Rainfall-runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme

Hilary McMillan[1] and Martyn Clark[1]

Received 15 July 2008; revised 24 November 2008; accepted 13 January 2009; published 22 April 2009.

[1] 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 and 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 trialed 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. Informal likelihoods are found to provide a more complete exploration of the behavioral regions of the response space and hence more accurate estimation of total uncertainty. Last, we demonstrate how information gained from the investigation of the response space, in conjunction with qualitative knowledge of system behavior, can be used to constrain the Markov chain trajectory.

Citation: McMillan, H., and M. Clark (2009), Rainfall-runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme, *Water Resour. Res.*, *45*, W04418, doi:10.1029/2008WR007288.

1. Introduction

[2] Improving availability and coverage of spatial data has driven developments in distributed, process-based catchment modeling; however, despite the correspondence between modeled and observed processes, it is not usually possible to determine model parameter values directly from field measurements. Instead, the values required are those of the ''effective parameters'' which represent integrated behavior at the model element scale. These values must be determined through some calibration method. As has been extensively discussed by *Beven* [1993, 2006], *Beven and Binley* [1992], *Wagener and Gupta* [2005] and others, 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 and 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 parameterizations which produce reasonable approximations to some aspects of true catchment behavior under some circumstances.

[3] 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 artifacts of model time step and solution techniques [*Kavetski et al.*, 2006a, 2006b]. 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, usually on the basis of expert judgment, 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.

[4] 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 optimization method [*Duan et al.*, 1992], using information

[1]National Institute of Water and Atmospheric Research, Christchurch, New Zealand.

0043-1397/09/2008WR007288

exchange between multiple sampler chains to improve convergence rates.

[5] 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 predicted and observed variable must be specified exactly; this specification 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].

[6] 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.*, 2003, 2006; *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 modeling 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.

[7] 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.

[8] 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). In the Rangitaiki catchment, 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. In response to the challenging problem of model calibration for the Rangitaiki, this paper focuses on three objectives:

[9] 1. To compare two strategies for model calibration using MCMC methods (in this case the SCEM-UA algorithm). The first strategy (section 3.1) uses a "formal" likelihood function based on strict assumptions about the error structure; the second strategy (section 3.2) uses an "informal" likelihood based on the modeler's judgment. The two approaches are assessed in terms of their success at full coverage of the response surface.

[10] 2. To investigate an extension to the standard "informal" likelihoods on the basis of sum of squared errors (e.g., Nash-Sutcliffe efficiency), by incorporating the timing error of the simulated hydrograph into the calibration objective function.

[11] 3. To test a "spatially informed" approach to MCMC calibration of a distributed rainfall-runoff model to overcome the difficulties of a multimodal parameter distribution caused by the heterogeneous geology of the catchment. Flow data from subcatchments are used to independently verify model success at reproducing the hydrological response.

2. Model and Data

2.1. TOPNET

[12] TOPNET was developed by combining TOPMODEL [*Beven and Kirkby*, 1979; *Beven et al.*, 1995], which is most suited to small watersheds, with a kinematic wave channel routing algorithm (D. G. Goring, Kinematic shocks and monoclinal waves in the Waimakariri, a steep, braided, gravel-bed river, paper presented at the International Symposium on Waves: Physical and Numerical Modelling, University of British Columbia, Vancouver, Canada, 21–24 August 1994) so as to have a modeling system that can be applied over large watersheds using smaller subbasins 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 subsurface storage controlling the dynamics of the saturated contributing area and base flow recession. To form a complete model, potential evapotranspiration, interception (based on the work of *Ibbitt* [1971]), infiltration (using a Green-Ampt mechanism [*Mein and Larson*, 1973]) and soil zone components were added. Kinematic wave routing moves the subbasin inputs through the stream channel network. Complete model equations are provided by *Clark et al.* [2008] and are not repeated here.

2.2. Catchment

[13] 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 $m^3 s^{-1}$. The river flows along a series of fault angle valleys which define a

