Identification of rainfall–runoff models for operational applications

THORSTEN WAGENER[1] **& NEIL McINTYRE**[2]

1 *Department of Civil and Environmental Engineering, The Pennsylvania State University,*
226B Sackett Bldg., University Park, Pennsylvania 16802, USA
thorsten@engr.psu.edu

2 *Department of Civil and Environmental Engineering, Imperial College London,*
London SW7 2AZ, UK

Abstract The problem of selecting appropriate objective functions for the identification of a lumped conceptual rainfall–runoff model is investigated, focusing on the value of the model in an operational setting. A probability-distributed soil moisture model is coupled with a linear parallel routing scheme, and conditioned on rainfall–runoff observations from three catchments in the southeast of England. Using an abstraction control problem, which requires accurate simulation of the intermediate flow range, it is shown that using the traditional *RMSE* fit criterion, produces operationally sub-optimal predictions. This is true in the identification period, when applied to a testing period, and to proxy catchment data. Using a second case study of the Leaf River in Mississippi (USA), where the focus changes to predicting flood peaks over a specified threshold, also suggests that the relevant flood threshold should govern the objective function choice. It is concluded that, due to limitations in the structure of the employed model, it would be counter-productive to try to achieve a good all-round representation of the rainfall–runoff processes, and that a more empirical approach to identification may be preferred for specific forecasting problems. This leaves us with the question of how far hydrological realism should be sacrificed in favour of purpose-driven objective functions.

Key words hydrological forecasting; model identification; objective functions; rainfall–runoff models; uncertainty

Identification de modèles pluie–débit pour des applications opérationnelles

Résumé Le problème du choix des fonctions objectif appropriées pour l’identification d’un modèle pluie–débit conceptuel global est étudié, dans une perspective opérationnelle. Un modèle probabiliste de distribution de l’humidité du sol est couplé avec un modèle linéaire parallèle de propagation, et est calé avec des données de pluie–débit de trois bassins versants du sud-est de l’Angleterre. En se plaçant dans une problématique de contrôle des prélèvements, qui nécessite une simulation précise des variations des débits intermédiaires, il apparaît que l’utilisation du critère d’ajustement classique de l’écart quadratique moyen aboutit à des prévisions opérationnelles sous-optimales. Cela est vérifié pour la période d’identification, lors de l’application à une période de test et lors de la transposition à des données de bassins voisins. L’étude de cas de la Rivière Leaf, au Mississipi (Etats Unis), où la problématique est la prévision des pics de crue supérieurs à un seuil spécifié, suggère que le seuil de crue pertinent doit présider au choix de la fonction objectif. Il apparaît que, en raison de limitations dans la structure du modèle utilisé, il serait contre-productif d’essayer d’atteindre une bonne représentation globale des processus de transformation pluie–débit et qu’une approche plus empirique d’identification peut être préférée en cas de problèmes de prévision spécifiques. Cela pose la question du degré de sacrifice du réalisme hydrologique au profit de fonctions objectif orientées par la problématique.

Mots clefs prévision hydrologique; identification de modèle; fonctions objectif; modèles pluie–débit; incertitude

MOTIVATION AND SCOPE

Water resource and flood management hydrological problems are more and more approached using continuous time rainfall–runoff modelling (e.g. Lamb, 2000;

Open for discussion until 1 April 2006

Cameron *et al.*, 2001, Blazkova & Beven, 2002), rather than traditional statistical or event-based models (e.g. NERC, 1975; Pilgrim, 1987; Institute of Hydrology, 1999). This is because representing the recent basin history can add to the accuracy of the model output, and because the temporal aspects of the flow regime (e.g. cumulative flow volumes, effects of successive rainfall events on flood peaks) are central to effective management. Also, continuous time models allow real-time control of floods and abstractions, and can interact with, for example, meteorological, water quality and supply system models. Furthermore, data availability, modelling expertise, software developments and computational resources now make continuous time modelling accessible and affordable even for operational purposes (Lamb & Calver, 2002).

A general feature of operational rainfall–runoff model applications is that, while it is generally considered advantageous to reproduce the continuous hydrograph, an emphasis is usually put on a certain flow regime. Applications usually tend to be related to high flows (e.g. flood risk), medium flows (e.g. abstraction control), or low flows (e.g. drought impact). One example application is abstraction control, which is usually defined by a lower flow threshold (below which the water company is not licensed to abstract, i.e. minimum environmentally acceptable flow) and an upper flow threshold (above which abstraction is limited by pump capacity rather than river flows). The reproduction of the intermediate flow regime might therefore be of particular importance in this case. This emphasis might shift to peaks above a certain threshold if the aim of the modelling study is flood risk assessment or other aspects of the hydrograph depending on the modelling objective. Research applications on the other hand often aim at reproducing all aspects of the hydrograph to show that the full hydrological system under study is reproduced well.

The class of model structures most commonly used for continuous rainfall–runoff modelling can be defined as conceptual (Wheater *et al.*, 1993). In a conceptual model, a series of interacting storage elements reflect the simplification of the basin processes and states preferred by modellers. Associated parameters define, for example, the size of these storage elements, drainage rates, and the distribution of fluxes. Such models range widely in complexity (in terms of the number of parameters), although simpler models, with relatively few parameters, are often preferred because the calibration (identification) load is lower and they often achieve performances equal to those gained from using more complex models (Jakeman & Hornberger, 1993), while having more identifiable parameters (Wagener *et al.*, 2002, 2004; Littlewood, 2003). More complex models attempt to explicitly represent many of the actual physical processes and inputs affecting streamflow, but data requirements, parameter identification problems and computational cost tend to make these models less practical (Wheater *et al.*, 1993; O'Connell & Todini, 1996).

There is a large number of conceptual models; each formulated with the aim of improving the representation of a particular basin, or a particular type of basin response (Singh & Frevert, 2002a,b). Some studies have indicated that specific structures perform consistently better for basins that show a very well-defined response, e.g. they are baseflow or quick flow dominated (Wagener *et al.*, 2004). Other studies have implied that differences in performance between model structures are often difficult to discern (Uhlenbrook *et al.*, 1999; Lee *et al.*, 2004, 2005), in particular when considering data errors and the freedom of fit given by empirical parameter calibration. Further studies recognize that model structural error is inevitable, and that the symptoms are best managed by careful attention to parameter estimation techniques and consideration of the range of ultimate tasks of the model (Gupta *et al.*, 1998). The benefits of being able to explore the significance of alternative model structures and their inadequacies has led to the development of modelling frameworks that allow the quick implementation and comparison of alternative structures (e.g. Leavesley *et al.*, 1996; Wagener *et al.*, 2002).

