
5. Mixtures of experts models.

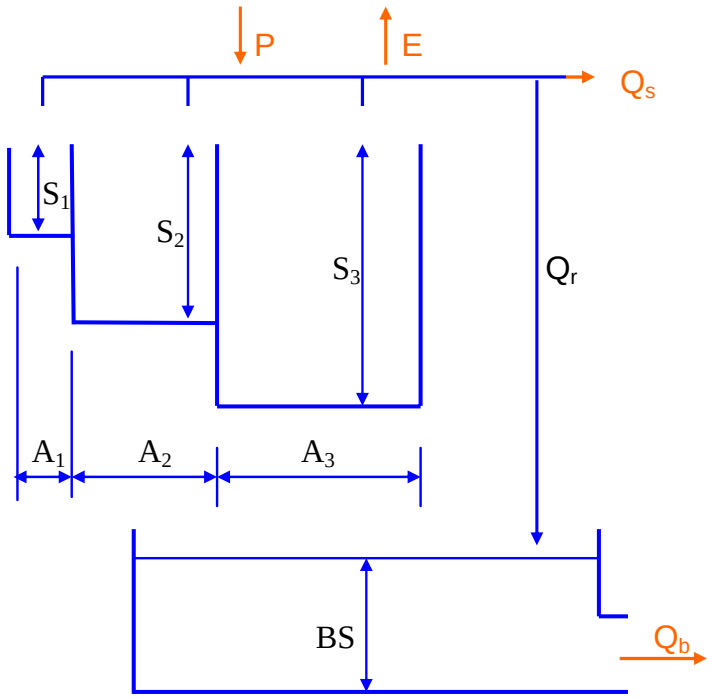
Rainfall-runoff models

Used for simulating catchment processes in response to a rainfall event.

Sources of uncertainty (Butts *et al.*, 2004, Journal of Hydrology):

1. Random or systematic errors in model inputs (boundary or initial conditions).
2. Random or systematic errors in the recorded output data.
3. Parameter uncertainty.
4. Model uncertainty.

Australian Water Balance Model (AWBM). Boughton (2004).



Quantifying uncertainty

Estimation of parameters (parameter calibration) in rainfall-runoff modelling is done by minimizing a response surface which may or may not be derived from a statistical model and a likelihood function.

The likelihood function is often complex with multiple modes. Bayesian approaches offer the possibility of rich inference, of exploring different interpretations of the data and using prior knowledge of hydrologists.

Traditional calibration in a statistical framework

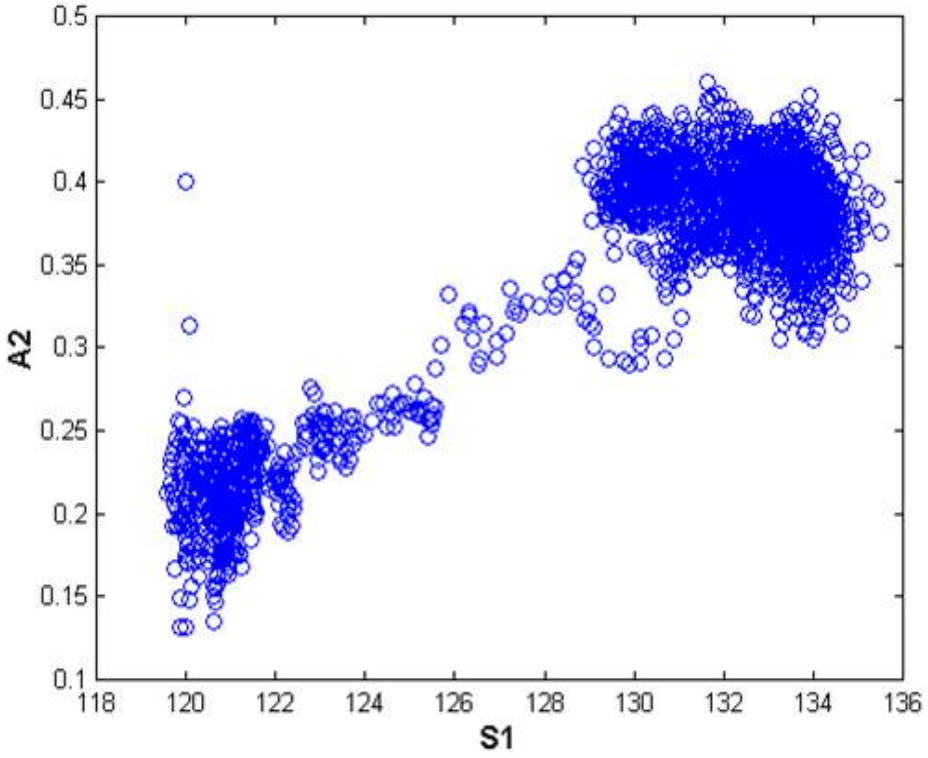
Model:

$$Q_t = f(x_t; \theta) + \epsilon_t$$

where Q_t is the response, x_t is the model inputs, θ is model parameters, $f(x_t; \theta)$ is the output of the runoff model for the given inputs and parameters, and ϵ_t is unexplained variation.

Usually ϵ_t is considered to have zero mean. We don't distinguish, for example, between measurement error and model inadequacy.

Quantifying uncertainty



Model uncertainty

The choice of model for a given application is a difficult problem.

"One is left with the view that the state of water resources modelling is like an economy subject to inflation – that there are too many models chasing (as yet) too few applications; that there are too many modellers chasing too few ideas; and that the response is to print ever-increasing quantities of paper, thereby devaluing the currency ..." (Robin Clarke, 1974).

Model uncertainty

Different models can provide much the same fit to the data. Model comparison in rainfall-runoff modelling as in other fields involves many considerations:

- Hydrologists have prior knowledge of how a catchment works, and what models capture the important physical processes.
- The use of the most physically realistic model may not be possible because of the data requirements of the model.
- Model choice should take into account the purpose for which the model is required.

