HYDROLOGICAL PROCESSES
Hydrol. Process. **24**, 2840–2850 (2010)
Published online 2 July 2010 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.7698

Diagnostic evaluation of conceptual rainfall–runoff models using temporal clustering

N. J. de Vos,[1]* T. H. M. Rientjes[2] and H. V. Gupta[3]

[1] *Water Resources Section, Delft University of Technology, PO Box 5048, 2600 GA Delft, The Netherlands*
[2] *Department of Water Resources, Twente University, PO Box 6, 7500 AA Enschede, The Netherlands*
[3] *Department of Hydrology and Water Resources, The University of Arizona, Tucson, AZ 85721, USA*

Abstract:

Given the structural shortcomings of conceptual rainfall–runoff models and the common use of time-invariant model parameters, these parameters can be expected to represent broader aspects of the rainfall–runoff relationship than merely the static catchment characteristics that they are commonly supposed to quantify. In this article, we relax the common assumption of time-invariance of parameters, and instead seek signature information about the dynamics of model behaviour and performance. We do this by using a temporal clustering approach to identify periods of hydrological similarity, allowing the model parameters to vary over the clusters found in this manner, and calibrating these parameters simultaneously. The diagnostic information inferred from these calibration results, based on the patterns in the parameter sets of the various clusters, is used to enhance the model structure. This approach shows how diagnostic model evaluation can be used to combine information from the data and the functioning of the hydrological model in a useful manner. Copyright © 2010 John Wiley & Sons, Ltd.

KEY WORDS calibration; catchment modeling; clustering; diagnostic evaluation

Received 2 June 2009; Accepted 18 March 2010

MOTIVATION

Modelling of catchment systems is among the most important and complex tasks in hydrological research. This complexity is the result of the various kinds of media in which water travels and the interactions among processes at various scales in time and space. A variety of model approaches have been suggested, ranging from simple empirical to complex physically based methods. Conceptual rainfall–runoff models are probably the most popular models for mimicking the transformation from precipitation over a catchment to streamflow, because they rely on simplified descriptions of physical processes and offer a combination of ease of development and use, good performance and transparency. Conceptual models commonly comprise a number of soil water reservoirs and routing routines which represent the different domains that mediate various runoff processes. Popular examples include the Sacramento soil moisture accounting model (Burnash, 1995) and the HBV model (Bergström, 1976; Lindström *et al.*, 1997). Determining *a priori* which conceptual model structure is most appropriate for a given situation remains a challenging problem in hydrology (Clark *et al.*, 2008).

The process of developing, calibrating and evaluating conceptual rainfall–runoff models carries a significant degree of subjectivity. This subjectivity follows mainly from various ways in which hydrological modellers can choose their preferred methods to evaluate model performance. Gupta *et al.* (2008) argued that the purpose of all evaluation must be diagnostic in focus, meaning that modellers need to identify those components of the model, which when assumed to be functioning normally, will explain the discrepancy between model output and observed data. This can be approached through the identification of the so-called signature information from the data. In the same article, the authors discuss how model evaluation consists of three aspects:

1. Quantitative evaluation of model output, including statistical measures that express the difference between model output and observation time series.
2. Qualitative evaluation of model consistency, such as model sensitivity tests and visual inspections of model behaviour.
3. Qualitative evaluation of model form and function, which implies a subjective expression of the degree in which the model structure and working comply with the real world system.

Quantitative model evaluation usually uses one or more statistical measures, in which commonly (i) the difference between a series of measurement data and a series of model outputs is transformed (e.g. power transformation, weighting) and then (ii) the series of transformed differences is aggregated into a single value (e.g. by taking the mean squared error). These operations reflect a subjective choice by highlighting specific aspects of the

*Correspondence to: N. J. de Vos, Water Resources Section, Delft University of Technology, PO Box 5048, 2600 GA Delft, The Netherlands. E-mail: njdevos@gmail.com hydrograph. Gupta *et al.* (1998) argued that using too few such aspects implies a loss of information because the calibration problem inherently involves many criteria. Following the development of effective and efficient algorithms (see Tang *et al.*, 2006), the power of the multi-criteria approach has been demonstrated in a number of hydrologic model calibration studies (Yapo *et al.*, 1998; Boyle *et al.*, 2000; Vrugt *et al.*, 2003a; Khu and Madsen, 2005; Kashif Gill *et al.*, 2006; de Vos and Rientjes, 2007, 2008). However, in both single-criterion and multi-criteria approaches, when the residuals are aggregated into each statistic the information regarding model behaviour and performance that is embedded in the time dimension is largely ignored. Arguably, it is questionable to ignore the time dimension given the predominantly dynamic nature of catchment runoff behaviour.

In rainfall–runoff modelling, complex dynamic relationships are commonly simplified and approximated using time-invariant parameters. For example, in many lumped storage-based conceptual models, the idea of changes in variable source areas that generate runoff is implicitly considered through the form of the storage–discharge relationship but changes in model parameter values to reflect dynamic changes that relate to such mechanisms are rarely (if ever) made. Another example of a dynamic process that is difficult to describe with time-invariant model parameters is macropore flow. After prolonged dry periods soil cracks can emerge, which cause a temporary increase in hydraulic conductivity of certain soil layers. Finally, many biological processes have a seasonal component and could be more adequately represented using time-invariant model parameters.

Given the structural shortcomings of conceptual rainfall–runoff models and the use of time-invariant model parameters, it seems reasonable to proceed with a hypothesis that these parameters may represent broader aspects of the rainfall–runoff relationship than merely the static catchment characteristics they are commonly supposed to quantify. By relaxing the common assumption of time-invariance of parameters, one can therefore attempt to obtain information from parameter variation about the dynamics of model behaviour and performance. Several studies have reported on the calibration and evaluation of models with time-variant parameters, and on the subsequent extraction of information from the results. Wagener *et al.* (2003) investigated the identifiability and evolution of model parameters over time for a very simple storage-based runoff model using the Dynamic Identifiability Analysis (DYNIA) approach. Using a moving time window of fixed length over which parameter sensitivity and model performance were assessed, the approach suggested significant time variation of parameters and also revealed that such information could be used to develop insight into the model form and function. Choi and Beven (2007) used temporal clustering to identify periods of hydrological similarity. They subsequently evaluated predictions of Monte Carlo realizations of TOPMODEL parameter sets both within these periods and on multiple objective functions. The behavioural parameter sets were shown to vary significantly over both clusters and criteria. Moreover, no set was found that performed well on all clusters or on all criteria, indicating deficiencies in model structure. Another approach in which the time-invariance assumption of model parameters is relaxed, is to build models for specific parts of the hydrograph and find optimal ways to combine the results of the local models (Hsu *et al.*, 2002, Oudin *et al.*, 2006, Fenicia *et al.*, 2007, Marshall *et al.*, 2007).