The spatial and temporal heterogeneity of processes in a basin are aggregated into conceptual elements as described above. One consequence of this aggregation process in conceptual modelling is that most model parameters lose their direct physical interpretation and have to be estimated through a process of calibration or identification. During this calibration process, observed and simulated basin responses (usually streamflow) are compared and the model parameters are adjusted until the best possible match between the two time series has been achieved. This match-up can be done manually (e.g. Smith *et al.*, 2003), using mainly visual inspection of the hydrographs, or, more commonly, using automated procedures (e.g. Duan *et al.*, 1992; Vrugt *et al.*, 2003a). Manual calibration is limited in value because of its high human resource cost, and because it cannot be used to thoroughly explore the parameter space to identify a population of candidate parameter sets and to estimate uncertainty (Boyle *et al.*, 2000). For automated calibration, the residuals, i.e. the differences between observed and simulated time series, are commonly aggregated into a statistical measure or objective function (e.g. Nash & Sutcliffe, 1970). The calibration aim is then to find the parameter set that minimizes (or maximizes) this objective function. In recent years, this process of calibration has been evolved into one of identification in which the range of parameter sets is found that provides acceptable predictions for the purpose at hand (Gupta *et al.*, 2005).

For a number of inter-related reasons, it is not necessarily true that the parameter set which gives the optimum objective function value during calibration or identification is that which is most useful for extrapolating the rainfall–runoff relationship in time. Each parameter set that is in some way optimal fits the data in a way that compensates for the particular combination of model structure error and data error present during calibration, and may lead to significantly biased parameter estimates in doing so (e.g. Michaud & Sorooshian, 1994; Andreassian *et al.*, 2001). Unless the biases have the same compensatory effects during model extrapolation, significant errors in forecasts would be expected. Errors in forecasts will also occur partly due to parameter equifinality (Beven, 1993), whereby different parameter sets yield equally good objective function values during calibration (equally good within the tolerance achievable considering data accuracy and model structure deficiency), leaving doubt about which set should be applied for extrapolation. Another, closely related reason is that the aggregation of all residuals into a single objective function during calibration does not provide detailed information about where the model is failing to perform well, only about the fit over the whole flow regime (Gupta *et al.*, 1998). This loss of information adds to the equifinality problem (Wagener *et al.*, 2001), and means that parameter sets which give good performance in the operationally relevant parts of the hydrograph remain unknown. While it would be preferred if models could accurately simulate all aspects of the hydrograph, several researchers have suggested that it is not possible to find an individual parameter set that fits both high and low flows using currently available model structures (e.g. Gupta *et al.*, 1998; Boyle *et al.*, 2000;

Wagener *et al.*, 2001; Lee *et al.*, 2004). In addressing the problem of model structure inadequacies, multi-objective analysis has been used to assess the trade-offs between meeting different modelling objectives and the associated parameter uncertainty, although the significance of this in improving model forecasts has not yet been explored.

What is the significance of the recent findings regarding model structural inadequacy for operational model use, when reliable representation of a certain flow regime is the primary concern? During identification, what balance should be given between fitting the most relevant flow regime and attempting to represent the rainfall–runoff processes via fitting of the entire hydrograph? What is the consequence for this in terms of the hydrological realism of the parameters derived? Two case studies using basins in the UK and in the USA are investigated here to address these questions. After some justification for the underlying assumptions regarding the modelling problems mentioned above (e.g. no single parameter set to fit all model aspects), an example of a water resources study and some preliminary investigations into the optimum strategy for calibrating a model for flood risk estimation purposes are presented. The principles of operational testing of hydrological models laid out by Klemeš (1986), e.g. using split-test samples and proxy basins, are used.

OPERATIONAL RAINFALL–RUNOFF MODELLING

Several studies conclude that current model structures are not capable of reproducing the whole hydrograph with a single parameter set. What are the consequences of this problem for the operational use of these models? Is it appropriate to use a model that is calibrated exclusively for a specific flow regime, without considering the overall performance, and therefore the overall representativeness for the basin? Dooge (1972) came to the conclusion that "*...in the context of water resources development a model is something to be used rather than something to be believed*". This implies that there is an inconsistency between models used for operational (engineering) purposes and those that withstand scientific scrutiny with respect to the realism of their system representation (Wagener, 2004). *For operational applications, the question remains - is it better to be more right for the wrong reasons, or is it better to be less right, but simulate an overall more realistic hydrograph response and hence assume more confidence in model application?*

Based on similar ideas, Klemeš (1986) suggested an operational testing scheme for hydrological models that directly relates tests to the anticipated modelling task. This scheme has since been adopted and modified by several researchers (e.g. Seibert, 1999; Uhlenbrook, *et al.*, 1999; Refsgaard, 1997; Refsgaard & Knudsen, 1996). The conceptual framework proposed by Klemeš (1986) is based on three premises:

(a) The hydrological model is intended for an operational application, not a pure scientific investigation (e.g. for planning, design and operational decision making).
(b) The criteria for the evaluation of the model performance are defined with respect to the operational tasks.
(c) The criteria are calculated by comparing model estimates with observations.

The same premises are used in this study. An attempt is made to define performance criteria that closely reflect the modelling purpose and the operational situation encountered. Klemeš (1986) derived a scheme of operational hierarchical validation using increasingly more difficult modelling tests from his underlying premises:

(a) Split-sample test: using two separate time periods of the same time series for calibration and testing. The two periods show similar characteristics (e.g. climate or land use).
(b) Proxy-basin test: transferring a model calibrated in one basin to a similar second basin without recalibration. Manual adjustments of parameters based on different basin characteristics are allowed.
(c) Differential split-sample test: The same as (a), but the periods are different in characteristics.
(d) Proxy-basin differential split-sample test: combining tasks (b) and (c).

The problem of geographical transferability of a hydrological model is, of course, of major importance with respect to the modelling of ungauged basins (Sivapalan, 2003; Sivapalan *et al.*, 2003); with general transferability, in space and time, being the ultimate objective of hydrological modelling following Klemeš (1986).

CASE STUDIES

Basins

The two case studies use three basins in the southeastern UK and one in the USA. The first case study is of the River Medway in Kent, plus two of its tributaries—the rivers Eden and Eastern Rother. At the Teston flow gauge, the Medway basin area is 1256 km^2, while the Eastern Rother at Udiam is 206 km^2 and the Eden at Penshurst is 224 km^2. The naturalized mean flows at these locations are 11.2, 1.86 and 2.22 m^3 s^{-1}, respectively. Daily flow and potential evaporation data and basin-average daily rainfall data are available for all three basins for the period 1990–1996. The second case study is of the River Leaf in Mississippi, USA. The Leaf River basin area at the flow gauge north of Collins, Mississippi, is 1950 km^2, with a mean flow of 62.3 m^3 s^{-1}. Daily flow and potential evaporation data, and basin-average daily rainfall data, are available for the period 1948–1988.

Modelling and analysis tools

The rainfall–runoff modelling (RRMT) and Monte Carlo analysis (MCAT) toolboxes (Wagener *et al.*, 2002) are used in this study. All model structures that can be built in the RRMT consist of a soil moisture accounting (SMA) module that produces effective rainfall, which is subsequently input into a routing module to introduce translation effects. The RRMT contains a library of conceptual rainfall–runoff model components that can be combined to form appropriate model structures. All models are lumped (i.e. inputs and states are averaged over the basin area, resulting in a single output of streamflow at the basin outlet). The models are supported by various tools for model identification and Monte Carlo analysis techniques. Using the latter, a number of different objective functions can be simultaneously evaluated: for example the root mean squared error (*RMSE*) using the residuals over the whole time series, or just with respect to high flows, medium flows, or low flows. The RRMT can be coupled with the MCAT, which allows model sensitivities, and parameter and output uncertainties to be analysed.