Bayesian statistics

Bayesian statistics is distinguished by the use of probability for quantifying all kinds of uncertainty.

Set of unknowns θ to learn about, data y.

Specify a full probability model for the data and unknowns

$$p(y, \theta) = p(\theta)p(y|\theta)$$

$p(\theta)$ is called the prior distribution and $p(y|\theta)$ is the likelihood function. The prior codes in probabilistic form what we know about the unknowns before observing data, and gives the opportunity for use of prior knowledge.

Bayesian statistics

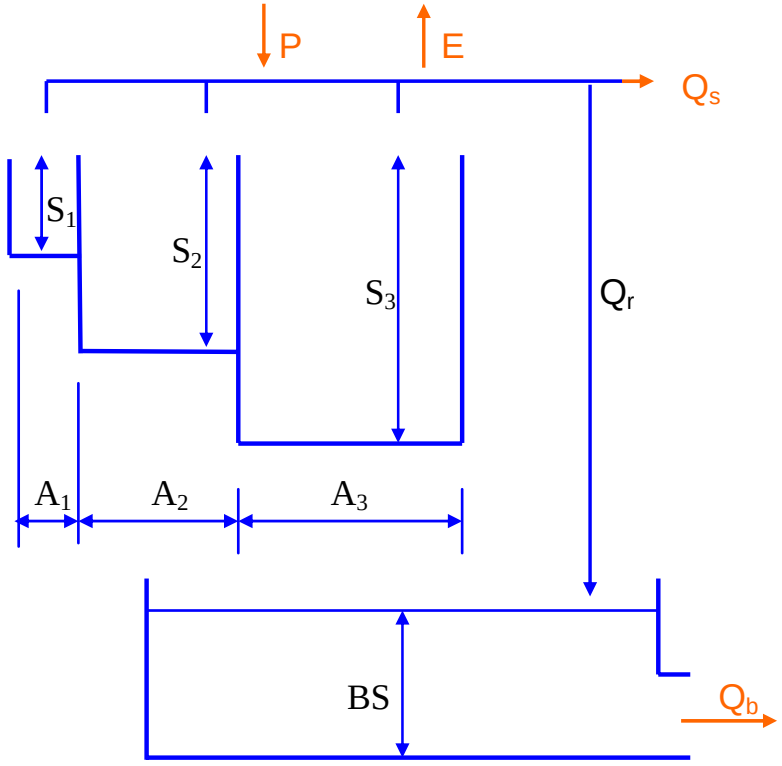
Conditioning on the observed data in this model we get

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

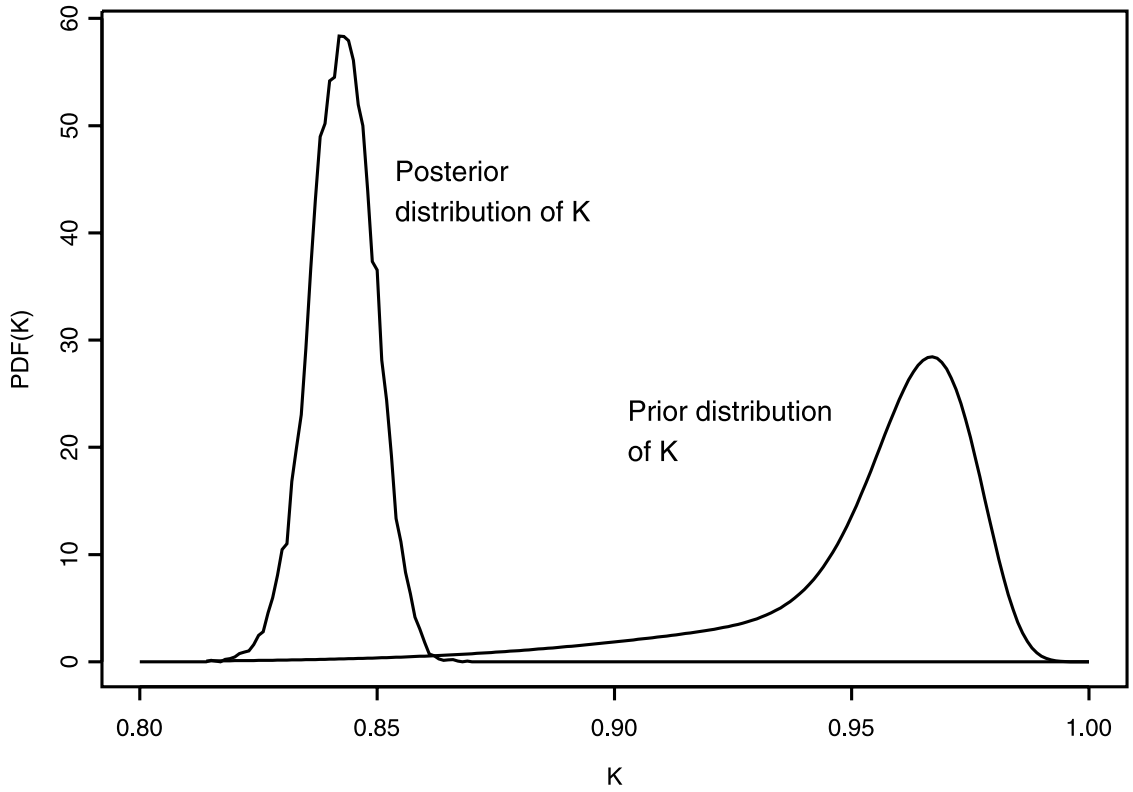
Here $p(\theta|y)$ is the posterior distribution expressing what we know about θ given the data y.

Inference is from the posterior distribution. In summarizing the posterior (by calculating probabilities or expectations) an integration over the parameter space is needed.

Example: AWBM



Example: AWBM



Predictive inference

Suppose predictions of future data y^* are required.

Predictive inference is based on

$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)d\theta.$$

In Bayesian inference predictive distributions are often the basis for informal methods of model criticism. How we go about model criticism usually depends on what the model will be used for.