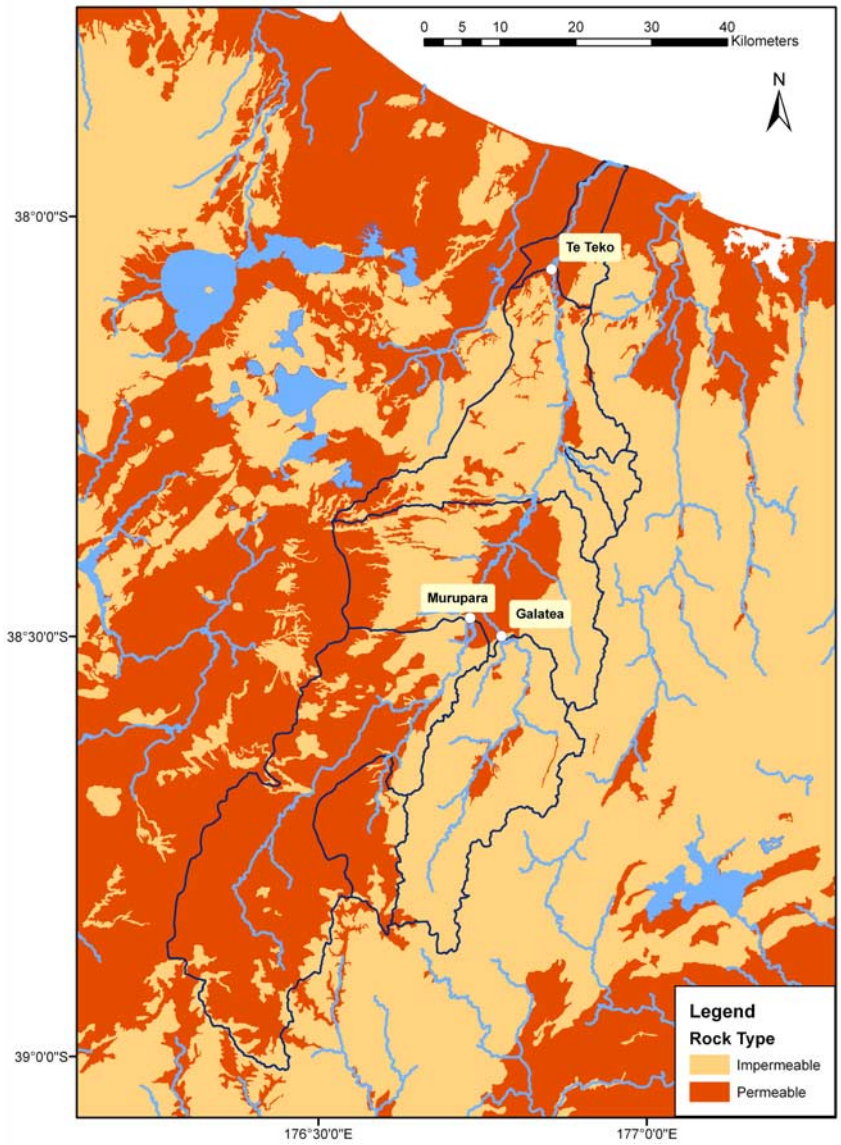

Figure 1. Geology of the Rangitaiki River basin, classified according to permeability. Gauging locations are marked.

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 metagreywackes [*Beanland and Haines*, 1998; *Manville et al.*, 2005]. These two parts of the catchment have strikingly different hydrological regimes: the porous tephras have a characteristic high stable base flow regime and subdued flood peaks; the steep and relatively impermeable greywacke responds quickly to rainfall with a peaked runoff pattern.

[14] 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).

2.3. Data

[15] Gauging data for the Rangitaiki is available at Te Teko, at the entrance to the coastal Rangitaiki Plains (Figure 1). The gauging station has a catchment area of 2890 km^2 and represents the combined flow of the greywacke and pumice areas: significant flood peaks are superimposed on a relatively sustained base flow. 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 mean annual mean flow is 21 $m^3\ s^{-1}$ and the mean annual flood is 40 $m^3\ s^{-1}$. 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 $m^3\ s^{-1}$, and the mean annual flood is 109 $m^3 s^{-1}$ [*McKerchar and Pearson*, 1989].

[16] 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°

Table 1. TOPNET Model Parameters

Parameter	Name	Estimation
	Subbasin Parameters	
f (m^{-1})	Saturated store sensitivity	Constant = 12.4 (multiplier calibrated)
K_0 (m h^{-1})	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 1/4 1 = f from soils (multiplier calibrated)
C	Soil zone drainage sensitivity	1
φ (m)	Wetting front suction	From soils
V (m s^{-1})	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 (deg C m^{-1})	Lapse rate	0.0065
	Channel Parameters	
N	Manning's 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

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.

[17] 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 subbasins 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 land cover 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.

2.4. Calibration via Parameter Multipliers

[18] 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, because of 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 behaviors. Poor forecasts would then be produced at internal catchment locations [*Clark et al.*, 2008].

[19] 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 subregions 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.

[20] The method used is to classify each Strahler 3 subcatchment 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 optimization process allows all multipliers to be calibrated simultaneously, such that the optimum combination of process descriptions in the two categories is found.

[21] 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). The internal check allows a test of model conditioning and parameter identification success; this is an important consideration because of the increased number of parameters used with this method.

3. MCMC Technique (Bayesian Uncertainty Framework)

3.1. Metropolis and Adaptive Metropolis Algorithms

[22] 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.

[23] The choice of proposal distribution q(x,y) has important consequences for the algorithm behavior. 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

[24] 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(σ,β), and hence the conditional posterior density can be derived to be of the form [*Box and Tiao*, 1973]

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

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 ($\beta = 1$ is used in this study), σ is the standard deviation of the measurement errors, T is the number of time steps, z is the transformed modeled discharge and υ(t) are the transformed errors.

3.3. Informal Likelihood Measures

[25] Second, 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 ''behaviorability 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 judgment of the modeler 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

[26] 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]:

$$\mathrm{NSE} = 1 - \frac{\sigma_{\varepsilon}^2}{\sigma_o^2} \quad (2)$$

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

[27] 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 nonnegative 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:

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

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

[28] 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 hypothesized to be due to two factors.

[29] 1. Poor representation of relative model performance; for example, 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

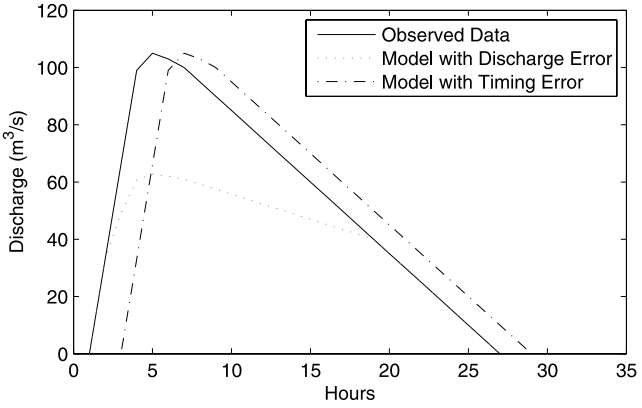

Figure 2. A synthetic example of hydrographs in which a model with minor (2 h) timing error (dash-dotted line; NSE = 0.62) is graded as having poorer performance than a model with 40% discharge error (dotted line; NSE = 0.66).

probability of move from 0.9 down to 0.8 as the posterior density ratio is 0.8/0.9 = 0.89.