The probability-distributed soil moisture model (PDM, see Moore, 1985), which has been found to give relatively good performance over a range of basin types (Young, 2002; Wagener *et al.*, 2004; Lee *et al.*, 2004) is selected as the SMA component in this study. This module represents the distribution of soil moisture capacity over the basin as a probability distribution. In this case a Pareto distribution is used, defined by shape parameter, *b*, and maximum soil moisture capacity over the whole basin, *cmax*. Evaporation is modelled as directly proportional to soil wetness, reaching a maximum of potential evapotranspiration at soil saturation, i.e. $c = cmax$. Runoff is generated by the distribution of exceeded capacities. The effective rainfall produced in this way is split (using parameter *q*) between fast and slow linear routing stores (with time constants *k*1 and *k*2). All the parameters, including their feasible ranges, are listed in Table 1.

Table 1 Ranges of parameter values sampled in identification.

Parameter	Lower bound	Upper bound	Unit
cmax	0	500	mm
b	0	2	-
q	0	1	-
*k*1	1	10	day
*k*2	10	400	day

The Medway, Eden and Eastern Rother basins

Identification and testing Uniform random sampling is used to explore the feasible parameter space for this parsimonious model structure and a number of objective functions are simultaneously evaluated. This involves a large number of parameter sets (20 000 in this case) being randomly sampled from an *a priori* feasible parameter space defined in Table 1. Using each sampled parameter set, the model is run and various measures of fit to the observed streamflow are calculated (see objective functions below). For each of these objective functions, the parameter set which provides the best fit is taken to be a good approximation of the optimum.

The data time series for the Medway, Eden and Eastern Rother are split into two periods to allow identification to be followed by a period for model testing or evaluation (Fig. 1). To better visualize medium- and low-flow behaviour, the time series has been transformed using the following Box-Cox power transformation (Kottegoda & Rosso, 1998):

$$y^*(\lambda) = \frac{y^\lambda}{\lambda} \qquad (1)$$

where y and y^* are the flow time series before and after transformation, respectively. The degree of transformation is defined by the value of λ for which a value of 0.3 has been selected as a suitable transformation for visualization purposes. Two equal periods of about three years are used in each case. The first period is initially used for

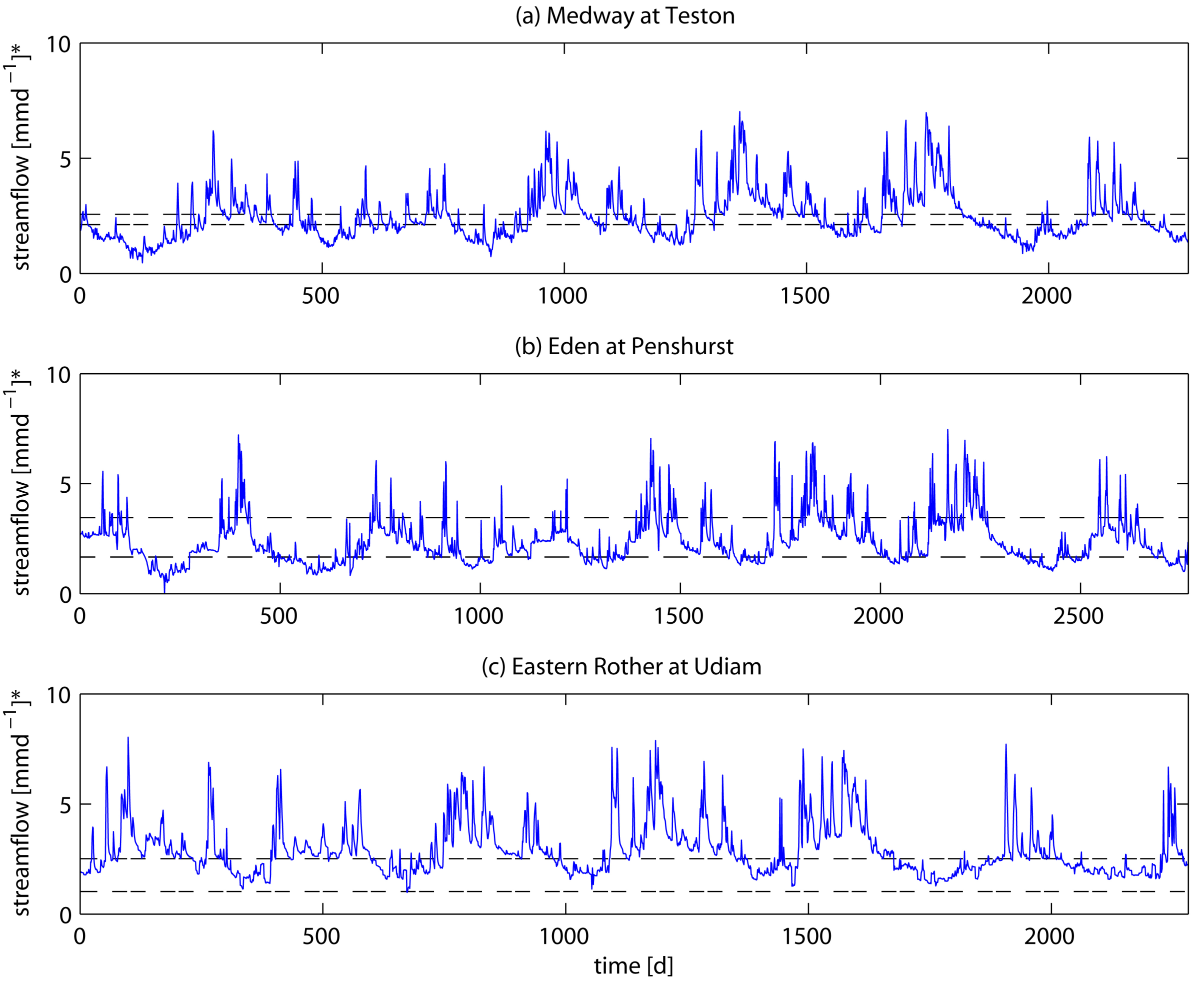

Fig. 1 Plots of the available time series for the rivers (a) Medway, (b) Eden and (c) Eastern Rother. The dashed lines define the medium flow range. The vertical dashed lines show the separation between the two different time periods. * Indicates that a Box-Cox transformation has been performed on the time series.

identification and the second for evaluating predictive performance, then *vice versa*. The issue of appropriate data lengths for model identification has been investigated by a variety of studies (e.g. Yapo *et al*., 1996; Jakeman *et al*., 1993; Sefton & Howarth, 1998), the general result being that the required length mainly depends on data quality, model complexity and climatic variability. For example, Yapo *et al*. (1996) state that an 8-year period is required for the result to be independent of climatic variability, whereas Jakeman *et al*. (1993) found that a 3-year period provides a good balance between a required minimum length for stable identification of the model parameters and changes in the system. For the present study, a 3-year period was deemed to be adequate for a useful comparison between alternative objective functions, given the parsimonious nature of the model and the focus on practical applications where calibration data are often limited.

