Rainfall-Runoff Model Calibration using Informal Likelihood Measures within a Markov Chain Monte Carlo Sampling Scheme

Hilary McMillan* and Martyn Clark

*h.mcmillan@niwa.co.nz

National Institute of Water and Atmospheric Research Ltd, 10 Kyle St, Riccarton, Christchurch, new Zealand

Abstract

This paper considers the calibration of a distributed rainfall-runoff model in a catchment where heterogeneous geology leads to a difficult and high-dimensional calibration problem, where the response surface has multiple optima and strong parameter interactions. These characteristics render the problem unsuitable for solution by uniform Monte Carlo sampling and require a more targeted sampling strategy. MCMC methods, using the SCEM-UA algorithm, are trialled using both formal and informal likelihood measures. Each method is assessed in its success at predicting the catchment flow response and capturing the total uncertainty associated with this prediction. The comparison is made at both the catchment outlet and at internal catchment locations with distinct geological characteristics. Lastly, we demonstrate how information gained from the exploration of the response space, in conjunction with qualitative knowledge of system behaviour, can be used to constrain the Markov Chain trajectory.

1 Introduction

Improving availability and coverage of spatial data has driven developments in distributed, process-based catchment modelling; however, despite the correspondence between modeled and observed processes, it is not usually possible to determine model parameters values directly from field measurements. Instead, the values required are those of the 'effective parameters' which represent integrated behaviour at the model element scale. These values must therefore be determined through a calibration methodology, via some search of the parameter space. As has been extensively discussed by Beven (1993; 2005; Beven and Binley, 1992) and others (Wagener and Gupta, 2005), the many sources of uncertainty in a hydrological model application lead to equifinality of parameter sets in providing acceptable model performance with reference to some observed data. These uncertainty sources may include, but are not limited to, input data uncertainty, initial condition uncertainty, model structural error, observed data uncertainty (Liu and Gupta, 2007). Indeed, since it is certain that our hydrological model does not fully represent the complexity of the natural catchment and is therefore 'wrong', we must expect that any calibration technique is a process of identifying some subset of model parameterisations which produce reasonable approximations to some aspects of true catchment behaviour under some circumstances.

The aim of a calibration technique should therefore be to enable an efficient search of the parameter space, identifying those regions where model performance is considered satisfactory. The task is made more difficult by the typically complex nature of the model response surface (Duan *et al.*, 1992; Sorooshian *et al.,* 1993) which may be exacerbated by artefacts of model timestep and solution techniques (Kavetski *et al.*, 2006a,b). Difficulties encountered may include multiple local optima in multiple regions of attraction, discontinuous derivatives, parameter interaction and flat areas (Duan *et al.*, 1992). The nature of these surfaces prohibits standard search mechanisms such as simplex- and Newton- type schemes. Alternative methods such as uniform random sampling suffer from a lack of sampling efficiency and can be extremely costly in terms of model evaluations. They also typically specify the sample space using minimum and maximum values for each parameter, based usually on expert judgement, physical interpretation of the parameter and previous model use. However with good model performance often occurring up to the boundary of the sample region, this technique may unjustifiably restrict the search.

In recent years, Markov Chain Monte Carlo (MCMC) methods have gained increasing popularity, in particular the Metropolis-Hastings (MH) Algorithm (e.g., Chib and Greenberg, 1995). These methods enable simulation of complex multivariate distributions by casting them as the invariant distribution of a Markov Chain. By finding an appropriate transition kernel which converges to this distribution, samples with the desired posterior distribution can be drawn from the Markov Chain. A popular version of the MH algorithm is the adaptive SCEM-UA algorithm (Vrugt *et al.*, 2003) which combines the MH sampler with the SCE-UA optimisation method (Duan *et al.,* 1992), using information exchange between multiple sampler chains to improve convergence rates.

All search techniques require a definition of the model response surface to be searched: this is usually couched in terms of ‘probability of model correctness given observed data’ and is assessed via a likelihood measure. The debate continues on the relative advantages of the informal likelihood measures used in the GLUE framework compared with parameter estimation via formal statistical likelihood estimation (e.g.

Mantovan and Todini, 2006; Beven *et al.*, 2007; Mantovan *et al.*, 2007; Thiemann *et al*, 2001; Beven, 2003; Gupta *et al.*, 2003; Clarke 1994). If statistical likelihood theory is to be used, the error model between model predictand and observed variable must be specified exactly; this may include information on heteroscedasticity and autocorrelation (e.g. Sorooshian, 1981; Sorooshian and Dracup, 1980) and may rely on hierarchical error models (Kuczera *et al.*, 2006). Under GLUE, the concept of a true model (and error model) against which to compare observations is rejected and it is accepted that many interacting sources of error, without well-defined formulations, combine to give total model error. Models are instead judged against informal likelihood measures, chosen by the hydrologist, which represent their expert perception of model performance in prediction of observed data (Beven, 2006).

Although MCMC methods have traditionally used formal likelihood measures to define the response surface (e.g. Arhonditsis *et al.*, 2008; Marshall *et al.*, 2004; Vrugt *et al.*, 2006; Vrugt *et al.*, 2003; Thiemann *et al.*, 2001), it is also possible to use informal likelihoods (e.g., Engeland and Gottschalk, 2002; Blasone et al., 2008; Vrugt et al., 2008). When informal likelihoods are used in MCMC methods, the main difference between MCMC methods and GLUE is that MCMC methods provide targeted sampling of the parameter space. Blasone et al. (2008) compared performance of the informal likelihoods in the SCEM-UA method with the traditional GLUE method and demonstrated that the targeted sampling resulted in better predictions of the model output (and that the uncertainty limits were less sensitive to the number of retained solutions). Vrugt et al. (2008) compared a formal Bayesian approach that attempts to explicitly quantify the individual sources of uncertainty in the hydrological modelling process with the traditional GLUE method that maps all sources of uncertainty onto the parameter space. They showed that while the estimates of total uncertainty were similar in both methods, the GLUE method produced large estimates of parameter uncertainty which can lead to erroneous conclusions on the identifiably of model parameters.

The formal Bayesian approaches for explicitly quantifying the individual sources of uncertainty suffer from two important limitations. First, as formulated by Vrugt et al. (2008) and Kavetski et al. (2006a; 2006b), the formal Bayesian methods require solving a high-dimensional optimization problem (i.e., separate multipliers for each storm); a problem that is intractable for distributed hydrological models where it is necessary to quantify uncertainty in the spatial pattern of precipitation events. Second, current methods for quantifying error in model structure are poorly developed—indeed, Vrugt et al. (2008) and Kavetski et al. (2006a; 2006b) essentially combine error in model inputs and model structure into a single error term. Informal likelihood measures therefore remain an attractive option.

This paper considers the calibration of a distributed rainfall-runoff model (described in Section 2.2) in an interesting case study catchment, the Rangitaiki in New Zealand (described in Section 2.1), where heterogeneous geology leads to a difficult and high-dimensional calibration problem, where the response surface has multiple optima and strong parameter interactions. These characteristics render the problem unsuitable for solution by uniform Monte Carlo sampling (as per standard GLUE) and require a more targeted sampling strategy. MCMC methods, using the SCEM-UA algorithm, are trialled using both formal (Section 3.1) and informal (Section 3.2) likelihood measures, and assessed in their success at full coverage of the response surface.

2 Model and Data

2.1 Catchment

The Rangitaiki River is located in the central North Island of New Zealand. It has a length of 155 km and mean flow in the lower reaches of around 25 m^3s^{-1}. The river flows along a series of fault-angle valleys which define a structural geological boundary. To the west are Quaternary volcanic rocks, comprising a series of partially overlapping, rhyolitic, welded ignimbrite sheets, overlain by thick tephra and pumice sequences; to the east are uplifted Jurassic basement greywackes and meta-greywackes (Beanland and Haines, 1998; Manville *et al.,* 2004). These two parts of the catchment have strikingly different hydrological regimes: the porous tephras have a characteristic high stable baseflow regime and subdued flood peaks; the steep and relatively impermeable qreywacke responds quickly to rainfall with a peaked runoff pattern.

Subcatchment Geology

Each subcatchment is classified according to its substrate geology as recorded in the New Zealand Land Resource Inventory (NZLRI). For the purposed of this study, a simple binary division was made between impermeable (greywacke, argillite, lava) and permeable (pumice, lapilli, tephra) geology. Although the two categories are broadly divided East and West of the Rangitaiki river in the upper catchment, there is some local variation (Figure 1).