Non-Bayesian approaches

GLUE (generalized likelihood uncertainty estimation).

Sample from “prior distributions” and then weight parameters by “generalized likelihood” measure for averaging in predictive inference.

“This appears to lead quite naturally to a Bayesian approach ... This is the essence of the GLUE approach.” (Beven, 2000).

“Underlying the development of the likelihood functions used in maximum likelihood approaches is the idea that there is a *correct* model.” (Beven, 2000).

“All models are wrong but some are useful.” (attributed to George Box).

Bayesian computation

In complex problems, posterior summarization is usually done with Markov chain Monte Carlo (MCMC) methods.

We have a posterior distribution $p(\theta|y)$. Construct a certain kind of random process (a Markov chain)

$$\{\theta^{(n)}; n = 0, 1, \ldots\}$$

so that the "long run" or equilibrium distribution of the process is $p(\theta|y)$.

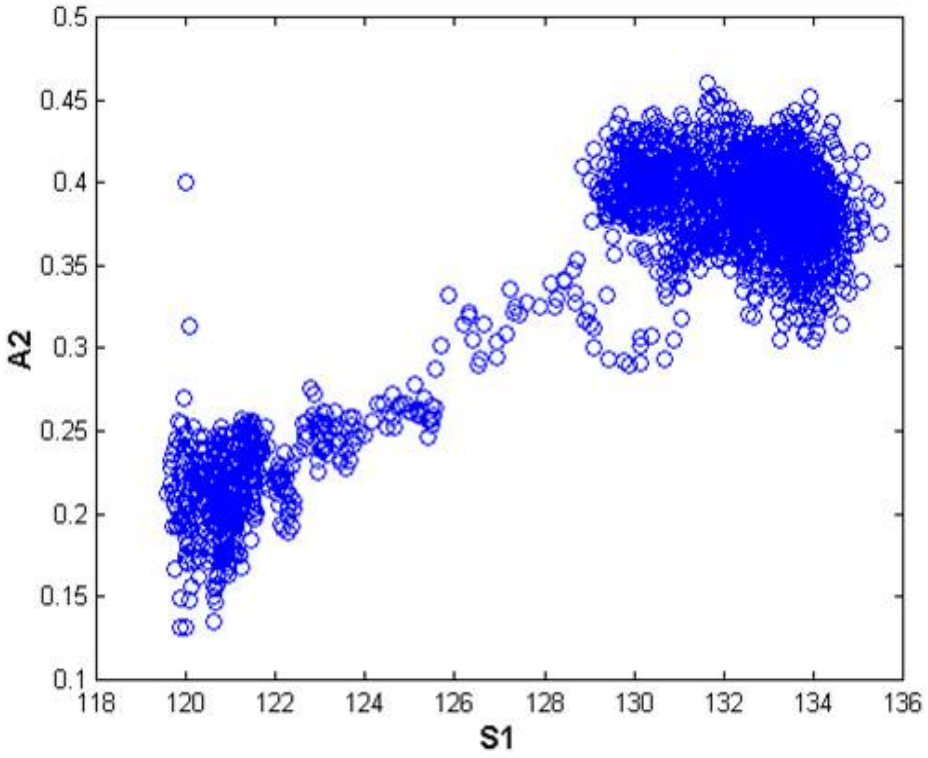
That is, the distribution of $\theta^{(n)}$ for n large is approximately $p(\theta|y)$.

Bayesian statistics

If we have such a process, we can simulate the process starting from a more or less arbitrary $\theta^{(0)}$ and then

- Wait until n is large enough that $\theta^{(n)}$ is approximately distributed as $p(\theta|y)$.
- Discard the initial period of simulations before time n (the "burn in" period).
- Take samples from the process after time n as a (correlated) sample from $p(\theta|y)$.

Bayesian statistics



Bayesian statistics

Suppose that $\theta^{(1)}, ...\theta^{(n)}$ represent a sample from the posterior distribution (the "burn in" period has already been discarded).

Then in predictive inference for future data y^*

$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)d\theta \approx \frac{1}{n}\sum_{i=1}^{n} p(y^*|\theta^{(i)}).$$

So an average of plug in predictive densities averaging over the posterior distribution for the parameters is used for predictive inferences and accounting for uncertainty in the Bayesian approach.

Bayesian statistics

Difficulties with MCMC:

- There isn't just one way to construct a Markov chain with given target posterior as the stationary distribution.
- Our "general recipes" for constructing chains with a given target posterior require some choices and tuning that are crucial to performance.
- When has convergence been reached, and how do we summarize the output?

Bayesian analysis of the AWBM

A Bayesian analysis of the Australian water balance model was given by Bates and Campbell (2001), *Water Resources Research*.

Development of a satisfactory MCMC sampling scheme is challenging – Bates and Campbell use a scheme that updates parameters one at a time or in small blocks, and the proposal steps that are made in their algorithm come from distributions that require preliminary runs for tuning.

There are order constraints among parameters and strong dependencies in the posterior distribution. How can computation be made more automatic?