The measures of fit used during identification and testing are the *RMSE* for low-, medium-, and high-flow ranges. These ranges are pre-defined by operational requirements for the different rivers and are mutually exclusive (see Table 2):

$$FL(\theta)=\sqrt{\frac{1}{L}\sum_{i=1}^{L}\left(m_i(\theta)-o_i\right)^2} \qquad (2)$$

$$FM(\theta)=\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(m_i(\theta)-o_i\right)^2} \qquad (3)$$

$$FH(\theta)=\sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(m_i(\theta)-o_i\right)^2} \qquad (4)$$

where θ is a parameter vector; *m* is the modelled flow; *o* is the observed flow; and *L*, *M* and *H* are the number of time steps for which the observed flow is within the low-, medium- and high-flow ranges respectively. For comparison with these three objective functions, two traditional objective functions are used that measure the overall fit: *RMSE* and the complement of the Nash-Sutcliffe Efficiency (*NSE*, Nash & Sutcliffe, 1970) using all *N* time steps (the complement of the *NSE* is used so that lower values are better):

$$RMSE(\theta)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(m_i(\theta)-o_i\right)^2} \qquad (5)$$

$$NSE(\theta)=\frac{\sum_{j=1}^{N}\left(m_j(\theta)-o_j\right)^2}{\sum_{j=1}^{N}\left(o_j-\bar{o}\right)^2} \qquad (6)$$

Table 2 Ranges of observed flows *o* used to define the *FL*, *FM* and *FH* objective functions.

Catchment	*FL* flow range (mm day^{-1})	*FM* flow range (mm day^{-1})	*FH* flow range (mm day^{-1})
Medway	$o \le 0.22$	$0.22 < o \le 0.42$	$o > 0.42$
Eden	$o \le 0.098$	$0.098 < o \le 1.13$	$o > 1.13$
Eastern Rother	$o \le 0.019$*	$0.019 < o \le 0.39$	$o > 0.39$

* There are only two time steps with flow values below this threshold. Therefore, *FL* cannot be used sensibly in this catchment.

Results and discussion The uniform random sampling procedure produced 20 000 parameters sets, plus the corresponding values of each of the five objective functions. These large samples allow statistical and visual analysis of inter-correlation between the objective functions. Figure 2 is a correlation matrix of the five objective functions (with the correlation coefficients shown in the upper diagonals and the data shown in the lower diagonals) using the results for the River Medway as an example. It shows that the (transformed) *NSE*, *RMSE* and *FH* objectives are well correlated and therefore using them both may add little to model identification. However, the other two objective functions (*FL* and *FM*) are much less correlated with *FH*, *NSE* and *RMSE*. Focusing on the relevant solutions (i.e. those with small values), for example for *FM versus RMSE*, a distinct trade-off between optimizing objectives is evident, whereby

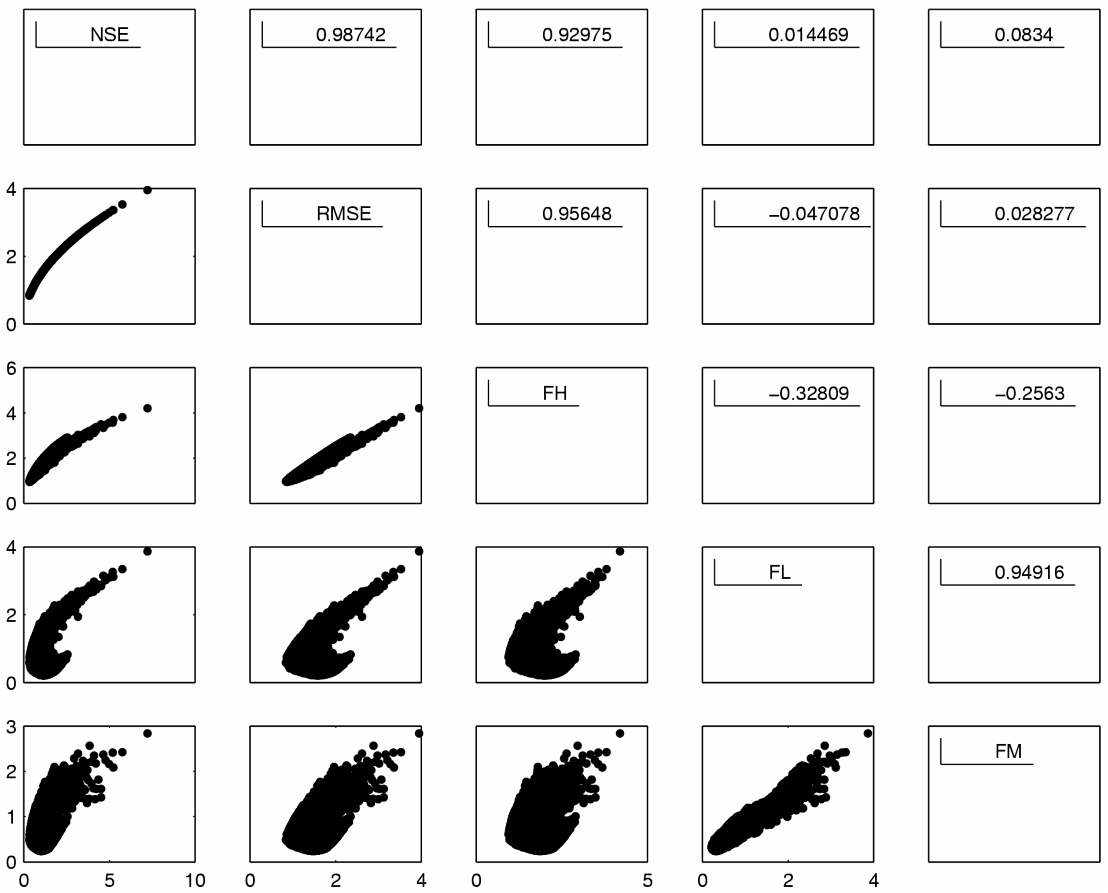

Fig. 2 Correlations of alternative objective functions derived using Monte Carlo sampling of 20 000 parameter sets for the identification period of the River Medway data. The individual parameter sets are plotted in the two-dimensional objective function space below the diagonal; above are the correlation coefficients.

improving on the value of one of them can only be done at the expense of degrading the other. Results are similar for the Eden and Rother basins.

These results confirm the previous findings of several researchers (e.g. Gupta *et al.*, 1998; Boyle *et al.*, 2000; Wagener *et al.*, 2001; Vrugt *et al.*, 2003b) that a conceptual model cannot be expected to deliver good calibration performance simultaneously across the full range of flows. To take the issue one step forward, the implications of this are now investigated for prediction rather than model identification, addressing the question “to what degree is the forecasting ability of the model improved (or deteriorated) by using an objective function that is exclusive to the range of flows that need to be forecast in operation?”. To do this, three forecasting tasks are defined—high flows, medium flows and low flows. As during identification, performance in achieving these tasks is measured using *FH*, *FM* and *FL* (defined in Table 2) during the 3-year test period, which can be considered a surrogate for a real forecasting problem. This is done independently for the rivers Medway, Eden and Eastern Rother.