GOALS AND SCOPE

In this article, we develop and examine an approach to diagnostic evaluation and improvement of a prior hydrological model structure by extracting temporal signature information via an augmented calibration procedure. The approach is based on the premise that deficiencies of the model structure cause the model parameters to vary with the hydrological modes of the system (if allowed to do so) to compensate for the effects of the model structural error. The main goals of this study are twofold: (1) to develop and test a method for identifying signature information in the form of time-variant model parameter values for a rainfall–runoff model of a meso-scale catchment and (2) to extract and use diagnostic information from the modelling results to improve the model structure.

To accomplish the first goal, a temporal clustering approach was devised to partition the historical data into several (here 12) periods of hydrological similarity. The model parameters were then permitted to vary with time in a discrete manner, taking on different values for each period of hydrological similarity, but remaining constant within each period, i.e. in our example each parameter can take on 1 of 12 different values over time, with the value corresponding to the temporal cluster mode active at that time. The goal of the approach is to see if the parameter variation can be related in some systematic manner to the magnitude of the system variables used to characterize periods of hydrological similarity, and to thereby make diagnostic inferences leading to improvements in the proposed hypothesis regarding the underlying structure of the system.

By applying the clustering procedure to observed data, we are able to use a physical basis in our dynamic analysis, effectively overcoming a main shortcoming of the previously mentioned approach by Wagener *et al.* (2003) who used an arbitrary time window of fixed length. The clustering procedure proposed here has similarity to the approach used by Choi and Beven (2007), but extends it by the logical step of actually interpreting the clustering results diagnostically so as to make improvements to the model structure. Note also that our main goal is different from the approach mentioned earlier on combining local models. Although using similar principles, our goal is to improve on a preconceived model structure rather than to identify and combine local model components.

A dataset from the Leaf River catchment, located north of Collins, Mississippi, USA was used for this study.

This humid catchment has a size of around 1944 km^2. Daily time series of precipitation (mm/day), potential evaporation (mm/day) and discharge (m^3/s) are available for the period October 1948 to September 1988. Roughly, a third of the data (1 October 1948 to 30 September 1962) was used for calibration, and the rest for model evaluation. In the absence of observations regarding moisture storage, a synthetic time series of soil moisture was generated using the simple soil moisture reservoir component of the GR4J lumped conceptual R–R model (Edijatno *et al.*, 1999; Perrin *et al.*, 2003). Rainfall and potential evaporation serve as model input to this model and a time series of lumped soil moisture was generated as output. In the GR4J approach the only parameter that requires estimation is the reservoir's maximum capacity, A, for which a value of 400 mm was chosen. The above procedure is similar to the one presented in de Vos and Rientjes (2005, 2007).

METHODS

Temporal cluster analysis

Introduction to cluster analysis. Cluster analysis is concerned with exploring datasets to assess whether they can be summarized meaningfully in terms of a relatively small number of clusters of objects which resemble each other and which are different in some respects from the objects in other clusters (Jain *et al.*, 1999; Everitt *et al.*, 2001). The concept has been applied in hydrology to cluster, e.g. precipitation fields (Lauzon *et al.*, 2006), watershed conditions (Liong *et al.*, 2000), hydrological homogeneous regions (Frapporti *et al.*, 1993; Hall and Minns, 1999), and also for regionalization purposes (Burn, 1989; Srinivas *et al.*, 2008) and for hydrological model evaluation and identification (Herbst *et al.*, 2009).

In this work, we attempt to find periods of hydrological similarity by temporal clustering of hydrological data. Other works that used clustering to this end include Hsu *et al.* (2002), Choi and Beven (2007), Reusser *et al.* (2009) and Toth (2009). By this approach, the information contained in the dimension of time is compressed into a discrete set of clusters and its information can be meaningfully and conveniently summarized.

Cluster inputs. Three time series were chosen for the first cluster analysis: (1) precipitation (P), (2) the 10-day moving average of the precipitation (P_{ma}) and (3) the GR4J-simulated soil moisture (S). These variables represent information regarding the recent input, memory and storage dynamics of the catchment, respectively. The input to the clustering algorithm consists of the simultaneous variable values at each time t, so the algorithm had three inputs.

Clustering algorithm. The k-means clustering algorithm involves the calculation of the centroid of a fixed number of clusters. This is usually done as proposed by Lloyd (1982). The method can be summarized as follows:

1. Randomly choose k initial centroids $Z = \{z_1, \ldots, z_k\}$.
2. Set each cluster N_i to be the points in X that are closer to z_i than to any other centroid.
3. Set each z_i to be the centroid of all points in N_i.
4. Repeat steps 2 and 3 until Z is stable.

The proximity measure used to determine the closeness of points to centroids was the Euclidian distance. The k-means++ seeding method (Arthur and Vassilvitskii, 2007) was used to choose the initial centroids with a probability proportional to the density of points. Arthur and Vassilvitskii (2007) show that this approach can significantly reduce errors and improve convergence speed of the algorithm.

The *a priori* choice of the number of clusters was made by running the clustering procedure for 2–30 clusters and studying the partition index (Bensaid *et al.*, 1996) and the Xie–Beni index (Xie and Beni, 1991), as shown in Figure 1. These are well known validity measures for expressing the quality of a clustering algorithm's partition of the data. The number of clusters was set to 12 based on the fact that both measures did not significantly improve for a larger number of clusters. In comparison, Choi and Beven (2007) found 15 clusters to be appropriate for their dataset.