2.2 Data

Gauging data for the Rangitaiki is available at Te Teko, at the entrance to the coastal Rangitaiki Plains. The gauging station has a catchment area of 2890 km^2 and represents the combined flow of the pumice and greywacke areas: a relatively sustained baseflow is overlain by significant flood peaks. The contrasting subcatchment flow regimes can be compared through the discharge records of two internal gauging stations at Murupara and Galatea. Murupara is situated on the main branch of the Rangitaiki, with a catchment of 1140 km^2 of the Kaingaroa Plateau. The average annual mean flow is 21 m^3s^{-1} and the mean annual flood is 40 m^3s^{-1}. Galatea is sited on the Whirinaki, and drains a 509 km^2 area of the greywacke ranges. Here the average annual mean flow is 14.5 m^3s^{-1}, and the mean annual flood is 109 m^3s^{-1} (McKerchar and Pearson, 1989).

The model uses input precipitation and climate data from Tait *et al.* (2006) who interpolated data from over 500 climate stations in New Zealand across a regular 0.05° latitude-longitude grid (approximately 5 km * 5 km). These data are provided at daily time steps, and are disaggregated to hourly data before use in the model. In this study we use data from the year 1998 when a large flood event occurred in the Rangitaiki catchment, allowing a test of the model response over a full range of discharge magnitudes.

To apply TopNet in the Rangitaiki, TopNet requires information on catchment topography, physical and hydrological properties. This information is available from a variety of sources. The New Zealand River Environment Classification (REC; Snelder and Biggs, 2002) includes a digital network of approximately 600,000 river reaches and related sub-basins for New Zealand. A 30 m Digital Elevation Model (DEM) provided topographic properties. Land cover and soil data is available from the New Zealand Land Cover Database (LCDB) and the New Zealand Land Resource Inventory (LRI; Newsome *et al*., 2000). The river basin was first disaggregated into individual subcatchments, each one of which becomes a model element. We use the Strahler 3 subcatchments from the REC, which have a typical size of 10 km^2, and split the Rangitaiki Basin into 308 elements. The REC also provides the geometrical parameters of the river network. Frequency distributions of the topographic wetness index and distance to streams are calculated from the DEM. Average soil and landcover parameters are derived from the LRI and LCDB respectively. In total, 12 parameters are required for each subcatchment, of which 6 may be specified using the information described above; the remaining 6 must be calibrated (Refer to Table 1 for descriptions of all the parameters). In addition, the Manning's n value for the subcatchment channel section must also be calibrated.

2.3 TOPNET

TOPNET was developed by combining TOPMODEL (Beven et al., 1979; Beven et al., 1995), which is most suited to small watersheds, with a kinematic wave channel routing algorithm (Goring, 1994) so as to have a modeling system that can be applied over large watersheds using smaller sub-basins within the large watershed as model elements (Ibbitt and Woods, 2002; Bandaragoda *et al.*, 2004; Clark et al., 2008). TOPNET uses TOPMODEL concepts for the representation of sub-surface storage controlling the dynamics of the saturated contributing area and baseflow recession. To form a complete model, potential evapotranspiration, interception (based on the work of Ibbitt, 1971), infiltration (using a Green-Ampt mechanism; Mein and Larsen, 1973) and soil zone components were added. Kinematic wave routing moves the sub-basin inputs through the stream channel network. Complete model equations are provided by Clark et al. (2008) and are not repeated here.

2.4 Calibration via Parameter Multipliers

In distributed rainfall-runoff models, the calibration problem is greatly complicated by the large number of model parameters: multiple model parameters for each model spatial element. Experience suggests that the integrated variables typically available to evaluate model performance, such as streamflow series, may hold insufficient information to determine all model parameter values (Beven, 2001). Various approaches have been applied to ease this discrepancy. Many studies assume that several parameters are spatially constant over the model domain, using a value determined either by expert opinion or by directly using values measured at point locations. Another popular approach is to apply a set of "parameter multipliers" to *a-priori* model element parameter values, significantly reducing the dimensionality of the calibration problem (Clark et al., 2008). However, due to the reliance on a previously determined spatial distribution of model parameters, there is a danger that distributed hydrological models calibrated using integrated data such as catchment outlet discharge may fail to properly represent the range of hydrological behaviours. Poor forecasts would then be produced at internal catchment locations (Clark et al., 2008).

This paper presents a model calibration strategy that provides correct representation of internal catchment processes. The calibration method is applied in the Rangitaiki, where two sub-regions of the catchment have significantly different hydrological characteristics. Our knowledge of catchment geology cannot be translated directly into values for model parameters; instead we seek to use the qualitative information to inform our calibration strategy.

Figure 1: Geology of the Rangitaiki River basin, classified according to permeability. Gauging Locations are marked.

The method used is to classify each Strahler 3 sub-catchment as either ‘permeable’ or ‘impermeable’ (refer to Section 2.1; note that in other catchments, three or more qualitative categories may be appropriate). *A-priori* model parameters are specified in each individual subcatchment using topography, soils and land-cover data (Table 1). Two sets of parameter multipliers are then allowed, one for each category. The optimisation process allows all multipliers to be calibrated simultaneously, such that the optimum combination of process descriptions in the two categories is found.

The Rangitaiki provides an ideal test location, as the model calibration can be implemented using only the outlet discharge gauged at Te Teko (Figure 1), but tested for diverse internal process representation using the two gauges at Murupara (pumice subcatchment) and Galatea (greywacke catchment). This internal check allows a test of model conditioning and parameter identification success; an important consideration due to the increased number of parameters used with this method.

3 MCMC technique (Bayesian Uncertainty Framework)

3.1 Metropolis and Adaptive Metropolis Algorithms

Markov Chain Monte Carlo provides a general approach to sampling from the posterior distribution. Classical Markov Chain theory specifies the transition kernel P(x,A) which gives the probability from moving from the point x to any point in the set A. A common question is then to determine whether the chain has an invariant distribution π which is unchanged by applying the transition kernel. The MCMC technique reverses the problem: the required posterior distribution is taken as the invariant π; instead we seek the appropriate transition kernel P(x,A) such that a chain using this kernel provides samples from the posterior. The Metropolis-Hastings algorithm, one of the most popular MCMC methods, provides a method for finding the required transition kernel. At each step of the Markov Chain, a new sample is drawn from a 'proposal distribution' q(x,y). However the chain only moves to this sample point according to a 'probability of move' $\alpha = \pi(y)/\pi(x)$, otherwise it remains at the previous sample point.

The choice of proposal distribution q(x,y) has important consequences for the algorithm behaviour. Where q(x,y) is too diffuse or does not properly represent interactions between parameters, α is often small and many candidate points are rejected, slowing the chain evolution. Where q(x,y) is too compact, the chain will move inefficiently around the search space, causing particular problems with spatially distal optima. The SCEM-UA algorithm (Vrugt *et al.*, 2003) seeks to avoid these problems by continually updating the proposal distribution using information gained about the nature of the posterior distribution. The proposal distribution becomes a multivariate normal with mean and covariance structure taken as the sample mean and sample covariance of different 'complexes' of points in the high-density region of the sample space. Although it is not proven that the SCEM-UA algorithm with adaptive proposal distribution provides an ergodic Markov Chain with the correct invariant distribution (Haario *et al.*, 1999; 2001), experimental investigations have shown that the algorithm performs well (Vrugt *et al.*, 2003).

3.2 Formal Bayesian Likelihood

The MCMC method is first carried out using a formal Bayesian Likelihood derivation for the posterior density. Following Thiemann *et al.* (2001), Vrugt *et al.* (2003), Bates and Campbell (2001), Marshall *et al.*, (2004) and others, we assume that measurement errors can be transformed via a one-to-one transformation to have the exponential power density $E(\sigma,\beta)$, and hence the conditional posterior density can be derived to be of the form (Box and Tiao, 1973)

$$p(z \mid \theta, \sigma, \beta) = \left[\frac{\omega(\beta)}{\sigma}\right]^{T} \cdot \exp\left[-c(\beta) \cdot \sum_{t=1}^{T} \left|\frac{\upsilon(t)}{\sigma}\right|^{\frac{2}{1+\beta}}\right] \qquad \text{(Equation 1)}$$

Where

$$c(\beta) = \left[\frac{\Gamma[3(1+\beta)/2]}{\Gamma[(1+\beta)/2]}\right]^{1/(1+\beta)}, \omega(\beta) = \frac{\{\Gamma[3(1+\beta)/2]\}^{1/2}}{(1+\beta) \cdot \{\Gamma[(1+\beta)/2]\}^{3/2}}$$

β is a scale parameter, σ is the standard deviation of the measurement errors, T is the number of timesteps, and $\upsilon(t)$ are the transformed errors.