Metropolis-Hastings algorithm

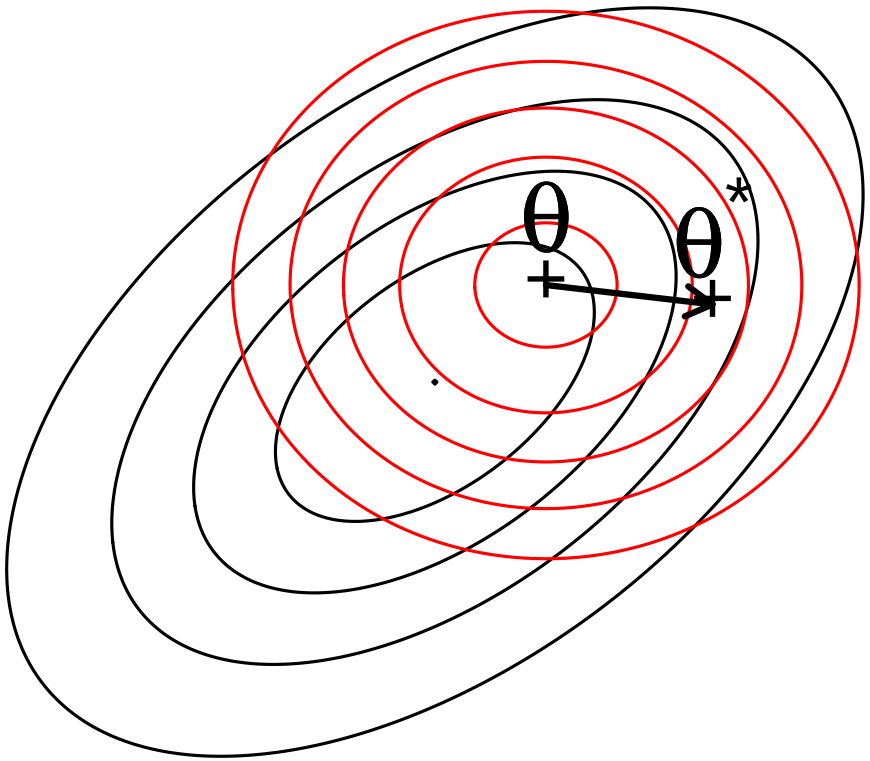
Almost all MCMC sampling schemes are variants of the Metropolis-Hastings algorithm.

Transitions are made by proposal of a new state θ^* given the current state θ according to a proposal distribution $q(\theta^*|\theta)$. Proposal is accepted with probability

$$\min\left\{1, \frac{p(\theta^*|y)q(\theta|\theta^*)}{p(\theta|y)q(\theta^*|\theta)}\right\}$$

otherwise θ is retained.

Metropolis-Hastings algorithm



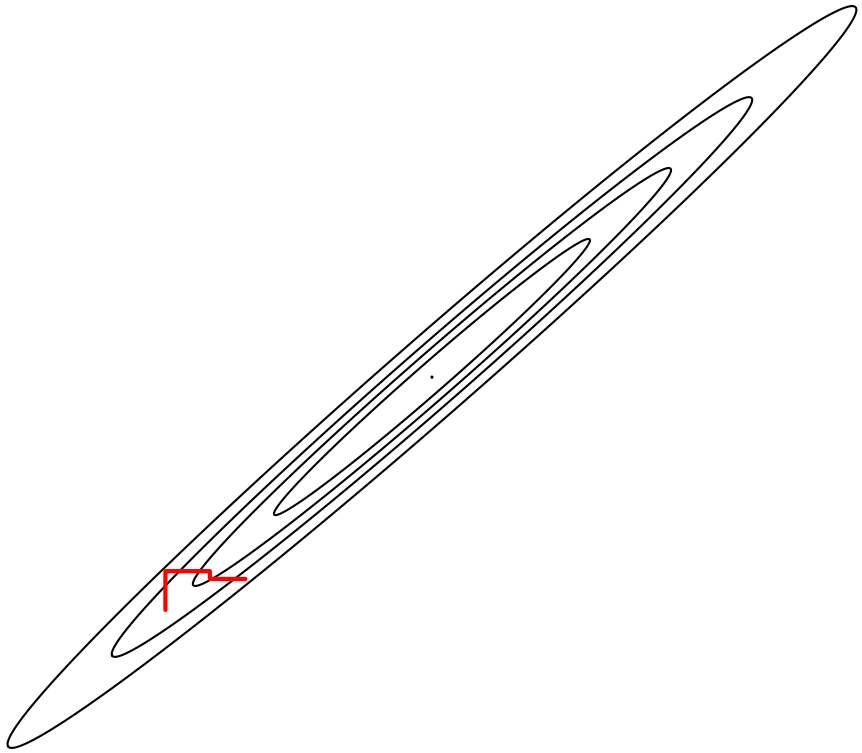
Metropolis-Hastings algorithm

The accept/reject rule maintains the desired stationary distribution.

There is an almost unrestricted choice for the proposal distribution to perturb the current state at each step.

The choice of proposal is crucial to performance.

Updates based on cycles



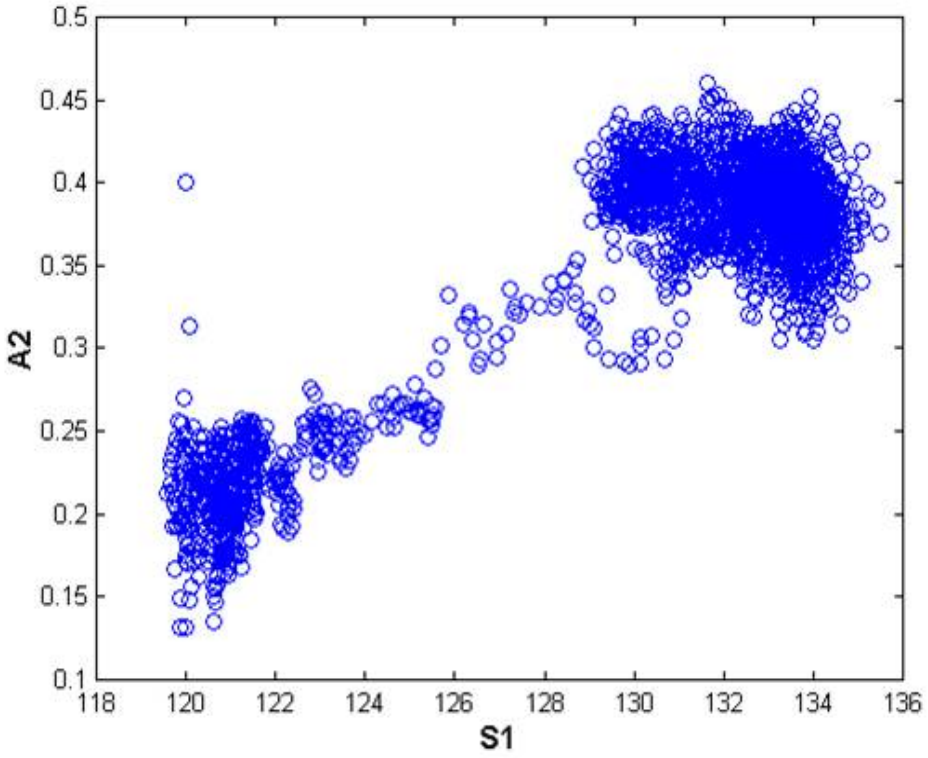
Updates based on cycles

Strong dependence between components of the posterior means that "one at a time" updating strategies may not work well.