Table 3 part (a) shows the best *RMSE*, *FH*, *FM* and *FL* performances during the identification (idn) period. It also shows the test period performance measured by each of these four objectives when (a) the same objective function has been employed

Table 3 Performances associated with changing from an *RMSE* objective function to a task-specific objective function.

River	Period	*RMSE*	*FH*	*FM*	*FL*
(a)					
Medway	1 [idn]	0.83	0.94 (0.95) *	0.24 (0.50)	0.21 (0.61)
	2 [test]	1.41	1.56 (1.61)	0.47 (0.54)	0.34 (0.75)
Eden	1 [idn]	1.43	1.66 (1.84)	0.39 (1.08)	0.12 (0.75)
	2 [test]	1.59	1.97 (2.10)	0.57 (1.07)	0.49 (0.81)
Rother	1 [idn]	1.41	1.52 (1.53)	0.17 (0.97)	-†
	2 [test]	1.36	1.45 (1.48)	0.30 (1.05)	
(b) After swapping identification and test periods:					
Medway	2 [idn]	1.28	1.42 (1.43) *	0.21 (0.68)	0.21 (0.85)
	1 [test]	0.95	1.05 (1.06)	0.40 (0.65)	0.38 (0.77)
Eden	2 [idn]	1.55	1.93 (1.99)	0.36 (1.09)	0.18 (1.02)
	1 [test]	1.54	1.73 (2.00)	0.59 (1.18)	0.59 (0.52)
Rother	2 [idn]	1.35	1.43 (1.45)	0.30 (1.11)	-†
	1 [test]	1.42	1.55 (1.52)	0.31 (1.04)	

* Values in brackets are derived using the parameter set with the best *RMSE* from the identification (idn) period.

† Not applied, see Table 2.

during identification, and (b) only *RMSE* has been employed during identification (shown in parentheses). Table 3 part (b) shows the result when the identification and test periods have been swapped around. Arguably, the results are not surprising, showing that performance in forecasting is consistently: (i) worse when the *RMSE* objective function has been used during identification rather than the task-specific objective function, and (ii) worse than the performance obtained during identification. The bias in cumulative volume was calculated over the identification period and is visualized in Fig. 3 to show the volumetric differences in results using the parameter sets optimal for the *FM* and those optimal for the *RMSE* objective functions. One can see that the absolute values for the bias increase by factors ranging between 3.5 and 5 for the different basins, and that using the *RMSE* leads to a severe underprediction of the flow volume in the region of interest (the medium-flow range).

The results in Table 3 and in Fig. 3 provide evidence that the best identification strategy for operational purposes, at least when a model is identified and applied in the same basin, might better be strictly defined by the task, rather than using a standard objective function such as *RMSE*. It might be expected that extrapolation to a different but similar basin would be a more demanding test of this principle (Klemeš, 1986). To test this, attention is focused on the *FM* objective function. Table 4 shows the forecasting *FM* performances when the models calibrated at each of the three basins (using both *FM* and *RMSE*) are extrapolated to the other two basins. This suggests the same thing—that the *RMSE* objective function should not be used for a task-specific application. Table 4 also suggests that, using *FM* as the identification objective function, the deterioration in performance associated with the extrapolation to new basins is minimal or (in the case of extrapolating the Eden and Medway models to the Rother) non-existent.

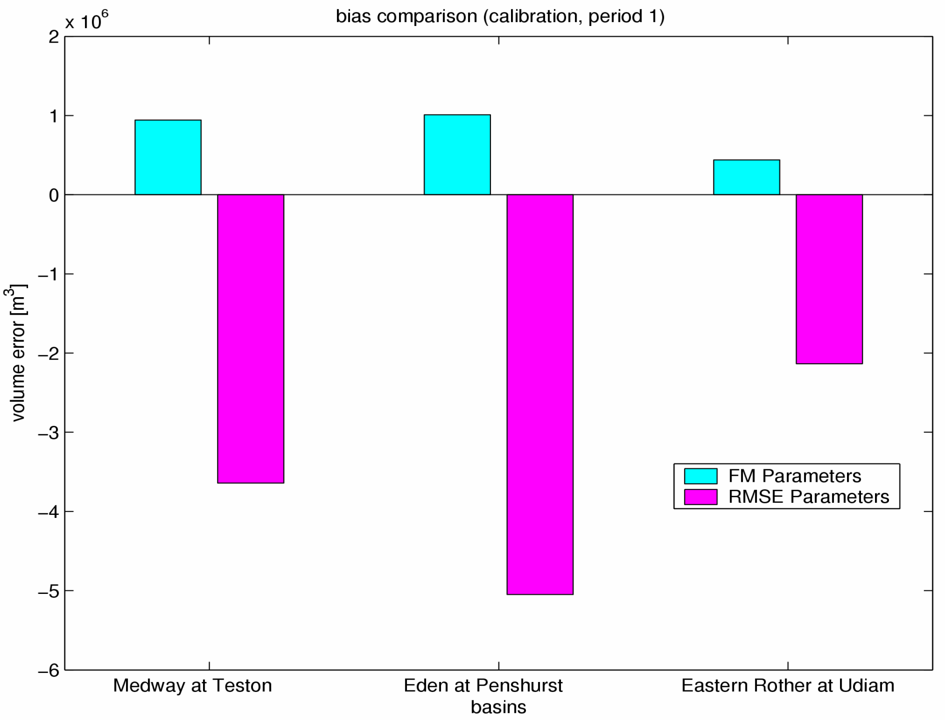

Fig. 3 Bias (i.e. volumetric error) comparison for the medium flow range for all three basins in period 1 (used as identification).

Table 4 Model *FM* performances in extrapolation to proxy catchments.

River used for identification	Period	*FM* performance at Rother	*FM* performance at Medway	*FM* performance at Eden
Rother	1 [idn]	0.17 (0.97) *	0.28 (2.10)	0.41 (2.97)
	2 [test]	0.30 (1.05)	0.39 (2.31)	0.52 (2.70)
Medway	1 [idn]	0.22 (0.35)	0.24 (0.50)	0.43 (1.21)
	2 [test]	0.35 (0.49)	0.47 (0.54)	0.62 (1.13)
Eden	1 [idn]	0.19 (0.43)	0.27 (0.51)	0.39 (1.08)
	2 [test]	0.35 (0.57)	0.47 (0.58)	0.57 (1.07)

* Values in brackets are derived using the parameter set with the best *RMSE* from the identification (idn) period; all other values are derived using the parameter set with the best identification period FM. Shaded areas show performances of parameter sets derived within the same catchment.