3.3 Informal Likelihood Measures

Secondly, the MCMC sampling is repeated using an informal likelihood measure as used under the philosophy of the GLUE system (Beven and Binley, 1992). This technique also requires the selection of a 'behaviourability threshold' such that when the likelihood measure falls below this value, the model is rejected. Although typically the choice of threshold has been based on the expert judgement of the modeller as to the error magnitude that is acceptable for the particular application, it may also be chosen objectively such that a set proportion of the observed values lie within the uncertainty bounds (Blasone *et al.*, 2008; Montanari, 2005).

3.3.1 Nash-Sutcliffe Likelihood

The Nash-Sutcliffe index of model efficiency (NSE; Equation 1) is one of the most commonly used descriptors of rainfall-runoff model performance Hall (2001).

$$\text{NSE} = 1 - \frac{\sigma_{\varepsilon}^{2}}{\sigma_{o}^{2}} \qquad \text{(Equation 2)}$$

Where σ_{ε}^2 is the error variance and σ_o^2 is the variance of the observed flow series.

Hence the NSE takes a value of 1 for a perfect model fit, a value of 0 for a model no better than the constant mean of the observed data. The Nash-Sutcliffe index is often used in the GLUE framework as an informal likelihood measure. In order for it to be used in SCEM-UA, it must be non-negative and monotonically increasing with improved performance. To meet the former condition, the NSE is set to zero when negative values are returned. The NSE is only used via the posterior density ratio R of two samples, which can be expressed in the following form:

$$R = \frac{1-\frac{\sigma_{\varepsilon 1}^2}{\sigma_o^2}}{1-\frac{\sigma_{\varepsilon 2}^2}{\sigma_o^2}} = \frac{\sigma_o^2 - \sigma_{\varepsilon 1}^2}{\sigma_o^2 - \sigma_{\varepsilon 2}^2} = \frac{K - SSE_1}{K - SSE_2} \qquad \text{(Equation 3)}$$

Where SSE_1 and SSE_2 are the sums of squared errors for the two samples and K is a constant.

After initial trials of a MCMC method using this index, it was found that the chain was initially slow to migrate to high performance regions of the sample space. This was hypothesised to be due to two factors:

the lack of ability to order poor model fits (as the NSE was set to zero whenever $\sigma_o^2 > \sigma_\varepsilon^2$) which prevented the chain from gradual movement towards high performance regions.(1) Poor representation of relative model performance, e.g. a NSE of 0.9 would typically be considered a significant improvement relative to a NSE of 0.8, however in this method there would be a high probability of move from 0.9 down to 0.8 as the posterior density ratio is 0.8/0.9 = 0.89.

(2) Lack of ability to order poor model fits (as the NSE was set to zero whenever $\sigma_o^2 > \sigma_\varepsilon^2$) which prevented the chain from gradual movement towards high performance regions.

In order to address this issue (1), the constant K may be adjusted to mimic the effect of the behavioural threshold and alter the ratio R; i.e. reducing K causes higher weight to be placed on small improvements in NSE. To address issue (2), the exact sum of squared error scores were retained such that all model fits could be correctly ordered, even though this information was not used to calculate the ratio R. A combination of these two measures This was found to significantly improve the Markov Chain efficiency.

3.3.2 Extended Nash-Sutcliffe

Despite the perennial popularity of error variance measures such as the Nash-Sutcliffe score, there are occasions when an approach base on the sum-of-squared-errors is likely to produce counterintuitive results when assessing the fit of modelled and observed hydrographs. Of particular concern is the relative importance assigned to discharge magnitude errors versus timing errors. It is a common occurrence for rainfall-runoff models to incorrectly predict the timing of a flood peak; however due to the timestep-by-timestep comparison in an SSE analysis, timing errors can cause extremely poor performance measure values (Figure 2).

Figure 2: A synthetic example of hydrographs in which a model with minor (2 hour) timing error is graded as having poorer performance than a model with 40% discharge error

A generalised version of the Nash-Sutcliffe likelihood is suggested in order to address these concerns, by allowing discrepancies between observed and modelled data points to be considered as a combination of discharge and timing errors. This is achieved by using the modeller’s judgment on relative importance of discharge and timing errors to determine the shape of an oval search window (Figure 3). The error at each timestep is defined as the minimum distance from the oval centre to the point on the oval boundary which intersects the opposing discharge curve. The squared error values are then summed and substituted directly into the standard Nash-Sutcliffe equation. Standard NS appears as a special case within the Extended NS when timing errors are considered infinitely worse than discharge errors and the search oval becomes a vertical line. A procedural description of calculation of the new error measure can be found in Appendix A.

Figure 3: Error magnitudes for the Extended Nash-Sutcliffe are found using an oval search window.

4 Results

4.1 Flow prediction

Formal Bayesian Likelihood

Model calibration was carried out using data from the year 1998, using the MCMC method described in Section 3.1 and a formal likelihood measure based on an exponential error distribution (Section 3.2). Ten parallel Markov Chains are run for a total of 5000 iterations; the first 1000 iterations are discarded as a ‘burn-in’ period for the chain. Gelman-Rubin convergence statistics are calculated to check the Markov Chain has converged to the stationary distribution representing the model posterior distribution.

Figure 4: 90% Uncertainty bounds on flow at Te Teko using formal likelihood measure to control MCMC search algorithm.

The resulting uncertainty bounds on the flow hindcast are shown in Figure 4; note that the bounds are sufficiently narrow to be hardly visible as distinct from the median calibrated prediction.

Informal Likelihood

The model calibration was repeated using the same Markov Chain set-up, but using in turn the Nash-Sutcliffe and Extended Nash-Sutcliffe likelihood measures. The resulting flow hindcasts are shown in Figures 5 and 6 respectively. It is clear that using an informal likelihood measure suggests a much greater uncertainty in the flow forecast, with uncertainties greatest during peak flow periods.

Figure 5: 90% Uncertainty bounds on flow at Te Teko using Nash-Sutcliffe informal likelihood measure to control MCMC search algorithm.

Figure 6: 90% Uncertainty bounds on flow at Te Teko using Extended Nash-Sutcliffe informal likelihood measure to control MCMC search algorithm.

A study of the Markov Chain behaviour can be used to provide additional information about the model response surface, and the success of the MCMC algorithm in fully exploring the surface (Vrugt *et al.*, 2003). Figures 7 and 8 allow a comparison of the sequential values of the Topmodel f parameter when using formal vs. informal likelihood measures. Figure 7 shows that in the case of the formal likelihood measure, the distribution quickly collapses to a single optimum, and the remainder of the parameter space is not explored. In contrast, Figure 8 shows that the informal likelihood measure produces a continuing wide dispersal of behavioural parameter values, and therefore a flatter response surface. Other model parameters showed similar trends. It is also interesting to note in Figure 8 that there is a distinct higher-density band for f in the range [0, 0.1], coupled with a more disperse band in the range [0.2, 0.6]. This suggests the possibility of a bi-modal distribution for f, with only the more peaked lower optimum found by the formal likelihood measure: this issue is discussed more fully in the following section.

Figure 7: Topmodel 'f' parameter value over successive iterations of the MCMC search algorithm using formal Bayesian (exponential error model) likelihood measure.

Figure 8: Topmodel 'f' parameter value over successive iterations of the MCMC search algorithm using informal 'Extended Nash-Sutcliffe' likelihood measure.

4.2 Calibration Constraints using Qualitative Geological Information

Internal Catchment Flow Gauging

By using the informal likelihood measure (Section 4.1) the Markov Chain revealed a dispersed posterior response surface, with the possibility of dual optima suggested by distinct bands in the parameter mixing diagrams when using the informal Extended Nash-Sutcliffe likelihood measure (Figure 8). Given the division of the catchment into dual 'permeable' and 'impermeable' areas, it seemed logical that these two phenomena might be related. The issue was investigated further using flow data from the two internal catchment gauges which had not previously been used in model calibration (Figure 9).