Conceptual rainfall–runoff model

The five-parameter conceptual HyMod rainfall–runoff model, shown in Figure 2, was used for this study. This model is based on the probability distribution model by Moore (1985) and was introduced by Boyle (2000). It was applied more recently by Wagener *et al.* (2001) and Vrugt *et al.* (2003b) among others. HyMod consists of a simple two-parameter rainfall excess model, in which it is assumed that the soil moisture storage capacity C varies across the catchment and, therefore, that the proportion of the catchment with saturated soils varies over time. The following distribution function describes the fraction

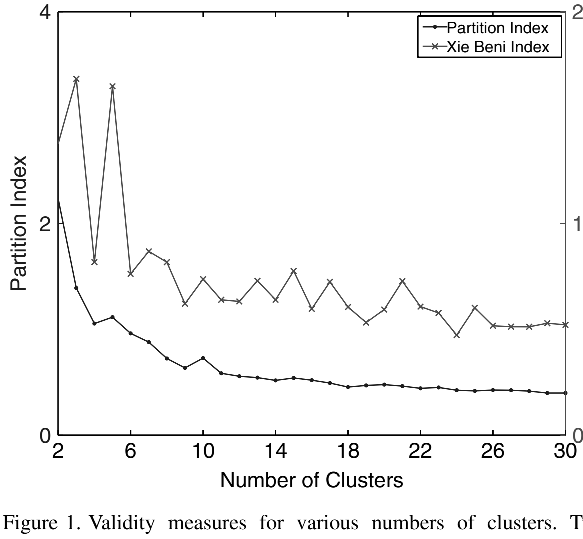

Figure 1. Validity measures for various numbers of clusters. Twelve clusters were chosen because both validity measures do not significantly decrease for a larger number of clusters

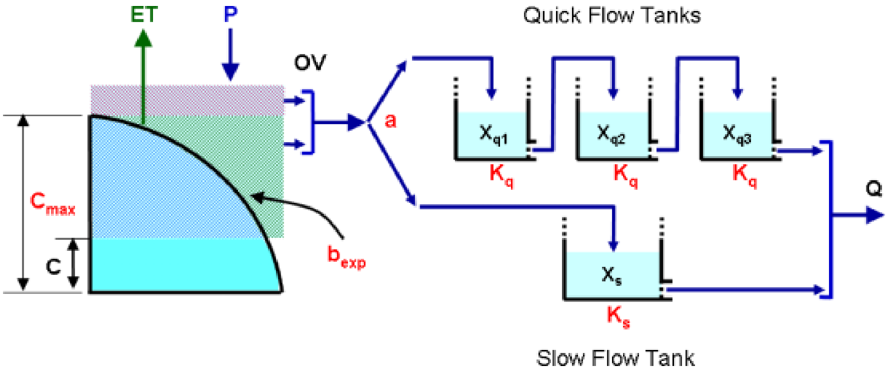

Figure 2. The HyMod model structure

of the catchment having a particular storage capacity C:

$$F(C) = 1 - \left(1 - \frac{C(t)}{C_{\max}}\right)^{b_{\text{exp}}} \tag{1}$$

where C is always smaller than $C_{\max}$. The routing component consists of a series of three linear reservoirs for quick flow and one linear reservoir for slow flow. Table I describes the HyMod parameters and presents reasonable ranges to be used in constraining their calibration (cf. Vrugt *et al.*, 2003b). The additional parameters that are mentioned in this table refer to improved versions of the original HyMod model which will be explained later in this article.

Model calibration

Traditional and dynamic calibration approaches. We calibrated all model structures three times, twice using a traditional calibration approach with time-invariant parameters (once using a single criterion and once using multiple criteria) and once using a dynamic calibration approach with time-variant parameters. In the latter procedure, we allow the model parameters to take on different values for each of the 12 different clusters, resulting in a number of degrees of freedom equal to the number of parameters times 12. However, the water balance was respected for the entire calibration period, effectively constraining the parameter variability between clusters. In this way, we attempt to capture reliable information about model functioning in the time-invariant parameters. The dynamic calibration approach results in 12 parameter sets, and may seem a rather brute force approach to calibration that seems to conflict with principles of parsimony. Note, however, that the goal of this calibration approach is to investigate the temporal variability of parameters, rather than to find the best-performing model *per se*.

Table I. HyMod model parameters

Name	Description and unit	Prior range
Parameters		
$C_{\max}$	Maximum soil moisture content (L)	10–1500
b_{exp}	Spatial variability of soil moisture capacity (–)	0·01–1·99
a	Quick/slow flow distribution factor (–)	0·01–0·99
K_s	Recession coefficient slow reservoir (T^{-1})	0·01–0·99
K_q	Recession coefficient quick reservoir(s) (T^{-1})	0·01–0·99
Additional parameters		
F_{RFC}	Rainfall correction factor (–)	0·5–1·5
b_s	Nonlinearity coefficient slow reservoir (–)	0·5–2
b_q	Nonlinearity coefficient quick reservoir (–)	0·5–2
L_q	Length transformation function quick reservoir (T)	1–10

For all calibration approaches, we minimized the normalized root mean squared error (NRMSE) objective function [Equaion (2)]:

$$\text{NRMSE} = \frac{\sqrt{1/K \sum_{k=1}^{K} \left(\hat{Q}_k - Q_k\right)^2}}{1/K \sum_{k=1}^{K} Q_k} \tag{2}$$

where K is the total number of data elements, and $\hat{Q}_k$ and Q_k are the simulated and the observed discharges at the kth time interval, respectively. The relative performance between different clusters can be compared because the NRMSE is a statistic that is normalized for the average discharge in each cluster.

The two objective functions used in the multi-criteria calibration are the NRMSE and the mean squared error of the log-transformed discharges (logMSE), which express errors on high and low flows, respectively.

Optimization algorithms. Single-criterion parameter estimation was performed using a self-adaptive variant of the differential evolution (DE) algorithm introduced by Storn and Price (1997). While relatively simple, the DE algorithm is powerful, and generally shows high accuracy and fast convergence on many test problems (see Vesterstrøm and Thomsen, 2004). It has been applied to hydrological model calibration by Shoemaker *et al.* (2007).

Several variants of the DE algorithm have been suggested (Storn and Price, 1997). We have selected the commonly used *DE/rand/1/bin* strategy, which can be summarized as follows. A population of N individuals $\mathbf{x}_{i,G}$, $i = 1, 2, \ldots, N$, each of which is a vector of D optimization parameters, is evolved for a number of generations (indicated by G). The evolution is defined as a process of three operations: mutation, crossover and selection. Each individual $\mathbf{x}_{i,G}$ is mutated according to

$$\mathbf{v}_{i,G+1} = \mathbf{x}_{r_1,G} + f \cdot (\mathbf{x}_{r_2,G} - \mathbf{x}_{r_3,G}), \quad r_1 \neq r_2 \neq r_3 \neq i \tag{3}$$

with randomly chosen indices r_1, r_2, $r_3 \in [1, N]$. $f \in [0, 2]$ controls the amplification of the difference vector $\left(\mathbf{x}_{r_2,G} - \mathbf{x}_{r_3,G}\right)$ and is one of the two main control parameters of the algorithm. If any component of a mutant vector falls outside the acceptable parameter bounds, it is set to the bound value. Crossover is performed using the individuals and their mutants according to