The Leaf River basin

Identification and testing For the Leaf River basin, where the focus is on high flows, only the *FH* objective (equation (2)) is used for identification. Again, a uniform random sampling procedure is performed selecting 20 000 parameter sets from the feasible parameter space given in Table 1. Eleven different realizations of *FH* are defined using a series of eleven thresholds (T_i; i = 1, 2, …, 11), resulting in eleven realizations of the optimum parameter set (α_i; i = 1, 2, …, 11). The largest threshold used (T_1) is marginally below the maximum flood observed in the identification period, and the lowest (T_{11}) includes all flows (equivalent to the *RMSE*). The remaining nine values are evenly distributed between T_1 and T_{11}. Two identification periods are used; the first five years in the data series (1948–1953, with a maximum flood peak of 1036 m^3 s^{-1}), and the last five years (1983–1988, with a maximum flood peak of 685 m^3 s^{-1}), and in each case the predictive performance is tested using the remaining 35 years (with a maximum flood peak of 1444 m^3 s^{-1}). This includes the likely problem that future predictions will often lie outside the range of response variability available for model identification.

The predictive performance of the alternative optimum models is measured as the magnitude of the difference between the observed number of peaks (N_o) over a defined flood threshold (T') during the test period (1948–1983) and the number predicted by the model (N_c), expressed as a percentage (P):

$$P(\alpha_j) = \frac{|N_o - N_c(\alpha_j)|}{N_o} \tag{7}$$

where subscript j refers to the jth realization of the optimum parameter set α. Performances under different values of T' are measured so that one can look for relationships between T' and the corresponding most appropriate value(s) of T (i.e. to investigate how specific to the flood forecasting task the definition of *FH* should be).

Results and discussion Figure 4 shows the variation of optimal parameter estimates over the eleven *FH* thresholds, using the 1983–1988 identification period. The parameter values have been converted to a [0, 1] scale (where the lower bound in Table 1 scales to 0 and the upper bound scales to 1). This indicates how the model has to adapt to new objective functions. Parameters *cmax* and *k*1 become notably lower as *T* increases, indicating that the model needs to store less water and respond faster when

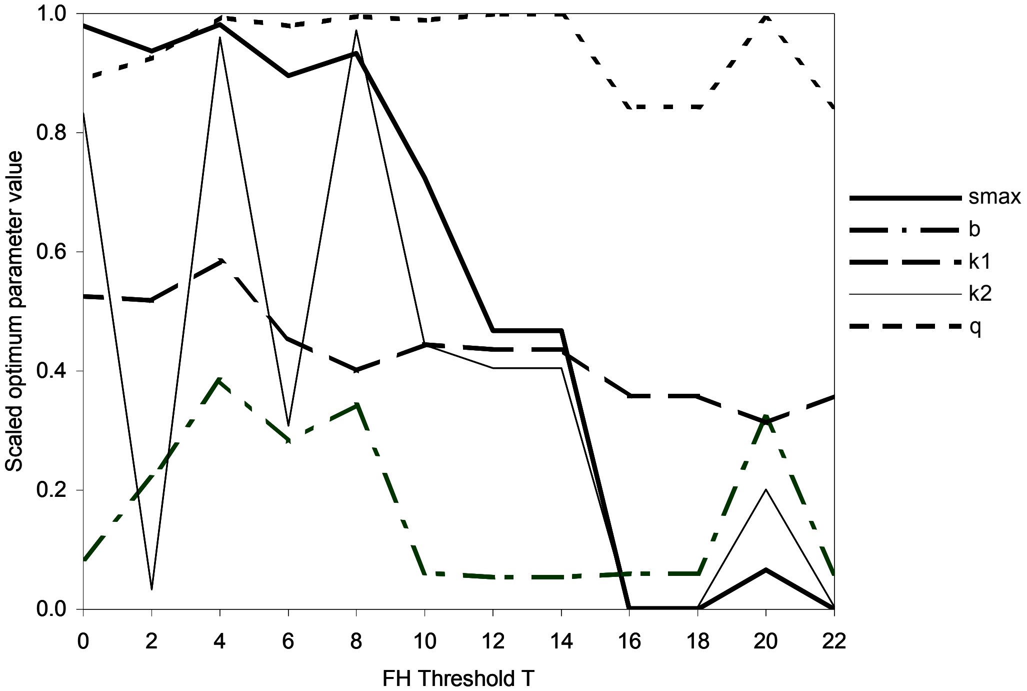

Fig. 4 Variation of calibrated parameter values (scaled) depending on the threshold used to define the *FH* objective function for the Leaf River (identification period 1983–1988).

it is required to match successively higher storm peaks only, excluding more of the recession periods. The variation in *cmax*, from its upper to its lower bound, is a good example of how the optimum parameter value can depend as much on the chosen objective function as it does on the physical basin characteristics. The more random fluctuations in *k*2, and less so in *b*, imply that these are not well identified using the *FH* objective function, which focuses on high-flow performance, while *q* is consistently high and therefore showing that a high proportion of effective rainfall is routed through the fast response store irrespective of the *T* value. The fact that *q* tends to be close to one for several thresholds suggests that a single store can be sufficient to reproduce the high-flow behaviour of the basin.

Table 5(a) and (b) shows the *P* performance values using the 1948–1953 and 1983–1988 identification periods, respectively, for every combination of *T* and *T′*. The most successful *T* for every *T′* is shaded, and values for which *T* = *T′* are outlined. Both sets of results generally imply that the threshold used to define *FH* at

Table 5 Performance of the Leaf River model obtained using different combinations of flood thresholds used to test the model and thresholds used to define the *FH* objective function.

(a) Leaf River, 1948–1953 identification period

Threshold used for model calibration (*T*)	Threshold used for testing model (*T′*)									
	40	36	32	28	24	20	16	12	8	4
50	–4	–8	–10	–29	–44	–86	–151	–281	–509	–685
40	0	3	–2	–8	–20	–34	–77	–165	–336	–479
36	1	3	4	0	–5	–18	–46	–123	–306	–468
32	1	3	4	0	–5	–18	–46	–123	–306	–468
28	1	3	4	0	–5	–18	–46	–123	–306	–468
24	1	3	4	0	–5	–18	–46	–123	–306	–468
20	1	3	4	0	–5	–18	–46	–123	–306	–468
16	1	3	4	0	–5	–18	–46	–123	–306	–468
12	2	5	6	7	7	9	–6	–46	–188	–361
8	2	5	6	6	10	10	11	2	–30	–22
4	2	5	6	8	10	13	19	15	–6	13
0	2	8	7	8	10	14	21	25	14	53

(b) Leaf River, 1983–1988 identification period

Threshold used for model calibration (*T*)	Threshold used for testing model (*T′*)									
	40	36	30	26	22	18	14	10	6	2
22	1	3	3	–1	–9	–28	–90	–198	–409	–332
20	0	3	–1	–6	–13	–25	–71	–142	–259	–235
18	1	3	3	–1	–9	–28	–90	–198	–409	–332
16	1	3	3	–1	–9	–28	–90	–198	–409	–332
14	2	4	5	3	7	–1	–6	–25	–73	39
12	2	4	5	3	7	–1	–6	–25	–73	39
10	2	5	5	4	7	5	1	–16	–35	38
8	2	5	5	2	8	11	11	–4	–59	–250
6	2	5	5	7	8	15	14	11	–30	–194
4	2	5	7	8	11	17	23	28	–13	–154
2	2	5	7	7	10	15	17	30	–2	–115
0	2	5	7	7	12	17	19	23	11	25

Notes: Best performance out of different identification thresholds is shaded grey. Cell representing the same identification and testing threshold is outlined.

identification should be close to, or slightly below, the threshold that defines the flood forecasting task. Using T_1 (equivalent to *RMSE*) as an identification objective function consistently gives unnecessarily poor predictions of flood frequency. In isolated cases it is noted that employing $T = T'$ would have resulted in an even worse prediction (e.g. $T' = 20$ in Table 5(a)). This might imply that a good representation of the basin state on the day (or days) before the flood event can be important.