Figure 9: Comparison of Internal Flow predictions at Murupara (pumice subcatchment) and Galatea (greywacke subcatchment) using formal and informal likelihood measures.

Striking differences were seen here between the formal and informal likelihood results. The informal likelihood results show a very large spread in possible internal flow distribution in the catchment, where the majority of the quickflow may be attributed to either pumice or greywacke areas (Figures 9C/9D). In reality, the pumice subcatchment provides a steady baseflow, with the greywacke catchment providing a peaked response to storm events (refer to Section 2.1) – however the unconstrained model calibration may assign 'pumice' vs. 'greywacke' characteristics to the sub-catchments in either order. In contrast, the calibration using a formal likelihood measure has collapsed to a single parameter allocation (Figures 9A/9B) which has incorrectly classified the subcatchments and in effect assigned 'greywacke-type' characteristics to the pumice sub-catchment, and vice-versa.

Constrained Calibration

It is natural to ask whether the calibration procedure may be constrained such that Markov Chains converge to the correct optimum such the flow characteristics are correctly assigned to the two geologically distinct sub-catchments. Although in the case of the Rangitaiki this could be achieved using multi-criteria calibration with additional data from the internal flow gauges, here we are interested in a strategy using only the catchment outlet flow gauge, such that the methodology would be transferable to other catchments with a single flow gauge.

The constraint process aimed to subdivide the parameter space in the simplest possible way into volumes representing 'pumice' or 'greywacke' behaviour. In order to be considered as constraints, parameters had to satisfy the dual criteria of having a physical interpretation, such that characteristics could be accurately assigned, and showing good discrimination between model realisations representing the two response types. The parameters that achieved this were (1) Topmodel f parameter - related to depth of soil profile and aquifer response time (2) $\Delta\theta_1$ – effective drained porosity (3) $\Delta\theta_2$ – root zone storage.

Multiplier ranges were defined for each of these based on separation of the observed marginal distribution by behavioural group. This was achieved by physical interpretation of the bi-modal parameter distributions, and resulting predicted flows, in the unconstrained calibration (Figure 8). Previous research in New Zealand demonstrates significant behavioural differences between pumice vs. non-volcanic regions, with pumice regions characterized by lower flood peaks (McKerchar and Pearson, 1989) and higher yields (Hutchinson, 1990). The bi-modal form is therefore compatible with an expectation that parameter multipliers for Topmodel “f”, $\Delta\theta_1$ and $\Delta\theta_2$ may need to be different for the two geology types to make targeted corrections to the default values. The two modes of the parameter distribution are classified as providing ‘Pumice-type’ and ‘Greywacke-type’ behaviour respectively. The resulting marginal distributions are shown in Figure 10: the Topmodel f parameter is seen to show non-intersecting ranges for the two parameters sets, the $\Delta\theta_1$ and $\Delta\theta_2$ parameters show defined ranges for the ‘greywacke-type’ parameters only. Other parameters (not shown) did not show good discrimination between behavioural types.

Figure 10: Multiplier ranges categorised by behavioural type for parameters: (a) Topmodel f (b) $\Delta\theta_1$ effective drained porosity (c) $\Delta\theta_2$ root zone storage. These plots were used to define constrained parameter ranges

The calibration was re-run using appropriate parameter ranges for each sub-catchment according to its geological classification. An informal likelihood measure was used as this is consistent with the analysis suggesting the presence of behavioural simulations within the constrained range: the formal likelihood measure in contrast rejected at the 90% level all simulations within the new constraints. The Extended Nash-Sutcliffe measure was used in order to allow proper consideration of both magnitude and timing errors.

Figure 11: Internal Flow predictions at Murupara (pumice subcatchment) and Galatea (greywacke subcatchment) using informal likelihood measures under a constrained calibration procedure.

The results for flow predictions at the two internal catchment flow gauges are shown in Figure 11. These results show accurate flow prediction in each subcatchment with substantially reduced uncertainty compared to the unconstrained calibration. We therefore conclude that imposing constraints on the 3 parameters f, $\Delta\theta_1$, $\Delta\theta_2$ was sufficient to guide the MCMC algorithm to the correct optimum.

5 Discussion and Conclusions

Where a catchment has sub-regions of contrasting hydrological behaviour, such as those caused by different geologies, there is a danger that distributed hydrological models calibrated using integrated data such as catchment outlet discharge may fail to properly represent the range of hydrological behaviours. Due to a wide range of possible distributions of flow within different branches of the catchment, the response surface representing the posterior distribution may have multiple optima and flat areas characteristic of complex equifinal behaviour. It is therefore important to use a calibration procedure which is capable of fully capturing and describing the behavioural regions of the parameter space.

MCMC algorithms such as the Metropolis-Hastings and its variants are popular choices for efficient exploration of complex response surfaces, however this paper has shown that the formal likelihood measures which are typically used within such algorithms may prevent the Markov Chain from fully exploring regions of the parameter space which might be considered behavioural when assessed using a standard performance measure such as the Nash-Sutcliffe statistic. Such formal Bayesian approaches assume that the model structure is correct, and therefore do not account for cases where the parameters compensate for weaknesses in model structure. This may lead to cases where, although parameter uncertainty is small, the optimised parameter values are in fact ‘wrong,’ as shown in Section 4.1 in the form of extremely poor flow predictions at internal locations.

By using instead an informal likelihood measure, we attempt to capture the total uncertainty in flow predictions due a range of known and unknown error sources. This methodology results in a greater volume of the parameter space being sampled, thus revealing more complete information about possible multiple optima or flat areas of the response surface. Of course, the posterior probability distribution to be sampled must reflect the hydrologist’s best understanding of the errors present in the modeling process; where these can be described very exactly a formal likelihood measure would be a more appropriate choice and would better represent the information on posterior parameter distribution which could be derived from the observed data. Unfortunately, however, it may often be the case that a formal likelihood measure which makes strong assumptions about model error distribution is used under conditions of incomplete information on error form.

Finally, this paper has shown how the additional information gained using an exploration of the response surface using an informal likelihood measure can be used to improve the calibration process in order to focus the Markov Chain trajectory on regions of the parameter space reflecting our qualitative knowledge of system behaviour. The ability to incorporate qualitative or ‘soft’ data into calibration algorithms is very valuable but may be more effectively deployed in conjunction with a description of the response surface which identifies threshold or boundaries between different response types.

Appendix A

Algorithm for Calculation of Extended Nash-Sutcliffe Performance Measure

1. Define ε_T as the timing error (e.g. in hours) which is considered ‘equally bad’ as a discharge error of 1 unit (typically 1 m^3s^{-1}), and τ the maximum allowable timing error.

For each timestep (t) in turn:

2. Identify the greater of the two discharge series (observed, modelled) at time t:

$$Q_1(t) = \max\{Q_{obs}(t), Q_{\mathrm{mod}}(t)\}$$

3. Create a vector of timesteps within the allowable time window:

$$\overline{T} = [t-\tau, \ldots, t-\Delta t, t, t+\Delta t, \ldots, t+\tau]$$

4. Create a vector of discharges corresponding to these time steps:

$$\overline{Q_2} = \begin{cases} [Q_{obs}(t-\tau), \ldots, Q_{obs}(t), \ldots, Q_{obs}(t+\tau)] & \text{where } Q_{\mathrm{mod}}(t) \geq Q_{obs}(t) \\ [Q_{\mathrm{mod}}(t-\tau), \ldots, Q_{\mathrm{mod}}(t), \ldots, Q_{\mathrm{mod}}(t+\tau)] & \text{where } Q_{obs}(t) > Q_{\mathrm{mod}}(t) \end{cases}$$

5. Calculate the squared error vector relating to this set of time steps:

$$\overline{SE} = \left(\frac{t-\overline{T}}{\varepsilon_{\mathrm{T}}}\right)^2 + \left(Q_1(t) - \overline{Q_2}\right)^2$$

6. Minimise the squared error over the time window:

$$\mathit{Squared\ Error}\,(t) = \min\{\overline{SE}\}$$

Having calculated the squared error for each timestep, return to the standard Nash-Sutcliffe method:

7. Calculate the error variance

$$\sigma_\varepsilon{}^2 = \frac{1}{n-1} \cdot \sum_t Squared\ Error\left(t\right)$$

8. Calculated the Extended Nash-Sutcliffe Score:

$$\text{Extended NSE} = 1 - \frac{\sigma_\varepsilon{}^2}{\sigma_o{}^2}$$

where σ_o^2 is the variance of the observed flow series.