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

[31] In order to address issue 1, the constant K may be adjusted to mimic the effect of the behavioral threshold and alter the ratio R; that is, reducing K causes higher weight to be placed on small improvements in NSE. To address issue 2, the exact sums 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 was found to significantly improve the Markov chain efficiency.

3.3.2. Extended Nash-Sutcliffe

[32] Despite the perennial popularity of error variance measures such as the Nash-Sutcliffe score, there are occasions when an approach based on the sum of squared errors is likely to produce counterintuitive results when assessing the fit of modeled 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, because of the time step–by–time step comparison in an SSE analysis, timing errors can cause extremely poor performance measure values (Figure 2).

[33] A generalized version of the Nash-Sutcliffe likelihood is suggested in order to address these concerns, by allowing discrepancies between observed and modeled data points to be considered as a combination of discharge and timing errors. This generalization is achieved by using the modeler's judgment on relative importance of discharge and timing errors to determine the shape of an elliptical search window (Figure 3). The error at each time step is defined as the minimum distance from the ellipse center to the point on the ellipse 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 not considered and the ellipse becomes a vertical line. A procedural description of calculation of the new error measure can be found in Appendix A.

4. Results

4.1. Flow Prediction

4.1.1. Formal Bayesian Likelihood

[34] 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. 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.

4.1.2. Informal Likelihood

[35] The model calibration was repeated using the same Markov chain setup, 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. The effects of using the extended NS performance measure are also demonstrated (Figure 6): the median model prediction has reduced discharge error at the flood peak (but increased timing error) when compared to the median prediction using standard NS (Figure 5).

[36] A study of the Markov chain behavior 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]. It is useful to compare the sequential values of the TOPMODEL f parameter when using formal versus informal likelihood measures, by plotting traces from the Markov chains produced by the SCEM-UA algorithm (Figures 7 and 8). In the case of the formal likelihood measure, the distribution quickly

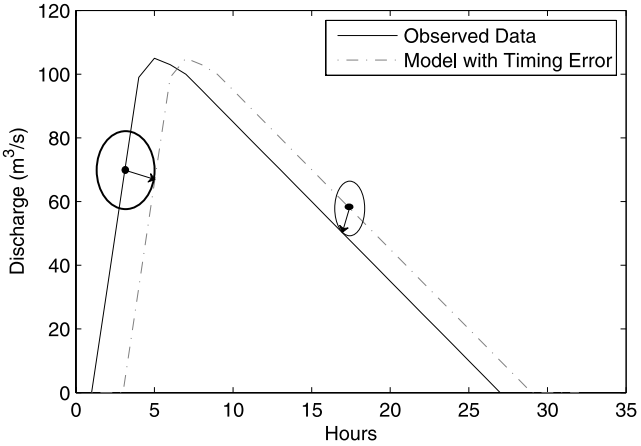

Figure 3. Conceptual diagram showing two examples of search window ellipses used to determine "distance" between observed and predicted flow values under the extended Nash Sutcliffe likelihood measure.

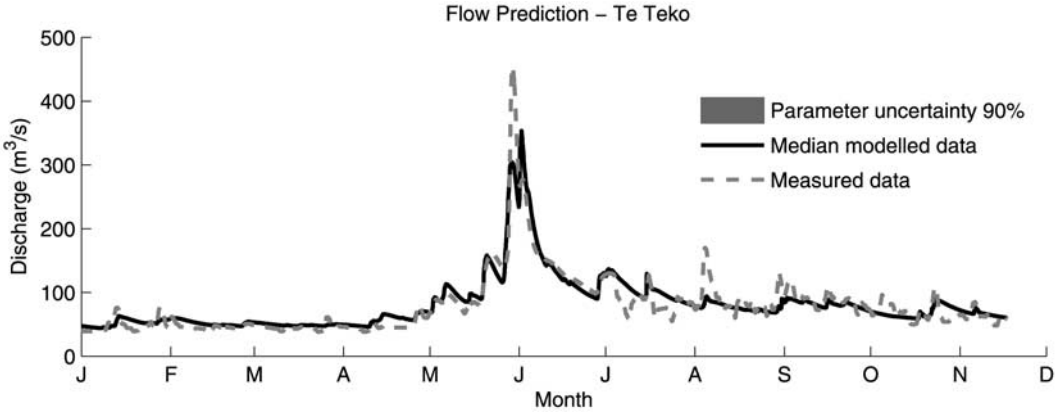

Figure 4. The 90% confidence limits on flow at Te Teko using formal likelihood measures to control the MCMC search algorithm. Note that the bounds are sufficiently narrow to be hardly visible as distinct from the median calibrated prediction.

collapses to a single optimum, and the remainder of the parameter space is not explored (Figure 7). In contrast, the informal likelihood measure produces a continuing wide dispersal of behavioral parameter values, and therefore a flatter response surface (Figure 8). 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 bimodal 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.

4.2. Calibration Constraints Using Qualitative Geological Information

4.2.1. Internal Catchment Flow Gauging

[37] 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).