$$\mathbf{u}_{i,G+1} = \left(u_{1i,G+1}, u_{2i,G+1}, \ldots, u_{Di,G+1}\right) \tag{4}$$

where

$$u_{ji,G+1} = \begin{matrix} v_{ji,G+1}, & \text{if } r(j) \leq c \text{ or } j = r_n(i) \\ u_{ji,G}, & \text{if } r(j) > c \text{ and } j \neq r_n(i) \end{matrix} \tag{5}$$

for $j = 1, 2, \ldots, D$. $r(j) \in [0, 1]$ is the jth output of a uniform random number generator. $c \in [0, 1]$ is the crossover constant and is the second main control parameter of the DE algorithm. $r_n(i) \in (1, 2, \ldots, D)$ is a randomly chosen index which ensures that at least one element of $\mathbf{u}_{i,G+1}$ comes from $\mathbf{v}_{i,G+1}$. Selection is performed according to a greedy selection scheme:

$$\mathbf{x}_{i,G+1} = \begin{cases} \mathbf{u}_{i,G+1}, & \text{if } f\ \left(\mathbf{u}_{i,G+1}\right) \text{ is better than } f\ \left(\mathbf{x}_{i,G}\right) \\ \mathbf{x}_{i,G}, & \text{otherwise} \end{cases} \tag{6}$$

for $j = 1, 2, \ldots, D$. This way, the old individual is replaced only if the objective function value of the new individual is better.

A crucial issue for the efficiency and efficacy of the DE algorithm is the choice of values for its control parameters f and c (Liu and Lampinen, 2002; Brest *et al.*, 2006). Here, we follow the approach suggested by Brest *et al.* (2006) in which the control parameters are evolved by placing them inside the vector associated with each individual. In that same article, this self-adaptive version was shown to successfully find optimal control parameters for different problems and consequently outperform other implementations of DE.

The population size for the traditional calibration procedure was set to ten times the dimension of the optimization problem (i.e. the numbers of parameters to be optimized). The number of generations was limited to 250. For the dynamic calibration, we have opted for a population size of five times the dimension of the problem (i.e. the numbers of parameters to be optimized times the number of clusters) and 1000 generations. To initialize the second calibration procedure, we set the initial parameter values close to the optimum found during the traditional calibration, thereby speeding convergence.

Multi-criteria parameter estimation was done with the MOSCEM-UA algorithm (Vrugt *et al.*, 2003b). The settings of the algorithm that were used are as follows: the number of complexes equal to the number of parameters to be calibrated, 100 random samples per complex and total number of draws equal to 2000 times the number of complexes.

RESULTS AND ANALYSIS

Cluster analysis

Figure 3 shows the three-dimensional data separated into the 12 clusters found by the k-means algorithm. The clusters have been numbered according to their rank in terms of magnitude of soil moisture for convenience in presentation of subsequent results. The figure shows that an increase of P_{ma} (moving average of precipitation) generally results in an increase of S (soil moisture), as is to be expected. Although there is likely to be some overlap in the information content of these variables, the clarity (minimal overlap) of the clusters in this plot indicates that the clustering algorithm has selected these two dimensions as major variables for distinguishing between clusters. The clustering algorithm also appears to be using P (precipitation) actively as an indicator for distinguishing peak flow events (cluster 9, e.g. is poorly discernable in the S vs P_{ma} subplot but very pronounced in the two other subplots).

Figure 4 shows the identified temporal clusters plotted along with the hydrograph for the evaluation period. Given that the hydrograph information was not in the clustering procedure, and the fact that these are evaluation period results, the close concurrence of the patterns indicates that the temporal clustering has been quite successful. The periods of wetting and drying and the related runoff dynamics of the catchment are clearly reflected in the organization of the clusters on both the events and the seasonal scales.

Model diagnosis

In this section, we show how diagnostic information inferred from model calibration and evaluation is used in three subsequent model structural improvement stages.

Stage I. The HyMod model calibration and evaluation performance statistics for the traditional and dynamic calibration approaches are presented in Table II and Figure 5. Table II shows the NRMSE computed for the entire period, and the bars in Figure 5 shows the NRMSE performance for the portion of the evaluation data in each cluster. In interpreting these results, note that the larger cluster numbers indicate larger amounts of soil moisture storage (catchment wetness; Figure 3). Despite having more degrees of freedom, and producing smaller overall NRMSE (better overall model performance), the dynamic calibration has resulted in worse evaluation period performance on clusters 2–8 which correspond to drier periods. It appears that the dynamic calibration

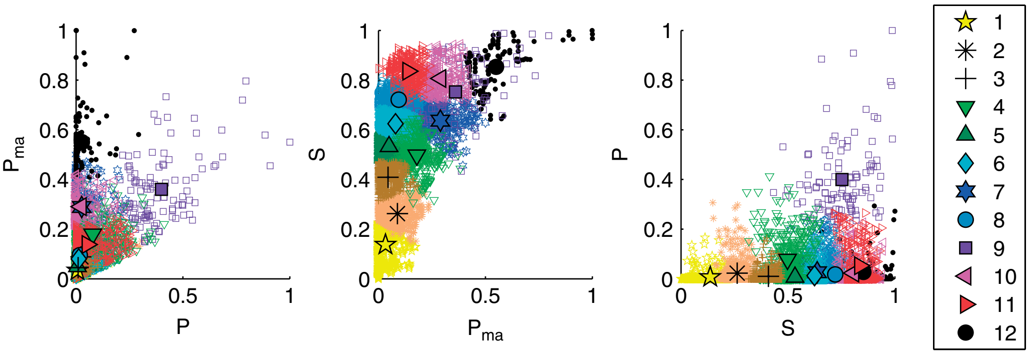

Figure 3. Two-dimensional projections of clustering results on three-dimensional dataset. Data are normalized between 0 and 1. The cluster centroids are indicated with large markers. The cluster numbers are sorted according to their rank in soil moisture for convenience