Note that at each timestep the oval search window is centred on the greater of the modelled and observed discharges: this avoids the situation where narrow, high discharge peaks which are not predicted correctly are not accounted for in the error calculation as the search window picks up low flows before or after these events. The reverse situation with a sudden trough in discharge levels would be extremely unusual in either a modelled or observed flow series.

References

Arhonditsis GB., Perhar G, Zhang W, Massos E, Shi M, Das A (2008), Addressing equifinality and uncertainty in eutrophication models, *Water Resources Research*, 44, W01420, doi:10.1029/2007WR005862.

Bandaragoda C, Tarboton DG, Woods R, 2004. Application of TOPNET in the distributed model intercomparison project. *Journal of Hydrology*, 298 (1-4) pp 178-201.

Bates BC and Campbell EP, 2001. A Markov Chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modelling. *Water Resources Research.* 37(4) 937-947.

Beanland S and Haines J, 1998. The kinematics of active deformation in the North Island, New Zealand, determined from geological strain rates, *New Zealand Journal of Geology and Geophysics* 41 (1998), pp. 311–324.

Beven KJ (1993) Prophecy, reality and uncertainty in distributed hydrologic modelling. *Advances in Water Resources.* 16:41–51

Beven KJ, 2001. *Rainfall-Runoff Modelling: The Primer*. Wiley: Chichester.

Beven KJ, 2003. Comment on "Bayesian recursive parameter estimation for hydrologic models" by M. Thiemann, M. Trosset, H. Gupta, and S. Sorooshian. *Water Resources Research*, 39 (5) 1116.

Beven KJ (2006) A manifesto for the equifinality thesis. *Journal of Hydrology*. 320 (1-2) 18-36

Beven KJ, Binley AM (1992) The future of distributed models: model calibration and uncertainty in prediction. *Hydrological Processes.* 6:279–298

Beven KJ and Kirkby MJ, 1979. A physically based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin* 24 1, pp. 43–69.

Beven KJ, Lamb R, Quinn P, Romanowicz R and Freer J, 1995. Topmodel. In: Singh, V.P., Editor, 1995. *Computer Models of Watershed Hydrology* 18, Water Resources Publications, Highlands Ranch, CO (Chapter 18); pp. 627–668.

Beven KJ, Smith P, Freer J, 2007. Comment on "Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology" by Pietro Mantovan and Ezio Todini. *Journal of Hydrology* 338 (3-4) 315-318

Blasone R-S, Vrugt JA, Madsen H, Rosbjerg D, Robinson MA, and Zyvoloski GA, 2008. Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. Advances in Water resources, 31, 630-648.

Box GEP and Tiao GC, 1973. *Bayesian Inference in Statistical Analysis.* Addison-Wesley-Longman: Reading, Mass.

Chib S, and Greenberg E, 1995. Understanding the Metropolis-Hastings Algorithm. *The American Statistician.* 49(4): 327 – 335.

Clark MP, Rupp DE, Woods RA, Zheng X, Ibbitt RP, Slater AG, Schmidt J and Uddstrom MJ, 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Advances in Water Research, in press.

Clarke RT, 1994. *Statistical Modelling in Hydrology*. Wiley: Chichester

Duan Q, Sorooshian S, and Gupta HV, 1992. Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models, *Water Resources Research*. 28(4), 1015–1031.

Engeland K and Gottschalk L, 2002. Bayesian estimation of parameters in a regional hydrological model. *Hydrology and Earth System Sciences.* 6 (5): 883:898.

Goring DG, 1994. Kinematic shocks and monoclinal waves in the Waimakariri, a steep, braided, gravel-bed river. *Proceedings of the International Symposium on Waves: Physical and Numerical Modelling*, University of British Columbia, Vancouver, Canada, 21–24 August, 1994 pp. 336–345

Gupta HV, Thiemann M, Trosset M, Sorooshian, S, 2003. Reply to comment by K. Beven and P. Young on "Bayesian recursive parameter estimation for hydrologic models", *Water Resources Research*, 39 (5) 1117.

Haario H, Sakman E, Tammimien J, 1999. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics.* 14(3): 375-395.

Haario H, Sakman E, Tammimien J, 2001. An adaptive Metropolis algorithm. *Bernouilli.* 7(2): 223-242.

Hall MJ, 2001. How well does your model fit the data? *Journal of Hydroinformatics* 3 (1) pp 49-55.
Hutchinson PD, 1990. Regression Estimation of Low Flow in New Zealand. Publication No. 22 of the Hydrology Centre, DSIR Marine and Freshwater, Christchurch, N.Z. 51p. ISSN 0112-1197.

Ibbitt RP, 1971: Development of a conceptual model of. interception. Hydrological research progress report. No. 5. Wellington, Ministry of Works.

Ibbitt RP and Woods R, 2002. Towards rainfall-runoff models that do not need calibration to flow data. *In Friend 2002 – Regional Hydrology: Bridging the Gap between Research and Practice. IAHS Publication no. 274.* Eds van Lanen, HAJ and Demuth, S. pp 189 -196.

Kavetski D, Kuczera G, Franks SW, 2006a. Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts. *Journal of Hydrology*. 320 (1-2) 173-186.

Kavetski D, Kuczera G, Franks SW, 2006b. Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis. *Journal of Hydrology*. 320 (1-2) 187-201.

Kuczera G, Kavetski D, Franks S, Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*. 331 (1-2) 161-177.

Liu Y, and Gupta HV, 2007: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*. 43, W07401, doi:10.1029/2006WR005756.

Mantovan P, Todini E, 2006. Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of Hydrology*. 330 (1-2): 368-381.

Mantovan P, Todini E, Martina MLV, 2007. Reply to comment by Keith Beven, Paul Smith and Jim Freer on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology". *Journal of Hydrology* 338 (3-4) 319-324.

Manville V, Newton EH, White JDL, 2005. Fluvial responses to volcanism: Resedimentation of the 1800a Taupo ignimbrite eruption in the Rangitaiki River catchment, North Island, New Zealand. *Geomorphology*, 65 (1-2), pp. 49-70

Marshall L, Nott D, Sharma A, 2004. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modelling. *Water Resources Research.* 40 (2): W02501.

McKerchar AI and Pearson CP, 1989. Flood frequency in New Zealand, *Publication, Hydrology Section* vol. 20, Division of Water Sciences, Department of Scientific and Industrial Research, Christchurch 87 pp.

Mein RG and Larson CL, 1973. Modeling infiltration during steady rain. Water Resour. Res. 9, pp. 384–394

Montanari A, 2005. Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research.* 41(8): W08406.

Newsome PFJ, Wilde RH, Willoughby EJ, 2000. *Land Resource Information System Spatial Data Layers.* Technical Report. Landcare Research NZ Ltd, Palmerston North, NZ.

Snelder TH; Biggs BJF, 2002. Multi-scale river environment classification for water resources management. *Journal of the American Water Resources Association* 38(5): 1225–1240.

Sorooshian, S and Dracup, JA, 1980. Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: correlated and heteroscedastic error cases. *Water Resources Research.* 16 (2): 430-442.

Sorooshian S, 1981. Parameter estimation of rainfall run-off models with heteroscedastic streamflow errors – noninformative data case. *Journal of Hydrology.* 52 (1-2): 127-138.

Sorooshian S, Duan Q, and Gupta HV (1993), Calibration of Rainfall-Runoff Models: Application of Global Optimization to the Sacramento Soil Moisture Accounting Model, Water Resour. Res., 29(4), 1185–1194.

Tait AB, Henderson RD, Turner RW, Zheng X, 2006. Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. *International Journal of Climatology.* 2097-2115.

Thiemann M, Trosset M, Gupta H, Sorooshian S, 2001. Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research.* 37 (10) 2521-2535.

Vrugt JA, Gupta HV, Bouten W, Sorooshian S, 2003. A Shuffled Complex Evolution Metropolis algorithm for optimisation and uncertainty assessment of hydrologic model parameters. *Water Resources Research,* 39 (8) 1201.

Vrugt JA, Gupta HV, Dekker SC, Sorooshian S, Wagener T, Bouten W, 2006. Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model. *Journal of Hydrology.* 324: 288-307.