[38] 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 quick flow may be attributed to either pumice or greywacke areas (Figures 9c and 9d). In reality, the pumice subcatchment provides a steady base flow, with the greywacke catchment providing a peaked response to storm events (refer to section 2.1); however, the unconstrained model calibration may assign "pumice" versus "greywacke" characteristics to the subcatchments in either order. In contrast, the calibration using a formal likelihood measure has collapsed to a single parameter allocation (Figures 9a and 9b) which has incorrectly classified the subcatchments and in effect assigned "greywacke-type" characteristics to the pumice subcatchment, and vice versa.

4.2.2. Constrained Calibration

[39] 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 subcatchments. Although in the case of the Rangitaiki this could be achieved using multicriteria 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.

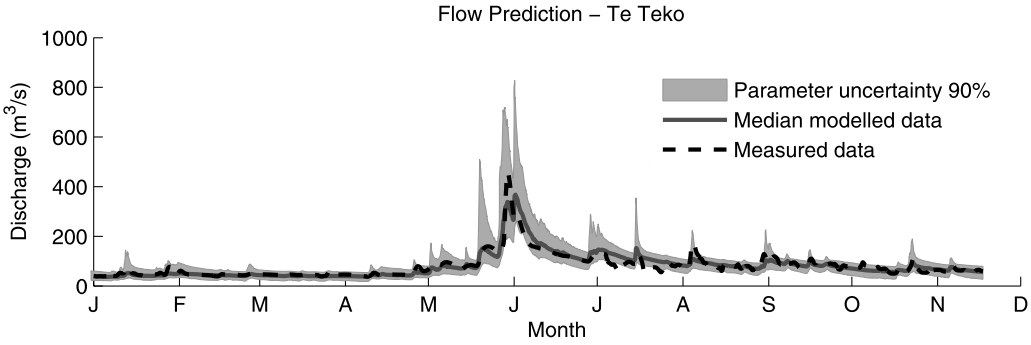

Figure 5. The 90% confidence limits on flow at Te Teko using Nash-Sutcliffe informal likelihood measures to control the MCMC search algorithm.

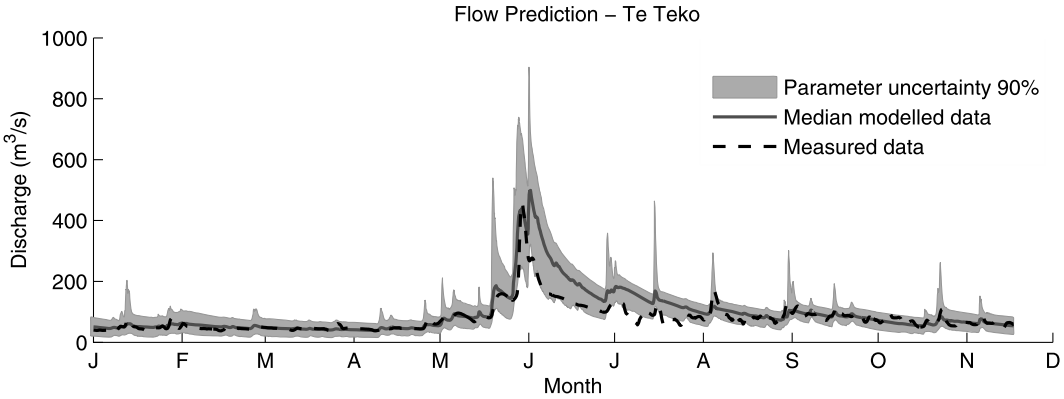

Figure 6. The 90% confidence limits on flow at Te Teko using extended Nash-Sutcliffe informal likelihood measures to control the MCMC search algorithm.

[40] The constraint process aimed to subdivide the parameter space in the simplest possible way into volumes representing "pumice: or "greywacke" behavior. 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 realizations 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, and (3) $\Delta\theta_2$, root zone storage.

[41] Multiplier ranges were defined for each of these on the basis of separation of the observed marginal distribution by behavioral group. This separation was achieved by physical interpretation of the bimodal parameter distributions, and resulting predicted flows, in the unconstrained calibration (Figure 8). Previous research in New Zealand demonstrates significant behavioral differences between pumice versus nonvolcanic regions, with pumice regions characterized by lower flood peaks [*McKerchar and Pearson*, 1989] and higher yields [*Hutchinson*, 1990]. The bimodal 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" behavior, respectively. The resulting marginal distributions are shown in Figure 10: the TOPMODEL f parameter is seen to show nonintersecting 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 behavioral types.

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

[43] Flow prediction results at the two internal catchment flow gauges show accurate predictions in each subcatchment with substantially reduced uncertainty compared to the unconstrained calibration (Figure 11). 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 a more reasonable optimum which predicts internal behavior consistent with observations.

5. Discussion and Conclusions

[44] Where a catchment has subregions of contrasting hydrological behavior, such as those caused by different

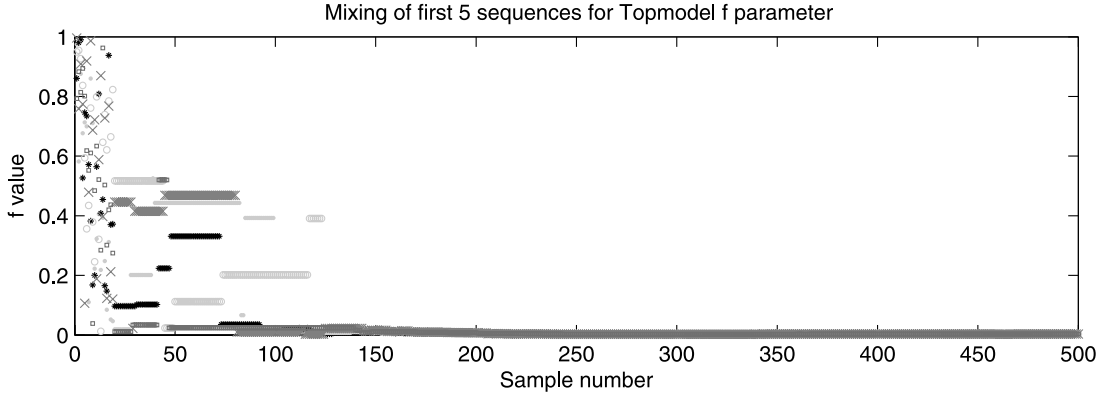

Figure 7. TOPMODEL "f" parameter value over successive iterations of five chains from the MCMC search algorithm using formal Bayesian (exponential error model) likelihood measures.