Thresholds $T' = 50$ and $T' = 60$ were also applied, but the model failed to identify the incidences of floods above these thresholds (one and two incidences respectively). This also leads to doubt about the adequacy of the observed rainfall, as well as about the employed lumped model. Alternative model structure, or adjustments of the rainfall input (e.g. Lamb, 2000) might help to reduce this problem.

One limitation of this numerical experiment is that it has been assumed that the inputs to the identification and testing periods are independent hydrological time series. Arguably, because the input time series are likely to have measurement biases that are consistent over both periods, then they may not be independent. The model may perform well in testing because the "optimal" model is compensating for the same data biases as it was during identification, so that $T \approx T'$ is hardly a surprising result. In that case, a further stage of analysis would be to assess robustness of objective functions to input biases.

CONCLUDING DISCUSSION

This paper has attempted to expose the importance of design of objective functions for lumped conceptual rainfall–runoff model identification in operational settings. Two sets of numerical experiments have been conducted—one focusing on forecasting an intermediate range of flows (using the Rother, Eden and Medway basins in the UK), and the other focusing on forecasting frequency of peaks over a threshold (using the River Leaf in the USA). In both cases, the probability distributed soil moisture accounting model was used to model effective rainfall and two parallel linear stores were used to route this to streamflow at the basin outlet. Generally speaking, the results imply that, at least when using relatively simple rainfall–runoff models, attempting to achieve a good all-round fit using traditional objective functions such as *RMSE* may be counter-productive for specified forecasting tasks.

Results from the first study showed that there was a consistent and substantial deterioration in predictive performance if a traditional objective function (*RMSE*) was employed, which is usually assumed to provide a reasonable overall fit, rather than one specifically designed for intermediate flows. This loss in performance was, in general, much greater than the loss associated with extrapolating the model over time and/or to neighbouring basins.

The second study illustrated that the range of floods considered during identification should be defined by the flood magnitude most relevant to the forecasting task. Results provided some evidence that the threshold used to define the identification objective function *FH* (equation (3)) should be set slightly lower than the relevant flood threshold, if possible. It is speculated that this allows the period immediately prior to the defined flood event to be well-represented, without unduly weighting *FH* to periods of non-flood flows. This raises the question whether further improvements could be made by using a more sophisticated version of equation (3) that includes only the rising limb of the floods rather than requiring the model to fit the recessions, and this seems a priority for further research.

There has been a general assumption in this study (and most other studies) that there are no significant errors in the streamflow data used to evaluate the models. Should this assumption be false, the flow data may not provide a good basis for measuring model performance. In particular, if the data are less accurate in particular flow ranges (e.g. due to difficulties in gauging and naturalizing low flows) then the differences in performance and parameter values between ranges may be due to data error, as well as model structure error. Similar issues would arise if rainfall biases were considered. The effect of flow and rainfall data errors on inferences about model reliability for different modelling tasks, and on objective function design, is a priority area for future research. Another issue not covered explicitly within this paper is parameter uncertainty. It has been assumed that the 20 000 samples provided a good enough approximation of the optimum parameter set for the parsimonious model used, for the purposes of comparing identification strategies. A possible future extension of the research would be to make the comparisons using stochastic representations of the models, considering confidence limits on results as well as deterministic performance, in order to consider the significance of parameter set equifinality in this context.

The observations made in this study arguably suggest that the employed rainfall–runoff model is acting empirically rather than representing the basin dynamics, due to its failure to optimally simulate different flow ranges using one parameter set. At the same time, the reasonable forecasting performance (relative to identification performance) indicates that this empiricism may be sufficient, at least for the limited complexity of modelling tasks specified here. Tasks that require a better representation of the continuous basin response, for example to model ecological responses to flow, or as part of a more integrated multi-purpose basin model, can be expected to pose a more challenging model identification problem. In general, there seems to be a discrepancy between optimal tools for engineering purposes and optimal tools for scientific investigations of the overall basin behaviour. The distance between the two is an indicator of the scientific progress that is still required in this field.

Acknowledgements Constructive criticism by two anonymous reviewers was appreciated and helped to improve the paper. Partial support for Thorsten Wagener was provided by a fellowship of the German Academic Exchange Service (DAAD) and by SAHRA (Center for the Sustainability of semi-Arid Hydrology and Riparian Areas) under the STC Program of the National Science Foundation (Agreement no. EAR-9876800). Thanks to Southern Water for the data set used in this study. The software tools used in this study can be downloaded free of charge from the Hydroarchive web site at http://www.sahra.arizona.edu/software. The Abstract was translated into French by Marie-Laure Segond of Imperial College London.

REFERENCES

Andreassian, V., Perrin, C., Michel, C., Usart-Sanchez, I. & Lavabre, J. (2001) Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models. *J. Hydrol.* **250**(1/4), 206–223.

Beven, K. J. (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. *Adv. Water Resour.* **16**, 41–51.

Blazkova, S. & Beven, K. J. (2002) Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty). *Water Resour. Res.* **38**(8). DOI 10.1029/2001WR000500.

Boyle, D. P., Gupta, H. V. & Sorooshian, S. (2000) Toward improved calibration of hydrological models: combining the strengths of manual and automatic methods. *Water Resour. Res.* **36**(12), 3663–3674.

Cameron, D., Beven, K. J. & Naden, P. (2001) Flood frequency estimation under climate change (with uncertainty). *Hydrol. Earth Syst. Sci.* **4**(3), 393–405.

Dooge, J. C. I. (1972) Mathematical models of hydrological systems. In: *Proc. Int. Symp. on Modelling Techniques in Water Resources Systems*, vol. 1 (ed. by A. K. Biswas), 171–189. Environment Canada, Ottawa, Canada.

Duan, Q., Gupta, V. K. & Sorooshian, S. (1992) Effective and efficient global optimisation for conceptual rainfall–runoff models. *Water Resour. Res.* **28**, 1015–1031.

Gupta, H. V., Sorooshian, S. & Yapo, P. O. (1998) Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resour. Res.* **34**(4), 751–763.

Gupta, H. V., Beven, K. J. & Wagener, T. (2005) Model calibration and uncertainty estimation. In: *Encyclopedia of Hydrological Sciences* (ed. by M. G. Anderson). John Wiley & Sons Ltd., Chichester, UK (in press).

Institute of Hydrology (1999) *Flood Estimation Handbook*. 5 vols. Institute of Hydrology (now CEH), Wallingford, UK.

Jakeman, A. J. & Hornberger, G. M. (1993) How much complexity is warranted in a rainfall–runoff model? *Water Resour. Res.* **29**(8), 2637–2649.