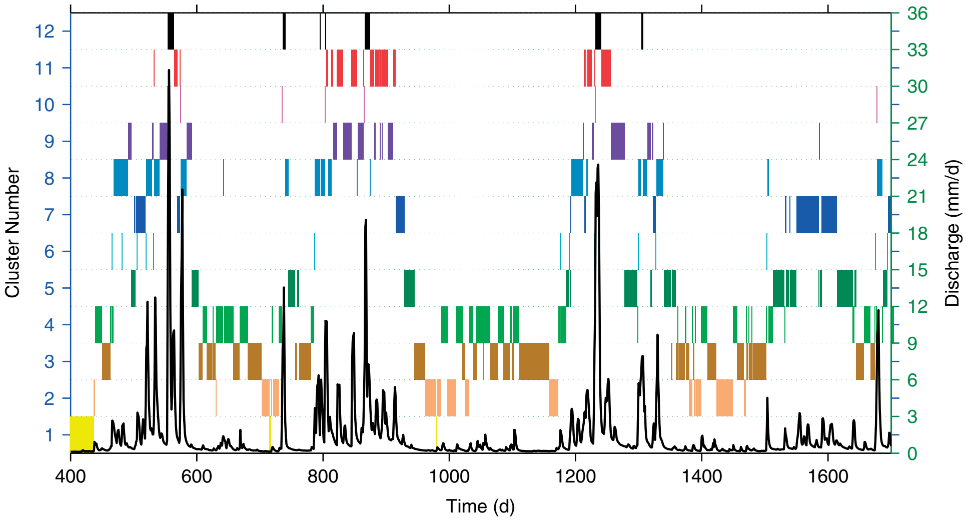

Figure 4. Representative detail of the clustering results from the evaluation period, showing the clusters in time, along with the observed hydrograph

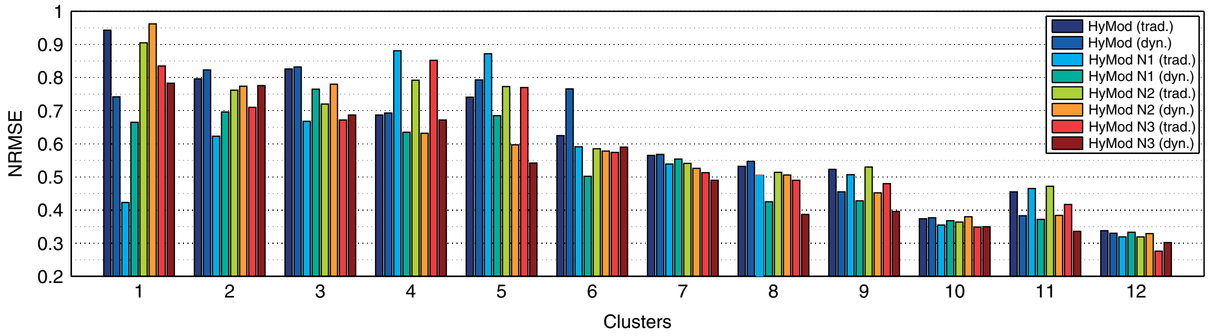

Figure 5. Performance of the various model structures per cluster on the evaluation data, for both the traditional and the dynamic calibration

has found a solution that provides better predictions for the high flows than the low flows. In fact, a detailed examination of the hydrographs for each cluster (not shown here to save space) shows severe underestimation during the dry periods. Moreover, the total discharge volume of the calibrated model was about 10% less than the observed volume. The above indicates a shortage of water, especially in the slow flow reservoir, and we therefore assume that the model receives too little net inflow by meteorological forcing. It appears that, in trying to minimize the overall period performance statistic for the current model structure, the calibration process consequently has settled on a trade off which emphasizes the fit of high flows over low flows.

To try and improve model performance, therefore, we first added a rainfall correction factor (an additional parameter F_{RFC}; Table I) to the model to compensate for measurement errors through multiplication with the rainfall [Equation (7)]; the enhanced version of the model is given the name 'HyMod-N1'. Such a parameter is not uncommon in hydrological models and is used in, for example, the HBV–96 model (Lindström *et al.*, 1997). Recalibration using the traditional calibration method resulted in an optimal value of $P_{\mathrm{corr}} = 1{\cdot}057$, indicating

Table II. Performance comparison

	Calibration NRMSE		Evaluation NRMSE	
	Traditional calibration	Dynamic calibration	Traditional calibration	Dynamic calibration
HyMod	0·836	0·702	0·793	0·773
HyMod-N1	0·800	0·668	0·763	0·749
HyMod-N2	0·806	0·657	0·773	0·723
HyMod-N3	0·737	0·616	0·709	0·694

Table III. Optimal parameter values found using traditional calibration for all HyMod model variants

Parameter	HyMod	HyMod-N1	HyMod-N2	HyMod-N3
$C_{\max}$	210	241	263	272
$b_{\exp}$	0·400	0·425	0·504	0·554
A	0·296	0·270	0·288	0·588
K_s	0·319	0·295	0·253	0·417
K_q	0·828	0·830	0·650	0·191
F_{RFC}	—	1·057	1·084	1·084
β_s	—	—	1·059	0·979
β_q	—	—	1·092	1·387
L_q	—	—	—	3·00

that the rainfall data likely underestimate the actual rainfall. Optimal values (using traditional calibration) for all parameters of the original HyMod model and the enhanced HyMod-N1 model are presented in Table III.

$$P_{\mathrm{corr}} = F_{\mathrm{RFC}} \cdot P \tag{7}$$

The five original parameters of the HyMod-N1 model were subsequently dynamically recalibrated in the same way as for the HyMod model, but with the additional F_{RFC} parameter kept constant. The assumption here is that the rainfall underestimation is of a more structural nature, and we want to avoid that the dynamic calibration abuses the potentially influential F_{RFC} parameter to compensate for other errors. The smaller calibration and evaluation period errors achieved by the dynamic calibration, both overall and for each cluster, indicate that the rainfall correction factor has fulfilled its purpose and is now allowing the dynamic character of the calibration process to be better exploited. The multi-criteria results presented in Figure 7 confirm that the trade-off between high and low flow performances has decreased.

Stage II. In the next stage of our diagnostic approach, we make use of the information about the functioning of the model structure that is implicitly contained in the variability of the dynamic HyMod-N1 parameters over the 12 clusters. The subplots in Figure 6 show the optimal values for each parameter for each temporal cluster, found using the dynamic calibration procedure; the horizontal line represents the time-invariant parameter value found using traditional calibration. Of course, many patterns might be hidden in these results, and useful information could be difficult to detect due to complex parameter interactions. Here, we begin with the simple test of hypothesis that useful information can be extracted from each individual dynamic parameter set and from their most obvious coincident patterns of variation. Two patterns stand out clearly in Figure 6: the general tendency for the K_s and K_q recession coefficients to be smaller than their 'normal' (time-invariant) value, and the tendency for both these parameters to increase with increasing wetness.