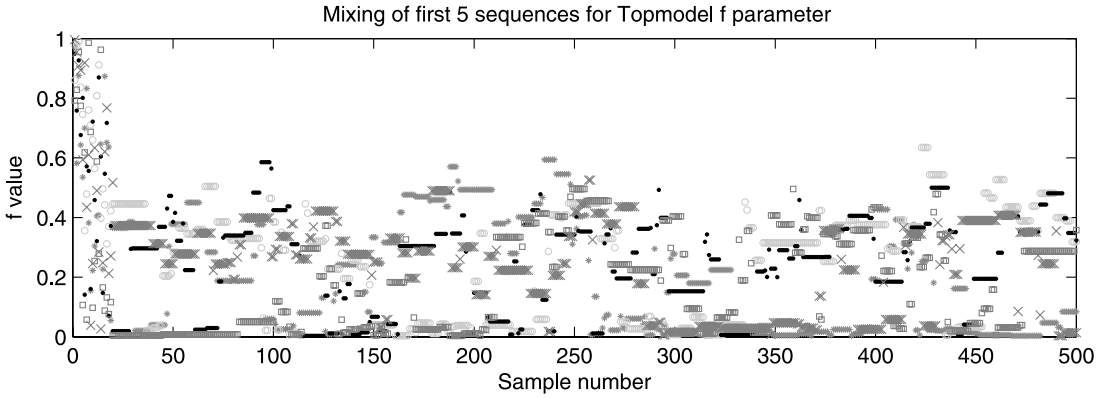

Figure 8. TOPMODEL "f" parameter value over successive iterations of five chains from the MCMC search algorithm using informal "extended Nash-Sutcliffe" likelihood measures.

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 behaviors. Because of 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 behavior. It is therefore important to use a calibration procedure which is capable of fully capturing and describing the behavioral regions of the parameter space.

[45] 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 behavioral 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 optimized parameter values are in fact "wrong," as shown in section 4.1 in the form of extremely poor flow predictions at internal locations.

[46] 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

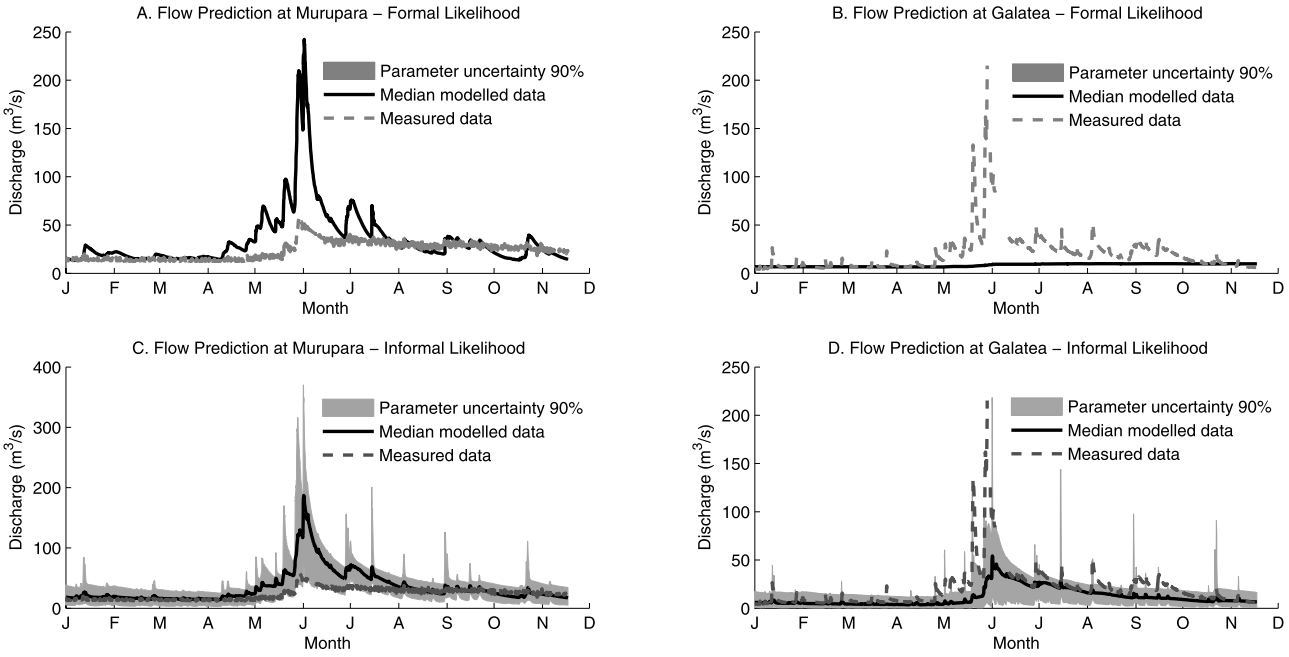

Figure 9. Comparison of internal flow predictions at Murupara (pumice subcatchment) and Galatea (greywacke subcatchment) using formal and informal likelihood measures.