Jakeman, A. J., Chen, T. H., Post, D. A., Hornberger, G. M., Littlewood, I. G. & Whitehead, P. G. (1993) Assessing uncertainties in hydrological response to climate at large scale. In: *Macroscale Modelling of the Hydrosphere* (Proc. Yokohama Symp., July 1993), 37–47. IAHS Publ. 214, IAHS Press, Wallingford, UK.

Klemeš, V. (1986) Operational testing of hydrological simulation models. *Hydrol. Sci. J.* **31**(1), 13–24.

Kottegoda, N. T. & Rosso, R. (1998) *Statistics, Probability, and Reliability for Civil and Environmental Engineers.* McGraw-Hill, New York, USA.

Lamb, R. (2000) An approach to the calibration of a conceptual rainfall–runoff model for flood frequency estimation by continuous simulation. *Water Resour. Res.* **35**, 3103–3114.

Lamb, R. & Calver, A. (2002) Continuous simulation as a basis for national flood frequency estimation. In: *Continuous River Flow Simulation: Methods, Applications and Uncertainties* (ed. by I. Littlewood), 67–75. British Hydrological Society Occasional Paper no. 13, Wallingford, UK.

Leavesley, G. H., Restrepo, P. J., Markstrom, S. L., Dixon, M. & Stannard, L. G. (1996) The modular modelling system (MMS): user's manual. *US Geol. Survey Open-File Report 96–151*, Denver, USA.

Lee, H., McIntyre, N., Wheater, H. & Young, A. (2005) Selection of conceptual models for regionalisation of the rainfall–runoff relationship. *J. Hydrol.* in press.

Lee, H., McIntyre, N., Wheater, H., Young, A. & Wagener, T. (2004) Assessment of rainfall–runoff model structures for regionalisation purposes. In *Hydrology—Science and Practice for the 21st* Century, vol. 1 (Proc. British Hydrological Society Int. Conf., London, July 2004), 302–308. British Hydrological Society, Wallingford, UK.

Littlewood, I. G. (2003) Improved unit hydrograph identification for seven Welsh rivers: implications for estimating continuous streamflow at ungauged sites. *Hydrol. Sci. J.* **48**(5), 743–762.

Michaud, J. D. & Sorooshian, S. (1994) Effect of rainfall-sampling errors on simulations of desert flash floods. *Water Resour. Res.* **30**(10), 2765–2775.

Moore, R. J. (1985) The probability-distributed principle and runoff production at point and basin scales. *Hydrol. Sci. J.* **30**(2), 273–297.

Nash, J. E. & Sutcliffe, J. V. (1970) River flow forecasting through conceptual models, Part I. A discussion of principles. *J. Hydrol.* **10**, 282–290.

NERC (Natural Environment Research Council) (1975) *Flood Studies Report*. NERC, London.

O'Connell, P. & Todini, E. (1996) Modelling of rainfall, flow and mass transport in hydrological systems: an overview. *J. Hydrol.* **175**(1/4), 3–16.

Pilgrim, D. H. (1987) *Australian Rainfall and Runoff, A Guide to Flood Estimation*. Inst. of Engrs of Australia, Canberra, Australia.

Refsgaard, J. C. (1997) Parameterisation, calibration and validation of distributed hydrological models. *J. Hydrol.* **198**, 69–97.

Refsgaard, J. C. & Knudsen, J. (1996) Operational validation and intercomparison of different types of hydrological models. *Water Resour. Res.* **32**(7), 2189–2202.

Sefton, C. E. M. & Howarth, S. M. (1998) Relationships between dynamic response characteristics and physical descriptors of catchments in England and Wales. *J. Hydrol.* **211**(1/4), 1–16.

Seibert, J. (1999) Conceptual runoff models—fiction or representation of reality? PhD Dissertation, University of Uppsala, Sweden.

Singh, V. P. & Frevert, D. K. (eds) (2002a) *Mathematical Models of Large Watershed Hydrology*, vol. 1. Water Resources Publications, Highlands Ranch, Colorado, USA.

Singh, V. P. & Frevert, D. K. (eds) (2002b) *Mathematical Models of Small Watershed Hydrology*, vol. 2. Water Resources Publications, Highlands Ranch, Colorado, USA.

Sivapalan, M. (2003) Predictions in ungauged basins: a grand challenge for theoretical hydrology. *Hydrol. Processes* **17**, 3163–3170. DOI 10.1002/hyp.5155.

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J. J., Mendiodo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S. & Zehe, E. (2003) IAHS Decade on Prediction in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrol. Sci. J.* **48**(6), 857–880.

Smith, M. B., Laurine, D. P., Koren, V. I., Reed, S. M. & Zhang, Z. (2003) Hydrologic model calibration in the National Weather Service. In: *Calibration of Watershed Models* (ed. by Q. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau & R. Turcotte). Water Science and Application 6, American Geophysical Union, Washington DC, USA.

Uhlenbrook, S., Seibert, J., Leibundgut, C. & Rohde, A. (1999) Prediction uncertainty of conceptual rainfall–runoff models caused by problems in identifying model parameters and structures. *Hydrol. Sci. J.* **44**(5), 779–797.

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

Vrugt J. A., Gupta, H. V., Bastidas, L. A., Bouten, W. & Sorooshian, S. (2003b) Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resour. Res.* **39**(8), 1214. DOI 10.1029/2002WR001746.

Wagener, T. (2004) Evaluation of catchment models. *Hydrol. Processes* **17**, 3375–3378.

Wagener, T., Boyle, D. P., Lees M. J., Wheater, H. S., Gupta, H. V. & Sorooshian, S. (2001) A framework for the development and application of hydrological models. *Hydrol. Earth Syst. Sci.* **5**(1), 13–26.

Wagener, T., Lees, M. J. & Wheater, H. S. (2002) A toolkit for the development and application of hydrological models. In: *Mathematical Models of Large Watershed Hydrology*. (ed. by Singh V. P. & D. K. Frevert), 91–140. Water Resources Publications LLC, Highlands Ranch, Colorado, USA.

Wagener, T., Wheater, H. S. & Gupta, H. V. (2004) *Rainfall–Runoff Modelling in Gauged and Ungauged Catchments*. Imperial College Press, London, UK.

Wheater, H. S., Jakeman, A. J. & Beven, K. (1993) Progress and destinations in rainfall–runoff modelling. In: *Modelling Change in Environmental Systems* (ed. by A. J. Jakeman, M. B. Beck & M. J. McAleer), 101–132. John Wiley & Sons Ltd, Chichester, West Sussex, UK.

Yapo, P. O., Gupta, H. V. & Sorooshian, S. (1996) Automatic calibration of conceptual rainfall–runoff models: sensitivity to calibration data. *J. Hydrol*. **181**, 23–48.

Young, A. (2002) River flow simulation within ungauged catchments using a daily rainfall–runoff model. In: *Continuous River Flow Simulation: Methods, Applications and Uncertainties* (ed. by I. Littlewood), 31–38. British Hydrological Society Occasional Paper no. 13, Wallingford, UK.

Received 15 March 2004; accepted 21 July 2005