Vrugt JA, ter Braak CJF, Gupta HV, and Robinson BA, 2008: Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modelling. *Stochastic Environmental Research and Risk Assessment*, in review.

Wagener T, Gupta HV, 1995. Model identification for hydrological forecasting under uncertainty, *Stochastic Environmental Research and Risk Assessment,* 19 (6): 378-387.

Figure 1: Geology of the Rangitaiki River basin, classified according to permeability. Gauging Locations are marked.

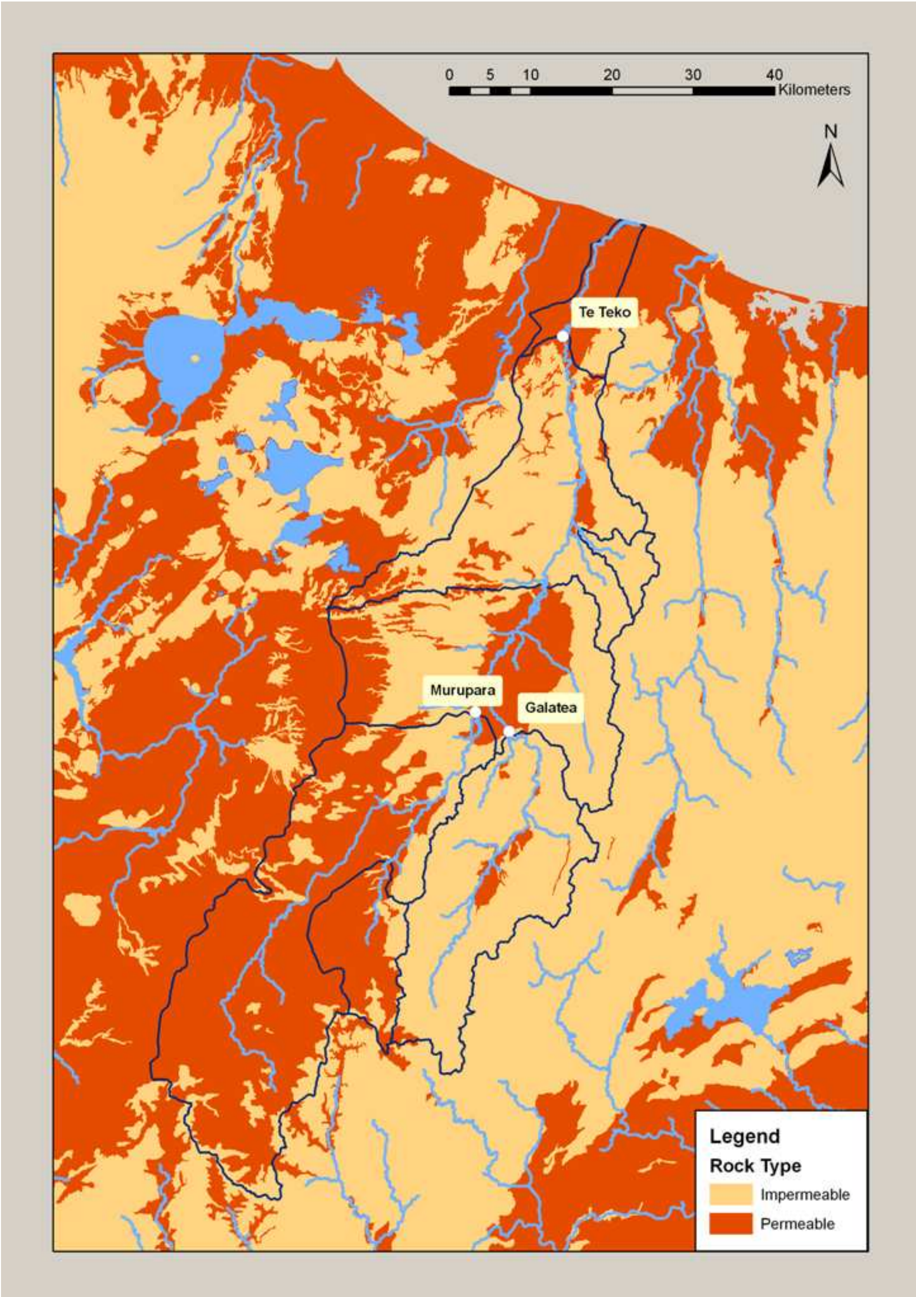

Figure 2: A synthetic example of hydrographs in which a model with minor (2 hour) timing error is graded as having poorer performance than a model with 40% discharge error

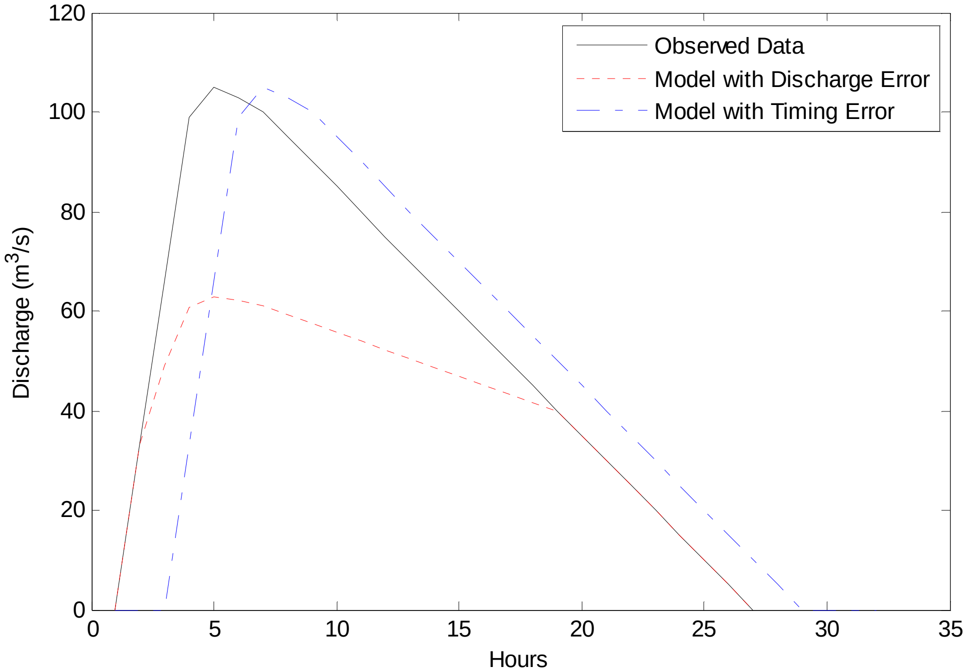

Figure 3: Search window to determine 'distance' between observed and predicted flow values under the Extended Nash Sutcliffe likelihood measure.

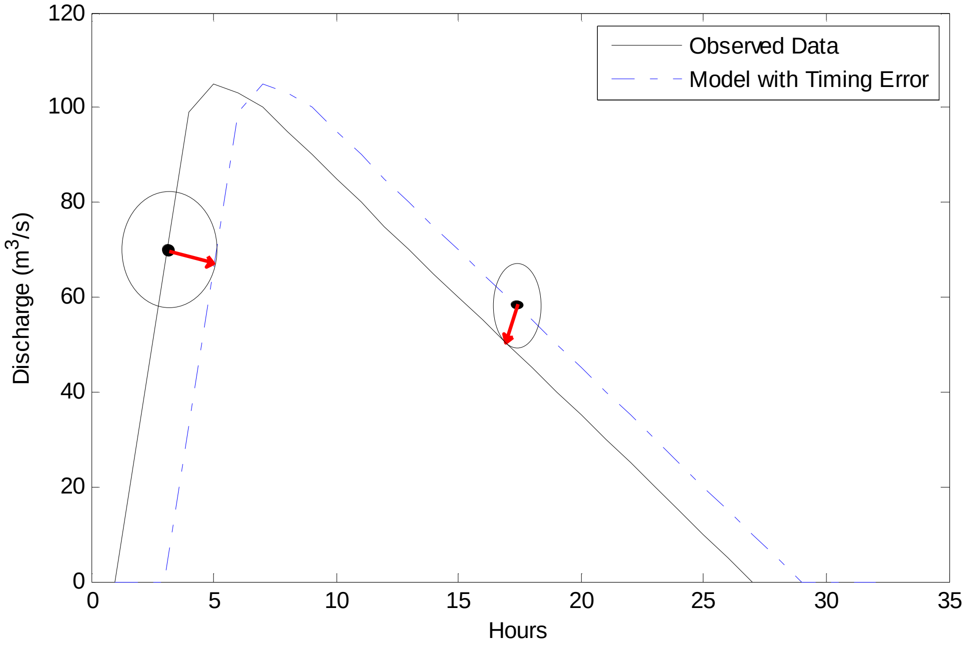

Figure 4: 90% Uncertainty bounds on flow at Te Teko using formal likelihood measure to control MCMC search algorithm. Note that the bounds are sufficiently narrow to be hardly visible as distinct from the median calibrated prediction.