These patterns suggest that the slow and the quick flow reservoirs may be functioning sub-optimally, and might better be represented using nonlinear recession rate dynamics. We have tested this hypothesis by using a nonlinear relationship between reservoir storage (X) and outflow (Q) for both the slow and the quick reservoirs (Equation (8)). Two nonlinearity coefficients (β_s and β_q) were introduced to this end (Table I).

$$Q = K \cdot S^{\beta} \tag{8}$$

When the HyMod-N2 model is recalibrated using the *traditional* (time-invariant parameter) calibration procedure we get the overall and individual cluster performance results shown in Table II and Figure 5. The optimal (time-invariant) parameter values are again shown in Table III and Figure 6 for easy comparison with the values obtained for the previous model structures. Compared to the traditionally calibrated HyMod-N1 model, the HyMod-N2 model shows a slight deterioration in terms of overall NRMSE performance. However, the performance of some dry clusters seems to improve. The multi-criteria comparison shown in Figure 7 explains these observations; the HyMod-N2 is comparable in terms of high-flow performance but is clearly better than HyMod-N1 when it comes to low flows. This suggests that the results in Table II and Figure 5 are indeed not fully representative of model performance because of the use of a single objective function which focuses on high flows. Finally, the smaller spread of the solutions of the HyMod-N2 model compared to the HyMod-N1 and HyMod models indicates that this model has a more reliable model structure.

The HyMod-N2 model was subsequently subjected to dynamic calibration. Note that the F_{RFC} parameter was again kept constant to avoid the dynamic calibration procedure abusing this powerful degree of freedom. The results, shown in Figure 6, indicate that the values of the HyMod-N2 parameters, especially K_s, seem to have become more consistently closer to the optimal values as indicated by traditional calibration. This suggests that nonlinearity of the slow flow reservoir is indeed a reasonable enhancement of the model. However, the values of K_q still show a clear difference between the dry and the wet clusters.

Stage III. Further exploring our main hypothesis that diagnostic model information can be extracted from the HyMod-N2 parameter values after dynamic calibration, we subsequently focus on the large differences between

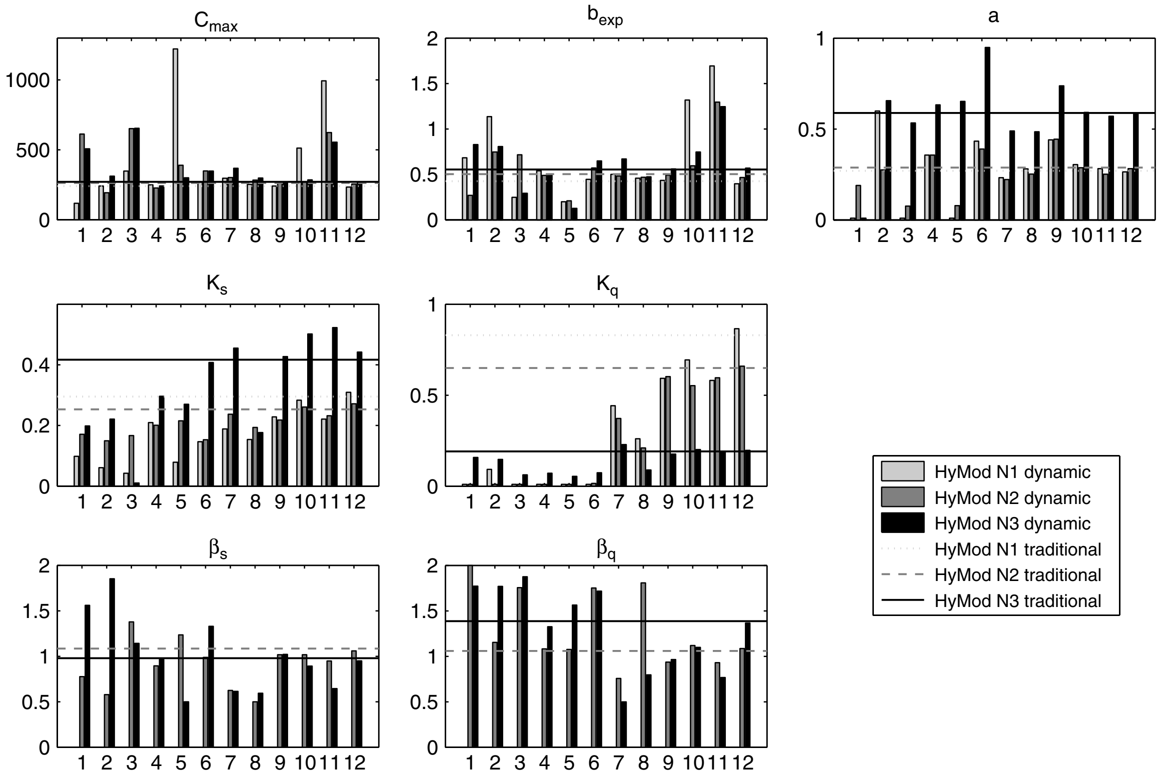

Figure 6. Parameter estimation results of the dynamic calibration of the various model structures. The horizontal lines show the optimal values found by the traditional single-criterion calibration procedures

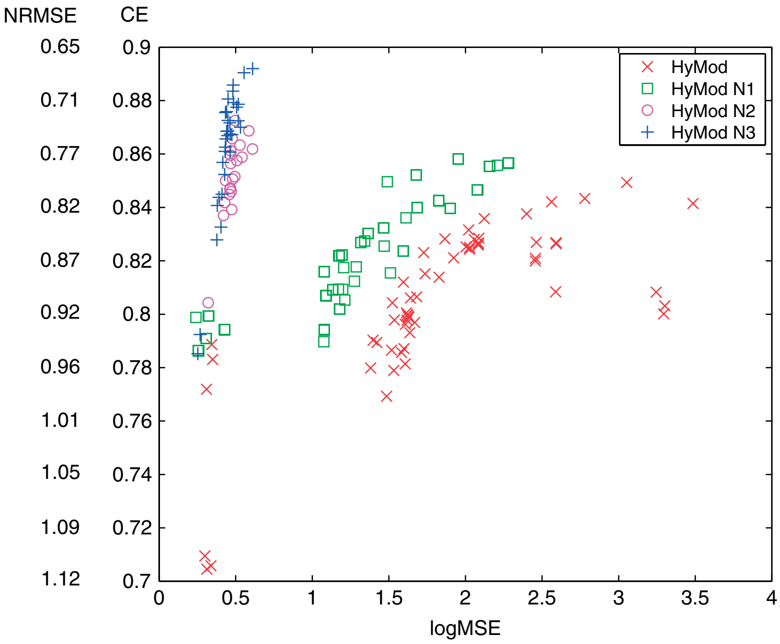

Figure 7. Multi-criteria evaluation results for the various model structures after calibration with the MOSCEM-UA algorithm on the NRMSE and logMSE objective functions. The accompanying values of the Nash–Sutcliffe coefficient of efficiency are also included on the *y*-axis