Fixing all components but one may strongly constrain the free component under strong dependence.

Well separated modes in the target posterior can also be problematic.

Australian water balance model



Block updating with random walk Metropolis

Random walk Metropolis with a normal proposal: generate θ^* by perturbing the current θ by a normal variate.

$$\theta^* \sim N(\theta, C).$$

A poor choice of the proposal covariance C gives poor performance.

Adaptive Monte Carlo

In generating $\{\theta^{(n)}; n \geq 0\}$ allow the proposal covariance at time n to depend on n and $\theta^{(1)}, ..., \theta^{(n-1)}$ (Haario, Saksman and Tamminen, 2001, *Bernoulli*).

This is no longer Markov chain Monte Carlo (but it still works, subject to some fine print).

The important thing is to adapt the proposal increasingly slowly as time goes on (Nott and Kohn, 2005, *Biometrika*).

Adaptive Monte Carlo

Adaptive random walk proposal covariance (Haario, Saksman and Tamminen, 2001). Choose $C = C^{(n)}$ as

$$C^{(n)} = \begin{cases} C_0 & n \leq n_0 \\ s\text{Cov}(\theta^{(1)}, ..., \theta^{(n-1)}) + \epsilon I & n > n_0 \end{cases}$$

where C_0 is an initial covariance which isn't adapted for the first n_0 iterations, s is a constant depending on the parameter dimension and ϵ is a small positive constant.

MCMC

Adaptive MCMC

Comparative performance of MCMC schemes

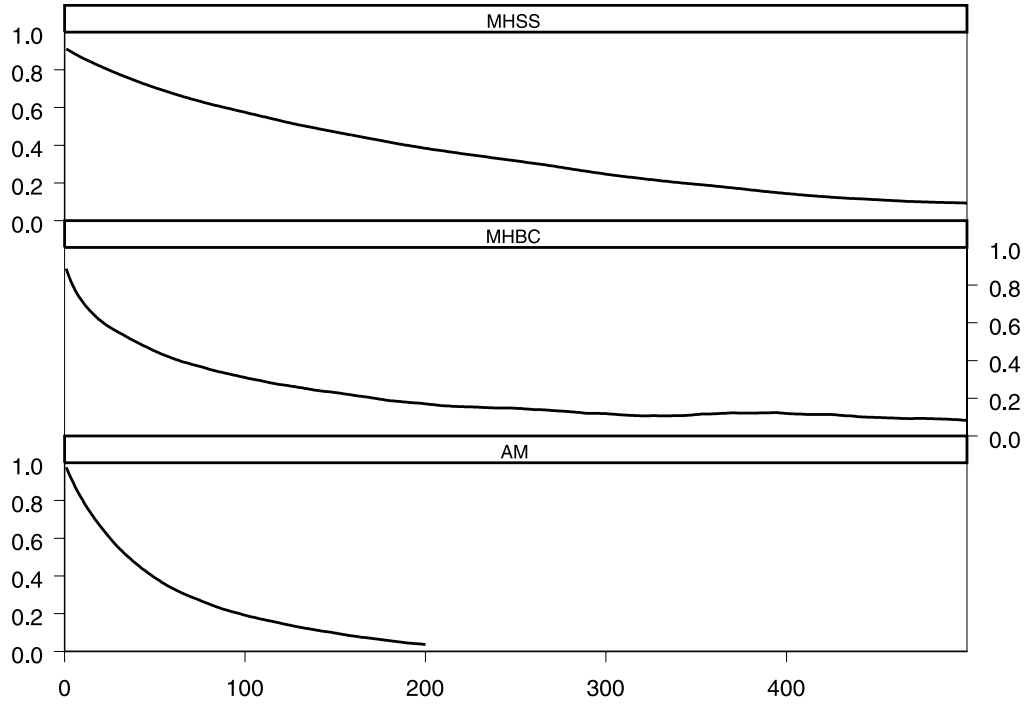
Bayesian analysis of the AWBM done by Bates and Campbell (2001), Water Resources Research.

Their MCMC scheme relies on one at a time and low-dimensional block updates.

The ideas behind their sampling scheme are not easily generalized to other models.

Comparison of their scheme with adaptive MCMC (Marshall, Nott and Sharma, 2002, *Water Resources Research*)

Autocorrelation functions



Model uncertainty in rainfall-runoff modelling

Characteristics of the problem of model choice in rainfall-runoff modelling:

1. The physical understanding of the hydrologist about how the catchment works should play an important role in model choice.

2. The complexity of the model that can be used will be limited by the data available to drive the model.

3. The use that will be made of the model should play an important role in model choice

Model uncertainty: the Bayesian approach

Bayesians (usually) describe model uncertainty probabilistically:

$$p(M|y) \propto p(M)p(y|M)$$

where

$$p(y|M) = \int p(y|\theta_M, M)p(\theta_M|M)d\theta_M.$$

The prior distribution on the model $p(M)$ and the prior distributions on model parameters $p(\theta_M|M)$ allow incorporation of prior knowledge.

Model uncertainty: the Bayesian approach

Model uncertainty is accounted for in predictive inference about future data y^* (or other predictive quantity of interest):

$$p(y^*|y) = \sum_M p(M|y)p(y^*|M,y).$$

Here predictive distributions from individual models $p(y^*|M,y)$ are combined by weighting according to posterior model probability.