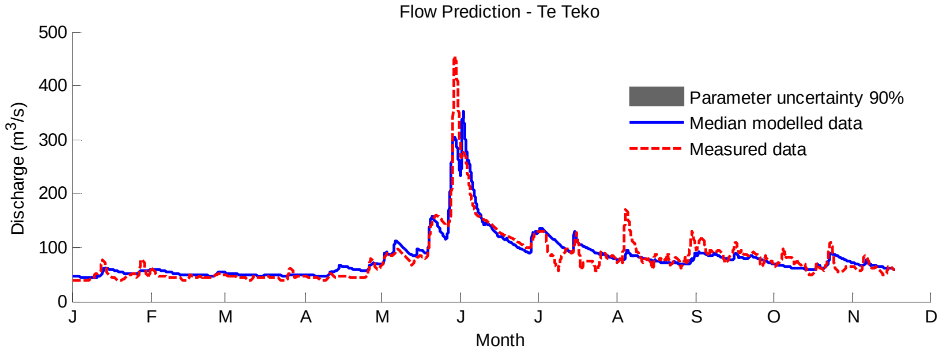

Figure 5: 90% Uncertainty bounds on flow at Te Teko using Nash-Sutcliffe informal likelihood measure to control MCMC search algorithm.

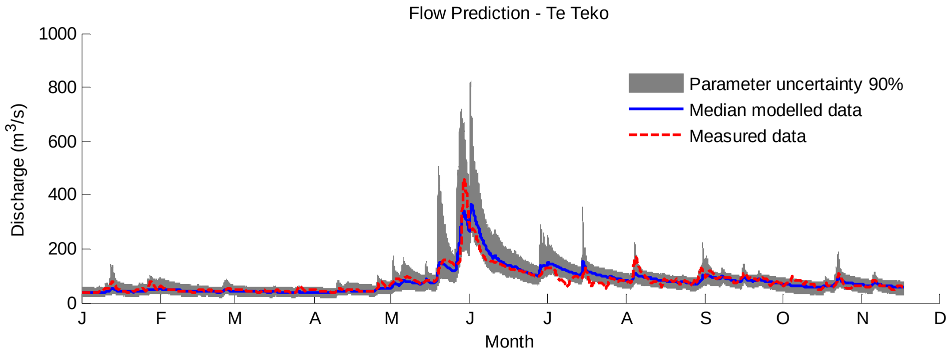

Figure 6: 90% Uncertainty bounds on flow at Te Teko using Extended Nash-Sutcliffe informal likelihood measure to control MCMC search algorithm.

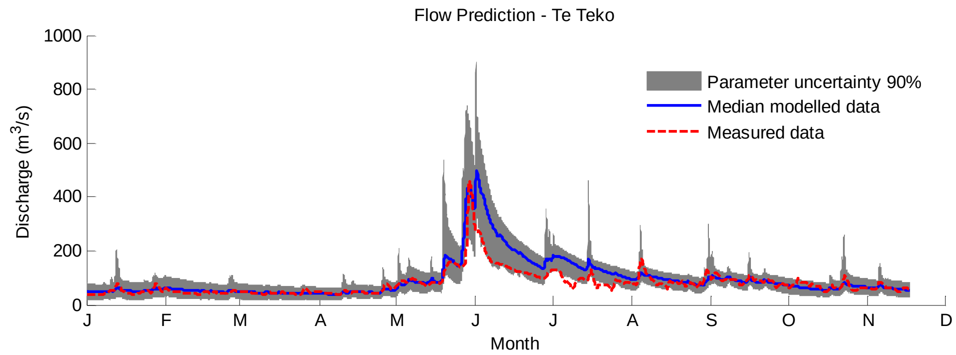

Figure 7: Topmodel ‘f’ parameter value over successive iterations of the MCMC search algorithm using formal Bayesian (exponential error model) likelihood measure

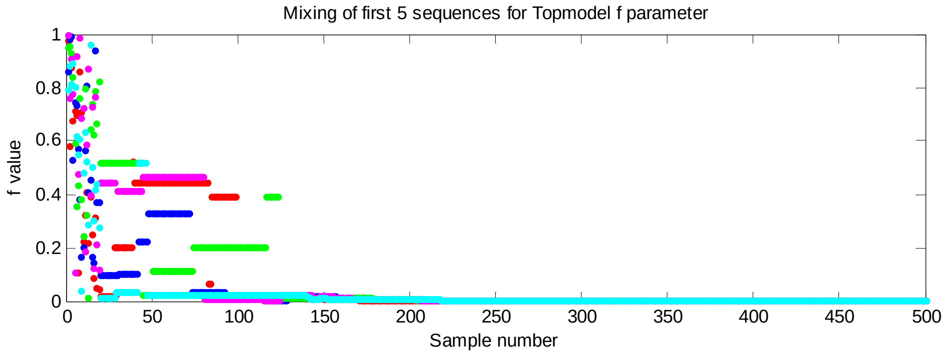

Figure 8: Topmodel ‘f’ parameter value over successive iterations of the MCMC search algorithm using informal ‘Extended Nash-Sutcliffe’ likelihood measure

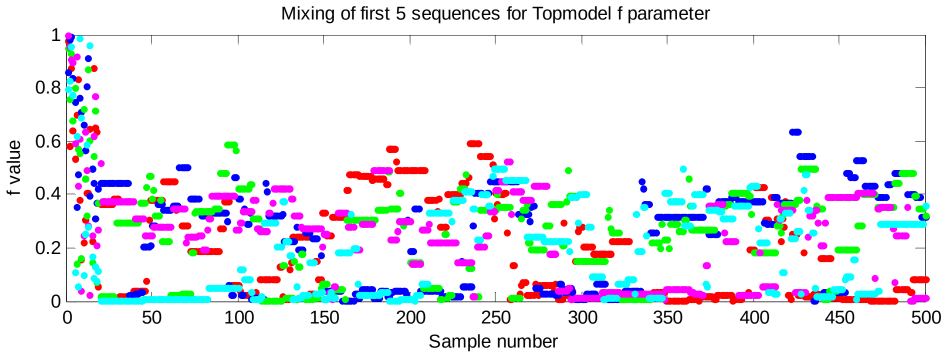

Figure 9: Comparison of Internal Flow predictions at Murupara (pumice subcatchment) and Galatea (greywacke subcatchment) using formal and informal likelihood measures