dry- and wet-period K_q values. The three-level cascade of quick flow reservoirs of the HyMod model is supposed to capture the faster component of the storage–discharge relationship of the catchment such as rapid (sub)surface flows. Additionally, the response function that this cascade model produces should account for runoff routing effects. As in the dynamic calibration the quick flow model component attempts to reduce outflow in the dry clusters to facilitate simulations of high flows in wet clusters, we assume the response function is not optimal. A triangular transformation function was therefore incorporated which should enable a more accurate response function, especially one that simulates runoff routing effects better. This is because such a function provides more flexibility in timing of reservoir inflow and enables a more gradual input to the reservoirs (a moderating of the instantaneous rainfall impulse shocks). A similar modification was proposed by Fenicia *et al.* (2008) to improve the structure of a conceptual rainfall–runoff model. By using the transformation function in combination with a single nonlinear reservoir, we still allow for a nonlinear storage–discharge relationship. Additionally, the transformation function plus a single reservoir is a more parsimonious solution in terms of states than the three-level cascade of reservoirs.

Implementation of this concept into HyMod-N2 results in the updated model structure HyMod-N3. An additional parameter, L_q, was introduced which represents the base length of the triangular function (Table I). The triangular transformation function is defined such that the reservoir inflow Y is transformed into $\hat{Y}$ as follows:

$$\hat{Y}_{t0} = Y_{t0} \cdot \omega_0 + Y_{t-1} \cdot \omega_1 + \cdots + Y_{t-\lceil L_q \rceil} \cdot \omega_{\lceil L_q \rceil} \quad (9)$$

where the coefficients are defined as:

$$\omega_i = \int_{i-1}^{i} \frac{2t}{{L_q}^2} \mathrm{d}t \quad \text{for} \quad i \in 1, 2, \ldots, \lceil L_q \rceil \quad (10)$$

The results in Figure 5 show that the HyMod-N3 outperforms the HyMod-N2 model on both the calibration and the evaluation data for almost all clusters. Interestingly, Table II even shows that the HyMod-N3 (with time-invariant parameters) performs even better than the dynamically calibrated HyMod-N2 model on the evaluation data, supporting the hypothesis that the new quick flow model component indeed results in a more consistent model of the rainfall–runoff process. The multi-criteria results presented in Figure 7 confirm that the HyMod-N3 model offers improvement over the HyMod-N2 model, especially on the fit of high flows. Interestingly, a, the parameter that controls the distribution of water between the slow and the quick flow reservoirs has increased significantly in value. This increase indicates that the HyMod-N3 model allocates significantly more water to the quick flow reservoir. Another observed change in the HyMod-N3 model that is related to this change in allocation, is that the optimal value for the nonlinearity coefficient of the slow flow reservoir is relatively close to 1 (Table III and Figure 6). This suggests that a linear storage–discharge relationship for slow flow may be appropriate after all.

Finally, the HyMod-N3 model was dynamically calibrated. The parameter values shown in Figure 6 deviate much less from the optimal values from traditional calibration than any of the other models. This reduced sensitivity to the variability of its parameters suggests that HyMod-N3 is a more robust model.

DISCUSSION AND CONCLUSIONS

This article has made a first step towards examining and understanding how to conduct diagnostic model evaluation by extracting temporal signature information via an augmented calibration procedure based on temporal clustering. Such an approach might be used to recursively achieve diagnostic improvements of hydrological models. The results show that consistent patterns of parameter variation do indeed show up through the application of this approach, and that analysis of these patterns can point towards potential model improvements. In this work, we have explored three iterations resulting in updates to the model structure, based on quite obvious patterns. More powerful alternatives to this simple diagnostic approach are likely to exist. Nevertheless, significant improvements were achieved on an already well-performing model. The research presented here illustrates one pathway to the development of a diagnostic approach to model evaluation. The study example shows how information from the data and the functioning of the hydrological model can be combined in a useful manner to achieve improvements to the working model hypothesis.

What can be concluded from this work is that, although process knowledge and perceptual models of reality will probably remain the most important source of information for model development and improvement, there is still much that can be gained through careful scrutiny of data and model functioning. The issue to be better understood is how such a diagnostic evaluation should be conducted—what are the strategies that will generally lead to uncovering information that can be reliably and readily used for model improvement. For this, a variety of model diagnostic approaches will need to be extensively tested and verified. Further research on the difficult step in which the calibration information is translated into model structural improvements could be especially beneficial. More advanced data analysis or pattern recognition techniques could prove to be particularly useful in this.

Other conclusions and recommendations based on this work are:

- The k-means clustering algorithm was shown to be effective in identifying hydrologically similar periods. Future research might benefit from the use of more sophisticated clustering techniques such as fuzzy clustering or random forests (Breiman, 2001). The rather subjective choice in the clustering procedure need to be further investigated to find appropriate settings for hydrological applications.
- The self-adaptive DE optimization algorithm was found to be effective in hydrologic model calibration. The algorithm obtained good solutions for both the traditional calibration and on the more complex (higher dimensional) dynamic calibration.
- Although the self-adaptive DE algorithm is powerful, it does not provide estimates of the parameter uncertainty of its results. A method that helps to assess the uncertainty in the parameter estimates would significantly benefit the diagnostic model evaluation.

ACKNOWLEDGEMENTS

The authors express their gratitude to Koray Yilmaz, Steven Weijs and the two anonymous reviewers for helpful comments and suggestions. The self-adaptive DE optimization algorithm code was adapted from the DE code on the website of Rainer Storn (http://www.icsi.berkeley.edu/∼storn/code.html) and the self-adaptive DE code provided to us by Janez Brest. Collaboration between the authors was made possible by a travel grant from The Netherlands Organization for Scientific Research (NWO).

REFERENCES

Arthur D, Vassilvitskii S. 2007. K-means++: the advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New Orleans, LA, USA; 1027–1035.