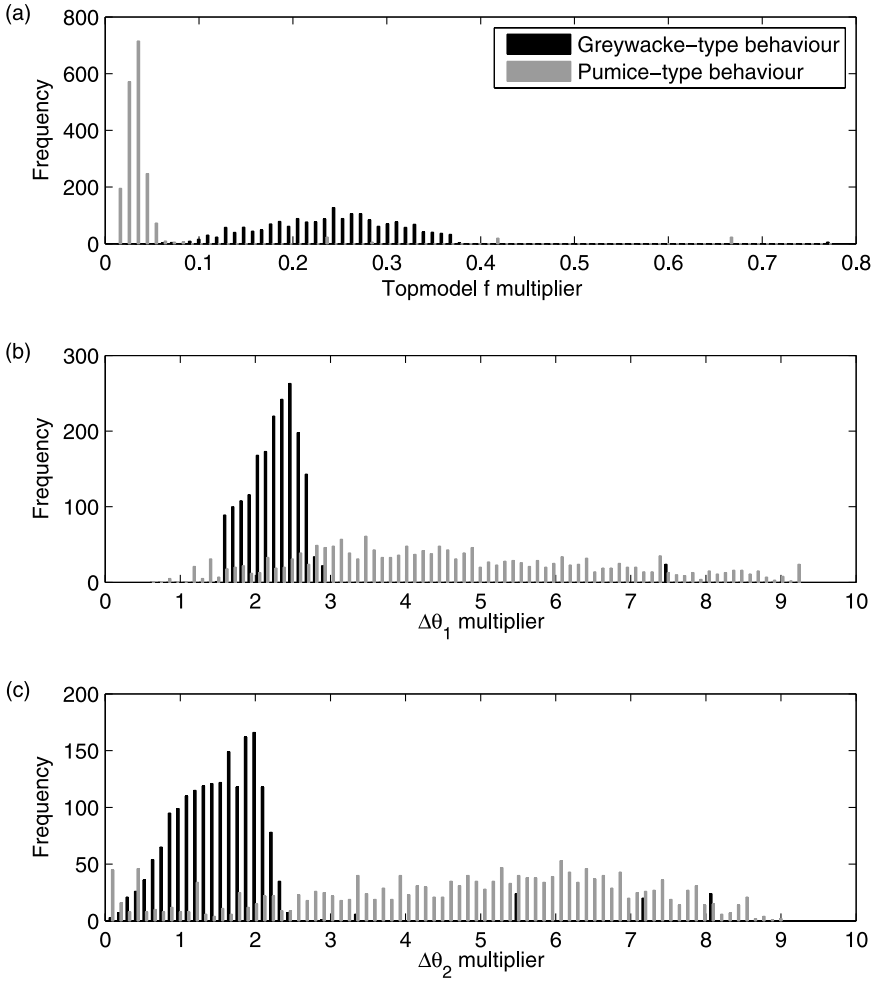

Figure 10. Multiplier ranges from unconstrained calibration, categorized by behavioral type for parameters: (a) TOPMODEL f, (b) $\Delta\theta_1$ effective drained porosity, and (c) $\Delta\theta_2$ root zone storage. These plots were used to define constrained parameter ranges.

better represent the information on posterior parameter distribution that 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.

[47] This paper also proposed an extension to the standard Nash-Sutcliffe efficiency to enable timing errors to be included in the model assessment. The "extended Nash-Sutcliffe" performance measure is tested as an alternative informal likelihood and is based on modeler judgment to weight the importance of timing errors versus discharge

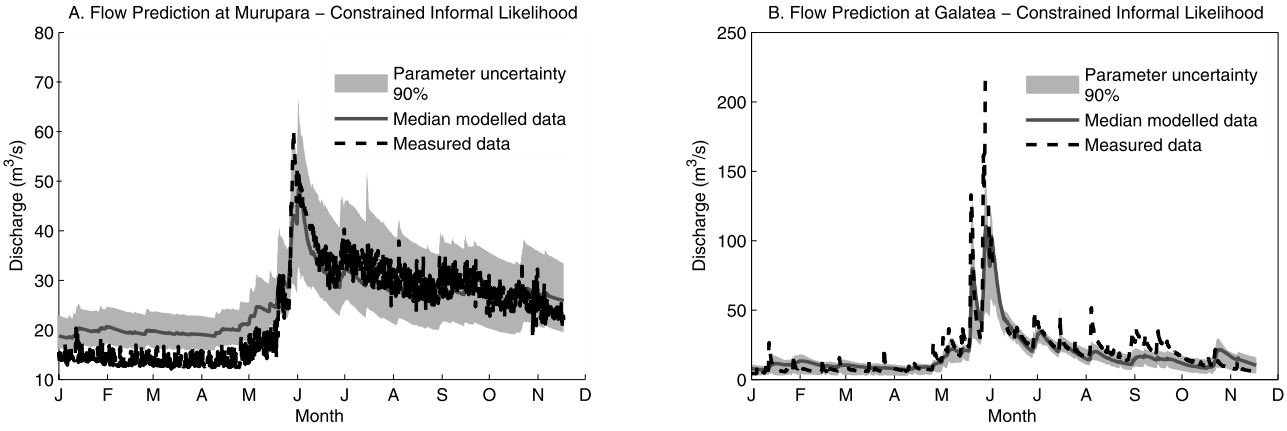

Figure 11. Internal flow predictions at (a) Murupara (pumice subcatchment) and (b) Galatea (greywacke subcatchment) using informal likelihood measures under a constrained calibration procedure.

errors. The method is found to have considerable advantages in improving estimation of peak flow magnitudes, while retaining the benefits of the standard Nash-Sutcliffe in providing a scalar objective function value whose value is intuitively understood by many hydrologists. In some cases, however, it may be more susceptible than standard NS to volume error in model predictions. This potential tradeoff demonstrates the difficulty of summarizing hydrological performance with a single statistic, a fact that has recently led to calls for more comprehensive model assessment via multiple "diagnostic signatures" [*Gupta et al.*, 2008].