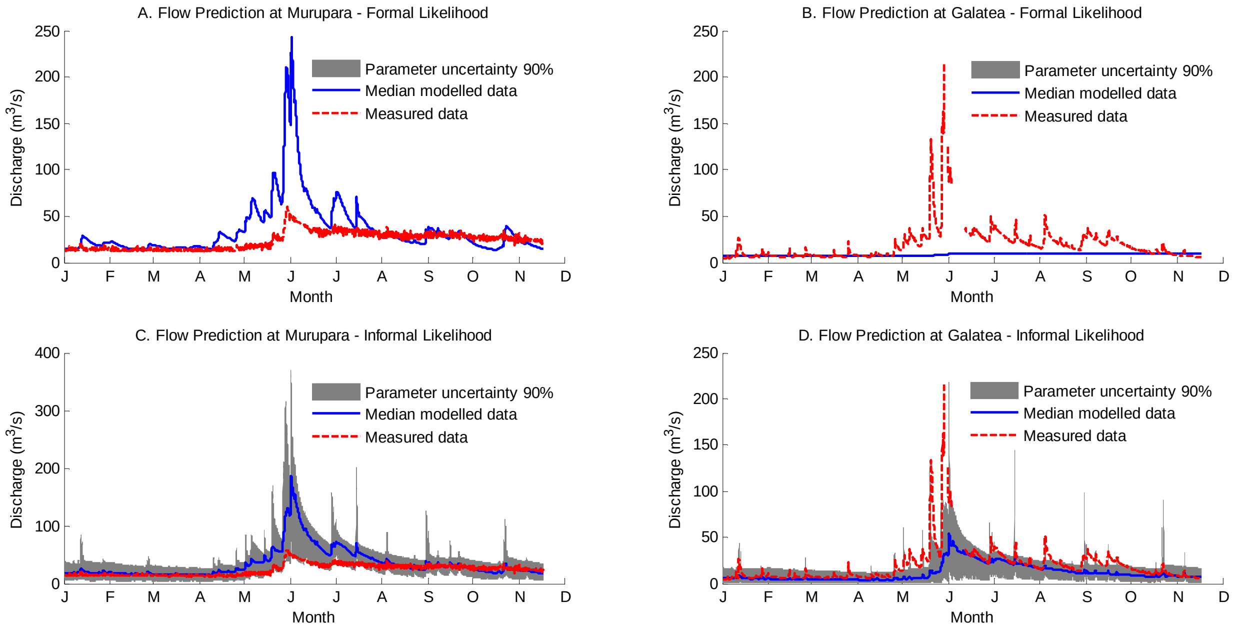

Figure 10: Multiplier ranges categorised by behavioural type for parameters: (a) Topmodel f (b) $\Delta\theta_1$ effective drained porosity (c) $\Delta\theta_2$ root zone storage. These plots were used to define constrained parameter ranges.

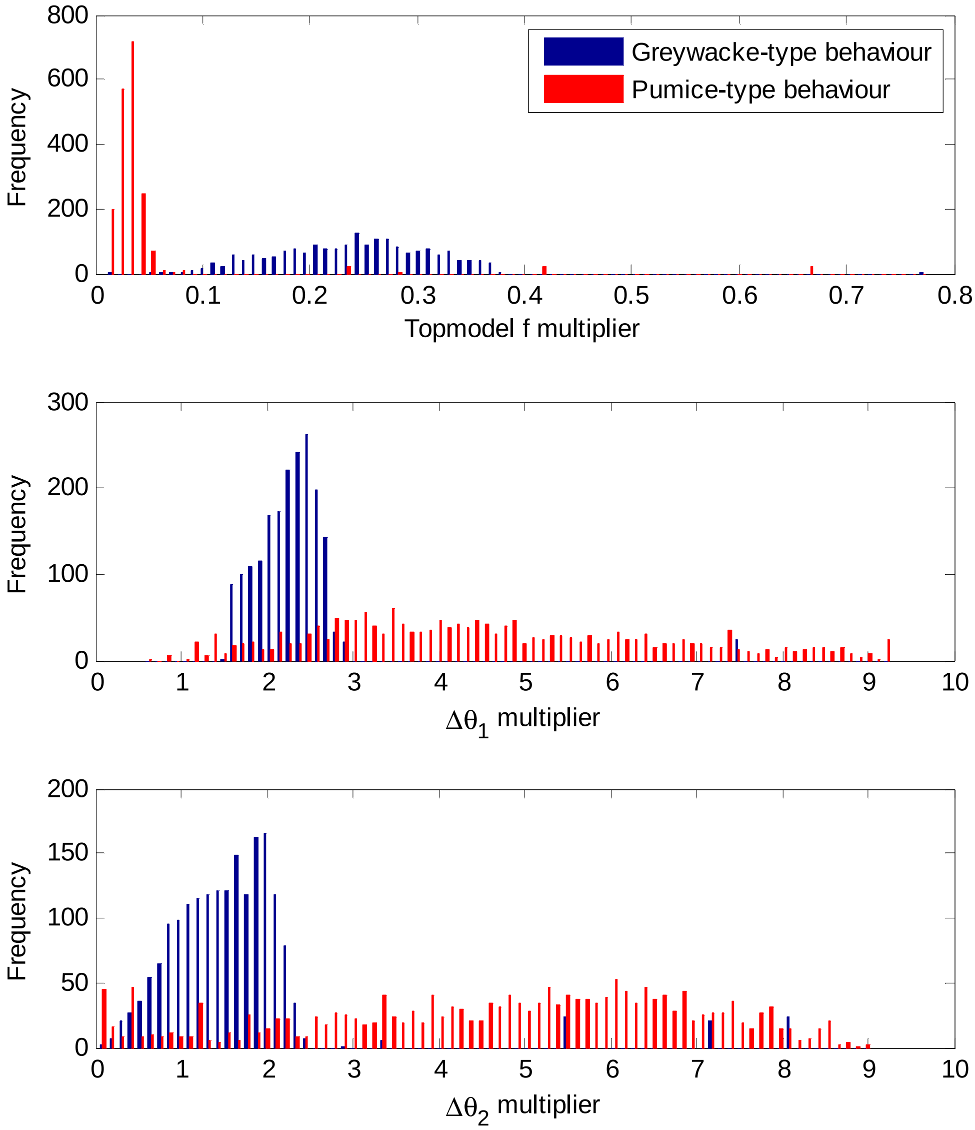

Figure 11: Internal Flow predictions at Murupara (pumice subcatchment) and Galatea (greywacke subcatchment) using informal likelihood measures under a constrained calibration procedure

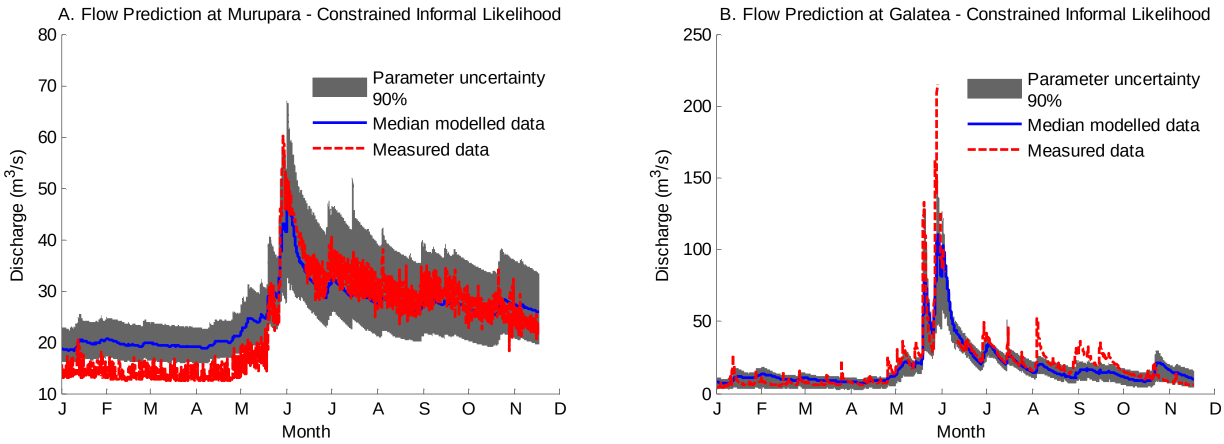

Table 1

TOPNET model parameters

	Name	Estimation
Sub-basin Parameters		
f (m^{-1})	Saturated store Sensitivity	Constant = 12.4 **(multiplier calibrated)**
K_0 (m/h)	Surface saturated hydraulic conductivity	Constant = 0.01 **(multiplier calibrated)**
$\Delta\theta_1$	Drainable porosity	From soils **(multiplier calibrated)**
$\Delta\theta_2$	Plant available porosity	From soils **(multiplier calibrated)**
D (m)	Depth of soil zone	Depth ¼ 1=f from soils **(multiplier calibrated)**
C	Soil zone drainage sensitivity	1
φ (m)	Wetting front suction	From soils
V (m/s)	Overland flow velocity	Constant = 0.1 **(multiplier calibrated)**
CC (m)	Canopy capacity	From vegetation
Cr	Intercepted evaporation enhancement	From vegetation
A	Albedo	From vegetation
Lapse (°C/m)	Lapse rate	0.0065
Channel parameters		
N	Mannings n	Constant = 0.024 **(multiplier calibrated)**
A	Hydraulic geometry constant	0.00011
B	Hydraulic geometry exponent	0.518
State variables		Initialization
z' (m)	Average depth to water table	Saturated zone drainage matches initial observed flow
SR (m)	Soil zone storage	0.02
CV (m)	Canopy storage	0.0005