This kind of predictive inference goes by the name of Bayesian model averaging.

Model uncertainty: the Bayesian approach

Computation of the model marginal likelihoods $p(y|M)$ is difficult.

MCMC methods have been developed (Green, 1995, Biometrika) that can traverse the model and parameter space jointly.

Model uncertainty: the Bayesian approach

In rainfall-runoff modelling applications if the number of models to be compared is small, an approach which estimates marginal likelihoods based on separate MCMC runs for different models is simple, reliable and more automatic.

Suppress conditioning on M in what follows writing $p(y) = p(y|M)$.

Model uncertainty: the Bayesian approach

Chib and Jeliazkov (2001), *J. Amer. Statist. Assoc.*

Rearranging

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}$$

we obtain

$$\log p(y) = \log p(y|\theta) + \log p(y|\theta) - \log p(\theta)$$

This formula holds for any value of θ.

Model uncertainty: the Bayesian approach

Evaluate the right hand side of this last formula at the posterior mode $\hat{\theta}$.

The only term difficult to compute is $\log p(\hat{\theta}|y, M)$.

Model uncertainty: the Bayesian approach

$$p(\hat{\theta}|y, M) = \frac{E_1(\alpha(\theta, \hat{\theta})q(\theta|\hat{\theta})}{E_2(\alpha(\hat{\theta}, \theta))}$$

where E_1 is the expectation with respect to the posterior distribution $p(\theta|y)$ and E_2 denotes expectation with respect to $q(\theta|\hat{\theta})$ and $\alpha(\theta, \theta')$ denotes the acceptance probability in a Metropolis-Hastings step from θ to θ'.

Model uncertainty: the Bayesian approach

With separate MCMC model runs for different models we can esitmate the marginal likelihood.

In hydrological applications there is usually only a small number of models to be compared. Methods which estimate marginal likelihoods directly based on separate MCMC runs may be more reliable than methods which explore the joint model and parameter space directly (Han and Carlin, 2001, *J. Amer. Statist. Assoc.*).

Model uncertainty: the Bayesian approach

Marshall, Sharma and Nott, 2006, *Water Resources Research*, consider the AWBM model formulated to have a variable number of soil moisture storages (1, 2 or 4 storages).

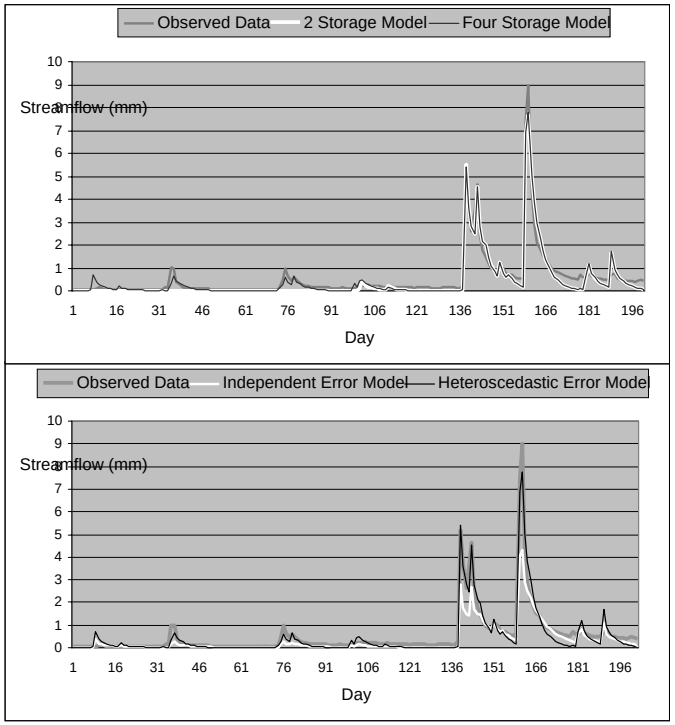
Model comparison: compare different numbers of storages, constant variance Gaussian errors versus heteroscedastic errors, independent versus autoregressive error structure.

Model uncertainty: the Bayesian approach

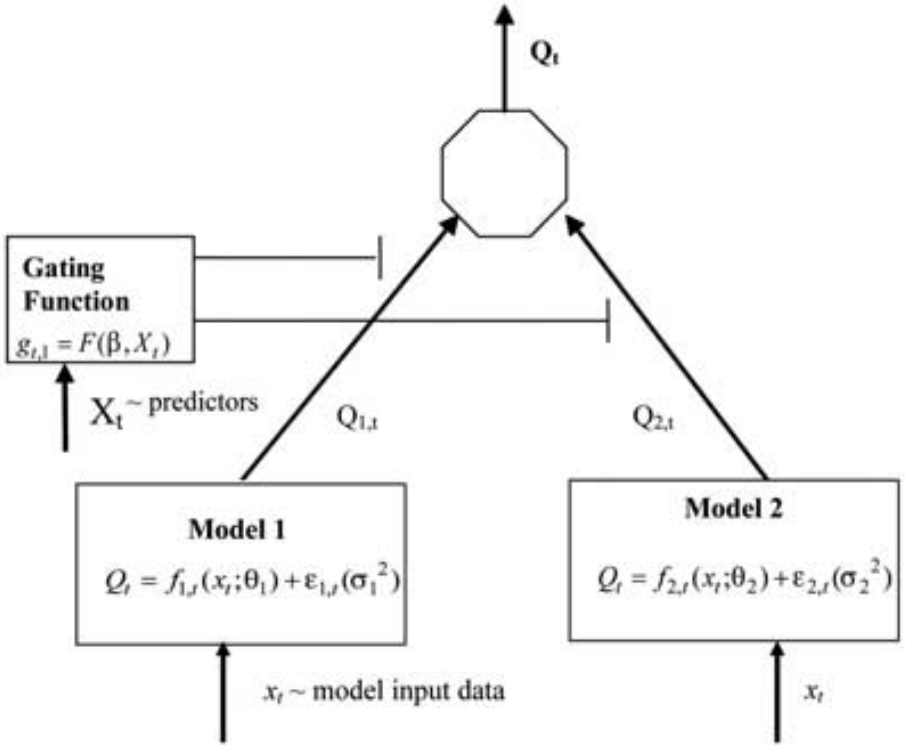
A formal Bayesian approach to model comparison is useful for incorporating prior knowledge and for penalizing model complexity with limited data.