[48] 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 behavior. In this instance, knowledge of the spatial distribution of geological types within the catchment, together with results from the exploratory MCMC run, are used to guide additional constraints placed on the calibration procedure. 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 thresholds or boundaries between different response types.

Appendix A: Algorithm for Calculation of Extended Nash-Sutcliffe Performance Measure

[49] 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.

[50] For each time step (t) in turn:

[51] 2. Identify the greater of the two discharge series (observed, modeled) at time t:

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

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

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

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

$$\mathbf{Q}_2 = \begin{cases} [Q_{obs}(t-\tau), \ldots, Q_{obs}(t), \ldots, Q_{obs}(t+\tau)] \text{ where } Q_{\mathrm{mod}}(t) \geq Q_{obs}(t) \\ [Q_{\mathrm{mod}}(t-\tau), \ldots, Q_{\mathrm{mod}}(t), \ldots, Q_{\mathrm{mod}}(t+\tau)] \text{ where } Q_{obs}(t) > Q_{\mathrm{mod}}(t) \end{cases}$$

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

$$\mathbf{SE} = \left(\frac{t-\mathbf{T}}{\varepsilon_T}\right)^2 + (Q_1(t) - \mathbf{Q}_2)^2$$

Where the first term represents the timing error contribution and the second term represents the discharge error contribution.

[52] 6. Minimize the squared error over the time window:

$$\text{Squared error}(t) = \min\{\mathbf{SE}\}$$

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

[53] 7. Calculate the error variance

$$\sigma_\varepsilon^2 = \frac{1}{n-1} \cdot \sum_t \text{squared error } (t)$$

[54] 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.

[55] Note that at each time step the elliptical search window is centered on the greater of the modeled 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 modeled or observed flow series.

[56] **Acknowledgment.** This work was supported by Foundation for Research, Science and Technology (FRST) grant C01X0812.

References

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

Bandaragoda, C., D. G. Tarboton, and R. Woods (2004), Application of TOPNET in the distributed model intercomparison project, *J. Hydrol.*, *298*(1–4), 178–201, doi:10.1016/j.jhydrol.2004.03.038.

Bates, B. C., and E. P. Campbell (2001), A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modelling, *Water Resour. Res.*, *37*(4), 937–947, doi:10.1029/2000WR900363.

Beanland, S., and J. Haines (1998), The kinematics of active deformation in the North Island, New Zealand, determined from geological strain rates, *N. Z. J. Geol. Geophys.*, *41*, 311–324.

Beven, K. J. (1993), Prophecy, reality and uncertainty in distributed hydrologic modelling, *Adv. Water Resour.*, *16*, 41–51, doi:10.1016/0309-1708(93)90028-E.

Beven, K. J. (2001), *Rainfall-Runoff Modelling: The Primer*, John Wiley, Chichester, U. K.

Beven, K. J. (2003), Comment on "Bayesian recursive parameter estimation for hydrologic models" by M. Thiemann, M. Trosset, H. Gupta, and S. Sorooshian, *Water Resour. Res.*, *39*(5), 1116, doi:10.1029/2001WR001183.

Beven, K. J. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1–2), 18–36.

Beven, K. J., and A. M. Binley (1992), The future of distributed models: Model calibration and uncertainty in prediction, *Hydrol. Processes*, *6*, 279–298, doi:10.1002/hyp.3360060305.

Beven, K. J., and M. J. Kirkby (1979), A physically based variable contributing area model of basin hydrology, *Hydrol. Sci. Bull.*, *24*(1), 43–69.

Beven, K. J., R. Lamb, P. Quinn, R. Romanowicz, and J. Freer (1995), TOPMODEL, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, chap. 18, pp. 627–668, Water Resour. Publ., Highlands Ranch, Colo.

Beven, K. J., P. Smith, and J. Freer (2007), Comment on "Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology" by Pietro Mantovan and Ezio Todini, *J. Hydrol.*, *338*(3–4), 315–318, doi:10.1016/j.jhydrol.2007.02.023.

Blasone, R.-S., J. A. Vrugt, H. Madsen, D. Rosbjerg, M. A. Robinson, and G. A. Zyvoloski (2008), Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling, *Adv. Water Resour.*, *31*, 630–648, doi:10.1016/j.advwatres.2007.12.003.

Box, G. E. P., and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, Mass.

Chib, S., and E. Greenberg (1995), Understanding the Metropolis-Hastings algorithm, *Am. Stat.*, *49*(4), 327–335, doi:10.2307/2684568.

Clark, M. P., D. E. Rupp, R. A. Woods, X. Zheng, R. P. Ibbitt, A. G. Slater, J. Schmidt, and M. J. Uddstrom (2008), Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model, *Adv. Water Resour.*, *31*, 1309–1324, doi:10.1016/j.advwatres.2008.06.005.

Clarke, R. T. (1994), *Statistical Modelling in Hydrology*, John Wiley, Chichester, U. K.

Duan, Q., S. Sorooshian, and H. V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031, doi:10.1029/91WR02985.