However, informal diagnostics are important for revealing specific model inadequacies.

Model uncertainty: the Bayesian approach



Combining models with mixtures



Simple single level mixture of experts

Streamflow Q_t, runoff model inputs x_t, gating function predictors w_t (t indexes time, and the model runs on a daily time step).

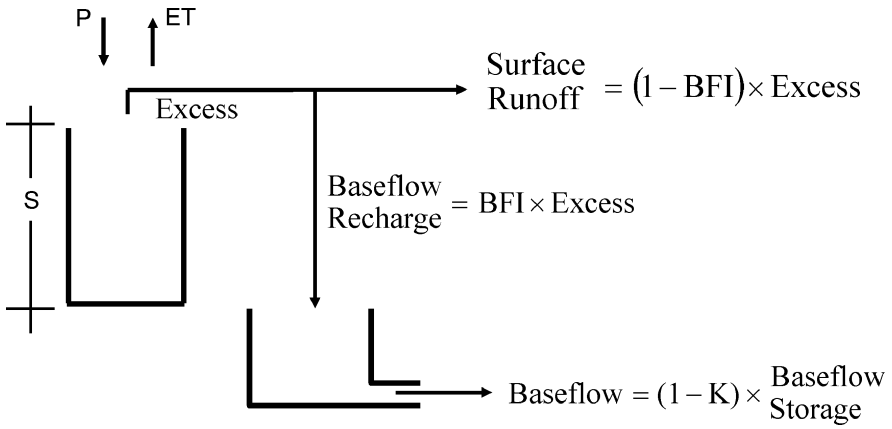
z_t - in my single level model with two experts z_t is a binary indicator for which "expert" model Q_t was generated from. z_t is unobserved.

Given $z_t = j$

$$Q_t = f_{j,t}(x_t; \theta_j) + \epsilon_{j,t}, \quad \epsilon_{j,t} \sim N(0, \sigma_j^2)$$

$$Pr(z_t = 1) = 1/(1 + \exp(-w_t^T \beta) \quad Pr(z_t = 0) = 1 - Pr(z_t = 1).$$

Simplified AWBM

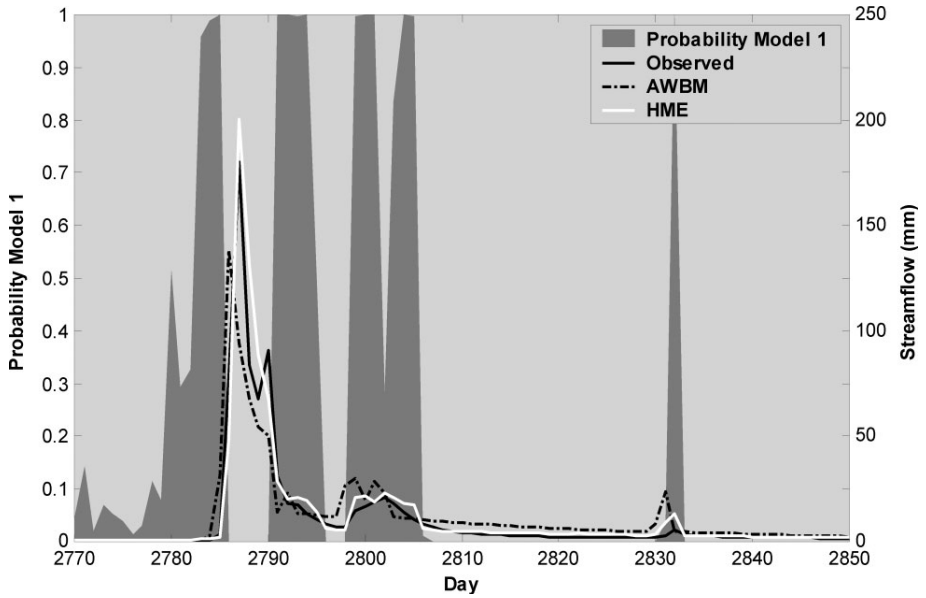


Simple single level mixture of experts

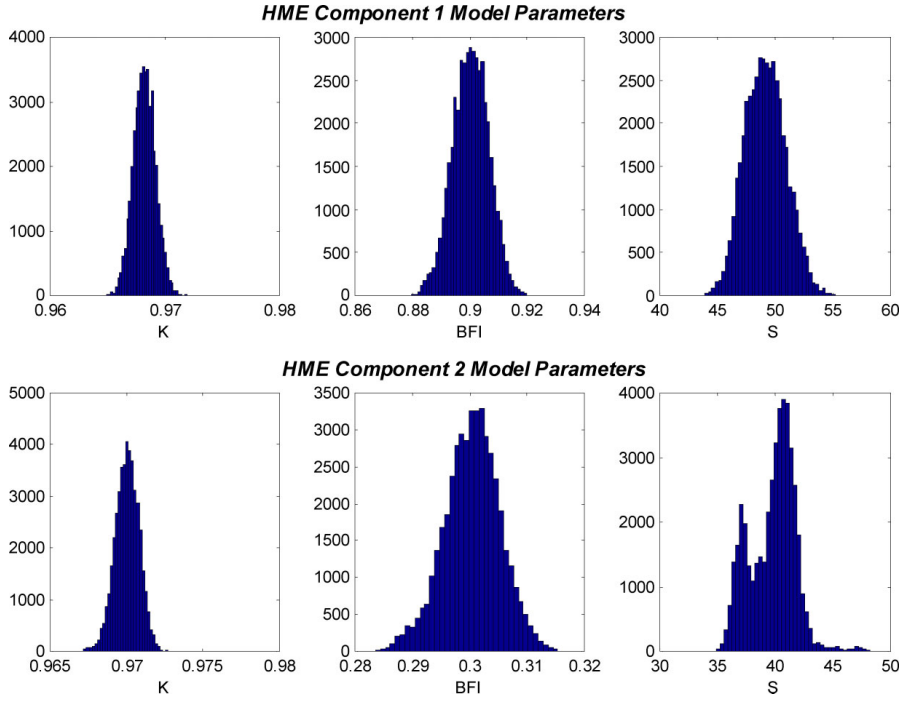
Marshall, Sharma and Nott (2007), *Hydrological Processes*, consider 10 catchments in Australia with different characteristics.