Engeland, K., and L. Gottschalk (2002), Bayesian estimation of parameters in a regional hydrological model, *Hydrol. Earth Syst. Sci.*, *6*(5), 883–898.

Gupta, H. V., M. Thiemann, M. Trosset, and S. Sorooshian (2003), Reply to comment by K. Beven and P. Young on "Bayesian recursive parameter estimation for hydrologic models," *Water Resour. Res.*, *39*(5), 1117, doi:10.1029/2002WR001405.

Gupta, H. V., T. Wagener, and Y. Q. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813, doi:10.1002/hyp.6989.

Haario, H., E. Sakman, and J. Tammimien (1999), Adaptive proposal distribution for random walk Metropolis algorithm, *Comput. Stat.*, *14*(3), 375–395, doi:10.1007/s001800050022.

Haario, H., E. Sakman, and J. Tammimien (2001), An adaptive Metropolis algorithm, *Bernoulli*, *7*(2), 223–242, doi:10.2307/3318737.

Hall, M. J. (2001), How well does your model fit the data?, *J. Hydroinf.*, *3*(1), 49–55.

Hutchinson, P. D. (1990), Regression estimation of low flow in New Zealand, *Publ. 22*, 51 pp., Hydrol. Cent., DSIR Mar. and Freshwater, Christchurch, New Zealand.

Ibbitt, R. P. (1971), *Development of a Conceptual Model of Interception*, *Hydrol. Res. Prog. Rep. 5*, Min. of Works, Wellington, New Zealand.

Ibbitt, R. P., and R. Woods (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*, edited by H. A. J. van Lanen and S. Demuth, *IAHS Publ., 274*, pp. 189–196.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, *J. Hydrol.*, *320*(1–2), 173–186, doi:10.1016/j.jhydrol.2005.07.012.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis, *J. Hydrol.*, *320*(1–2), 187–201, doi:10.1016/j.jhydrol.2005.07.013.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, *331*(1–2), 161–177, doi:10.1016/j.jhydrol.2006.05.010.

Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*, W07401, doi:10.1029/2006WR005756.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *J. Hydrol.*, *330*(1–2), 368–381, doi:10.1016/j.jhydrol.2006.04.046.

Mantovan, P., E. Todini, and M. L. V. Martina (2007), Reply to comment by Keith Beven, Paul Smith and Jim Freer on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology," *J. Hydrol.*, *338*(3–4), 319–324, doi:10.1016/j.jhydrol.2007.02.029.

Manville, V., E. H. Newton, and J. D. L. White (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), 49–70, doi:10.1016/j.geomorph.2004.07.007.

Marshall, L., D. Nott, and A. Sharma (2004), A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modelling, *Water Resour. Res.*, *40*, W02501, doi:10.1029/2003WR002378.

McKerchar, A. I., and C. P. Pearson (1989), *Flood Frequency in New Zealand*, *Publ. Hydrol. Sect.*, vol. 20, 87 pp., Div. of Water Sci., Dep. of Sci. and Ind. Res., Christchurch, New Zealand.

Mein, R. G., and C. L. Larson (1973), Modeling infiltration during steady rain, *Water Resour. Res.*, *9*(2), 384–394, doi:10.1029/WR009i002p00384.

Montanari, A. (2005), Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, *41*, W08406, doi:10.1029/2004WR003826.

Newsome, P. F. J., R. H. Wilde, and E. J. Willoughby (2000), Land resource information system spatial data layers, technical report, Landcare Res. NZ Ltd., Palmerston North, New Zealand.

Snelder, T. H., and B. J. F. Biggs (2002), Multi-scale river environment classification for water resources management, *J. Am. Water Resour. Assoc.*, *38*(5), 1225–1240, doi:10.1111/j.1752-1688.2002.tb04344.x.

Sorooshian, S. (1981), Parameter estimation of rainfall run-off models with heteroscedastic streamflow errors—The noninformative data case, *J. Hydrol.*, *52*(1–2), 127–138, doi:10.1016/0022-1694(81)90099-8.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442, doi:10.1029/WR016i002p00430.

Sorooshian, S., Q. Duan, and H. V. Gupta (1993), Calibration of rainfall-runoff models: Application of global optimization to the Sacramento soil moisture accounting model, *Water Resour. Res.*, *29*(4), 1185–1194, doi:10.1029/92WR02617.

Tait, A. B., R. D. Henderson, R. W. Turner, and X. Zheng (2006), Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface, *Int. J. Climatol.*, *26*(14), 2097–2115, doi:10.1002/joc.1350.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resour. Res.*, *37*(10), 2521–2535, doi:10.1029/2000WR900405.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

Vrugt, J. A., H. V. Gupta, S. C. Dekker, S. Sorooshian, T. Wagener, and W. Bouten (2006), Application of stochastic parameter optimization to the Sacramento soil moisture accounting model, *J. Hydrol.*, *325*(1–4), 288–307, doi:10.1016/j.jhydrol.2005.10.041.

Vrugt, J. A., C. J. F. ter Braak, H. V. Gupta, and B. A. Robinson (2008), Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling (online), *Stochastic Environ. Res. Risk Assess.*, doi:10.1007/s00477-008-0274-y.

Wagener, T., and H. V. Gupta (2005), Model identification for hydrological forecasting under uncertainty, *Stochastic Environ. Res. Risk Assess.*, *19*(6), 378–387, doi:10.1007/s00477-005-0006-5.

M. Clark and H. McMillan, National Institute of Water and Atmospheric Research, 10 Kyle Street, Riccarton, Christchurch 8011, New Zealand. (h.mcmillan@niwa.co.nz)