Results for one catchment (Never Never River at Glennifer Bridge):

Simple single level mixture of experts



Simple single level mixture of experts



Simple single level mixture of experts

One model is dominant at peak flow and in long dry spells.

The other model is dominant at other times.

There is a hydrological interpretation of the difference in parameters between the two component models and when they apply.

Simple single level mixture of experts

The mixture of experts fit and the parameter estimates in the expert component models is often revealing about what processes are not well captured by a single fixed model with the structure of one of the experts.

As well as being a predictive tool and a tool for accounting for complex data structure through mixing, the HME framework is also very useful for model criticism.

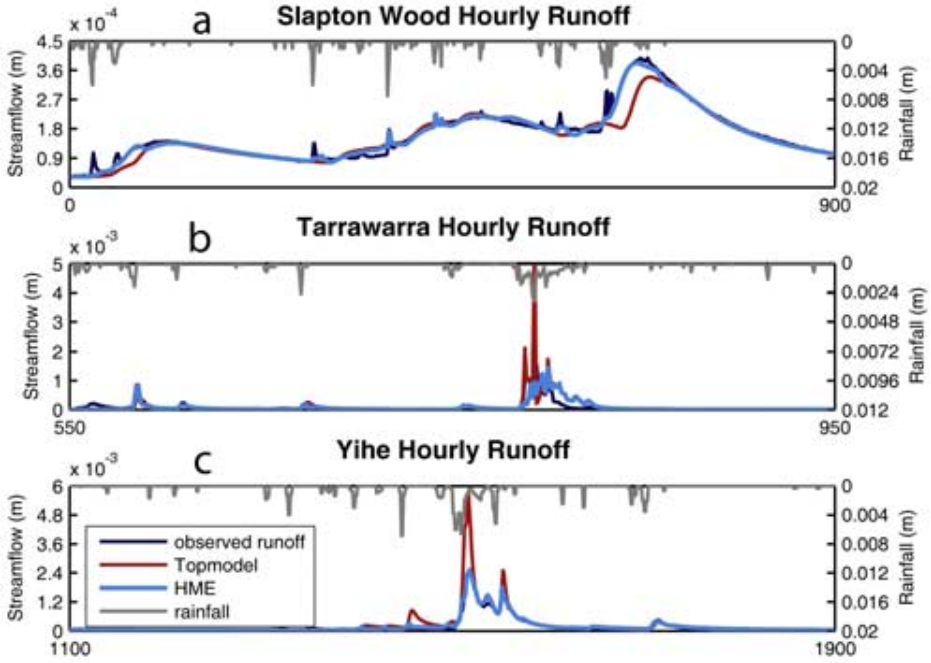
Simple single level mixture of experts

In the last case study the “expert” model was rather simple.

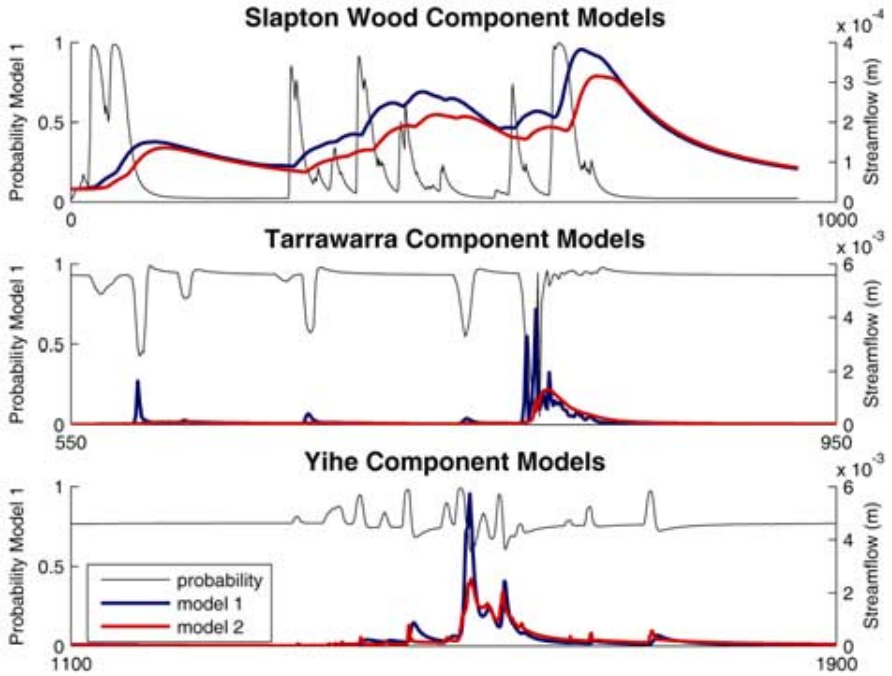
Case study where the expert model derives from TOPMODEL (Beven and Kirby, 1979, *Hydrol. Sci. Bull.*) Marshall, Sharma and Nott (2007) Geophysical Research Letters.

Single level mixture of two experts was considered with topmodel components for three catchments with very different characteristics.

Simple single level mixture of experts



Simple single level mixture of experts



Conclusion

Mixtures of experts models are a useful tool for prediction and for model criticism with a view to further model development.

Currently working on incorporating uncertainties on the model inputs into the modelling framework.