Bensaid AM, Hall LO, Bezdek JC, Clarke LP, Silbiger ML, Arrington JA, Murtagh RF. 1996. Validity-guided (re)clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems* **4**: 112–123.

Bergström S. 1976. *Development and application of a conceptual runoff model for Scandinavian catchments*. SMHI RH07, Norrköping.

Boyle DP. 2000. *Multicriteria calibration of hydrological models*. PhD dissertation, Department of Hydrology and Water Resources, University of Arizona: Tucson, USA.

Boyle DP, Gupta HV, Sorooshian S. 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic models. *Water Resources Research* **36**(12): 3663–3674.

Breiman L. 2001. Random forests. *Machine Learning* **45**(1): 5–32.

Brest J, Greiner S, Bošković B, Mernik M, Žumer V. 2006. Self-adapting control parameters in differential evolution: a comparative study on numerical benchmark problems. *IEEE Transactions on Evolutionary Computation* **10**(6): 646–657.

Burn DH. 1989. Cluster analysis as applied to regional flood frequency. *Journal of Water Resources Planning and Management* **115**(5): 567–582.

Burnash RJC. 1995. The NWS river forecast system—catchment modeling. In *Computer Models of Watershed Hydrology*, Singh VP (ed). Water Resources Publications: Colorado; 311–366.

Choi HT, Beven KJ. 2007. Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework. *Journal of Hydrology* **332**: 316–336.

Clark MP, Slater AG, Rupp DE, Woods RA, Vrugt JA, Gupta HV, Wagener T, Hay LE. 2008. Framework for Understanding Structural Errors (FUSE): a modular framework to diagnose differences between hydrological models. *Water Resources Research* **44**: W00B02. DOI: 10.1029/2007WR006735.

de Vos NJ, Rientjes THM. 2005. Constraints of artificial neural networks for rainfall–runoff modeling: trade-offs in hydrological state representation and model evaluation. *Hydrology and Earth System Sciences* **9**: 111–126.

de Vos NJ, Rientjes THM. 2007. Multi-objective performance comparison of an artificial neural network and a conceptual rainfall–runoff model. *Hydrological Sciences Journal* **52**(3): 397–413.

de Vos NJ, Rientjes THM. 2008. Multi-objective training of artificial neural networks for rainfall–runoff modeling. *Water Resources Research* **44**: W08434.

Edijatno N, Nascimento O, Yang X, Makhlouf Z, Michel C. 1999. GR3J: a daily watershed model with three free parameters. *Hydrological Sciences Journal* **44**(2): 263–277.

Everitt B, Landau S, Leese M. 2001. *Cluster Analysis*. Oxford University Press: NY, USA.

Fenicia F, Savenije HHG, Matgen P, Pfister L. 2008. Understanding catchment behavior through stepwise model concept improvement. *Water Resource Research* **44**: W01402.

Fenicia F, Solomatine DP, Savenije HHG, Matgen P. 2007. Soft combination of local models in a multi-objective framework. *Hydrology and Earth System Sciences* **11**: 1797–1809.

Frapporti G, Vriend P, van Gaans PFM. 1993. Hydrogeochemistry of the shallow Dutch groundwater: interpretation of the national groundwater quality monitoring network. *Water Resource Research* **29**(9): 2993–3004.

Gupta HV, Sorooshian S, Yapo PO. 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research* **34**(4): 751–763.

Gupta HV, Wagener T, Liu Y. 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes* **22**: 3802–3813.

Hall MJ, Minns AW. 1999. The classification of hydrologically homogeneous regions. *Hydrological Sciences Journal* **44**: 693–704.

Herbst M, Gupta HV, Casper MC. 2009. Mapping model behaviour using self-organizing maps. *Hydrology and Earth System Sciences* **13**: 395–409.

Hsu K, Gupta HV, Gao X, Sorooshian S, Imam B. 2002. Self-organizing linear output map (SOLO): an artificial neural network suitable for hydrologic modeling and analysis. *Water Resources Research* **38**(12): 1302.

Jain AK, Murty MN, Flynn PJ. 1999. Data clustering: a review. *ACM Computing Surveys (CSUR)* **31**(3): 264–323.

Kashif Gill M, Kaheil YH, Khalil A, McKee M, Bastidas L. 2006. Multiobjective particle swarm optimization for parameter estimation in hydrology. *Water Resources Research* **42**: W07417.

Khu ST, Madsen H. 2005. Multiobjective calibration with Pareto preference ordering: an application to rainfall–runoff model calibration. *Water Resources Research* **41**(3): W03004.

Lauzon N, Anctil F, Baxter CW. 2006. Clustering of heterogeneous precipitation fields for the assessment and possible improvement of lumped neural network models for streamflow forecasts. *Hydrology and Earth System Sciences* **10**: 485–494.

Lindström G, Johansson B, Persson M, Gardelin M, Bergström S. 1997. Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology* **201**: 272–288.

Liong SY, Lim WH, Kojiri T, Hori T. 2000. Advance flood forecasting for flood stricken Bangladesh with a fuzzy reasoning method. *Hydrological Processes* **14**(3): 431–448.

Liu J, Lampinen J. 2002. On setting the control parameter of the differential evolution method. *Proceedings of the 8th International Conference on Soft Computing (MENDEL 2002)*: Brno, Czech Republic; 11–18.

Lloyd SP. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**(2): 129–136.

Marshall L, Nott D, Sharma A. 2007. Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework. *Hydrological Processes* **21**: 847–861.

Moore RJ. 1985. The probability-distributed principle and runoff production at point and basin scales. *Hydrological Sciences Journal* **30**(2): 273–297.

Oudin L, Andréassian V, Mathevet T, Perrin C, Michel C. 2006. Dynamic averaging of rainfall–runoff model simulations from complementary model parameterizations. *Water Resources Research* **42**: W07410.

Perrin C, Michel C, Andréassian V. 2003. Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology* **279**: 275–289.

Reusser DE, Blume T, Schaefli B, Zehe E. Analysing the temporal dynamics of model performance for hydrological models. *Hydrology and Earth System Sciences* **13**: 999–1018.

Shoemaker CA, Regis RG, Fleming RC. 2007. Watershed calibration using multistart local optimization and evolutionary optimization with radial basis function approximation. *Hydrological Sciences Journal* **52**(3): 450–465.

Srinivas VV, Tripathi S, Rao AR, Govindaraju RS. 2008. Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. *Journal of Hydrology* **348**: 148–166.

Storn R, Price K. 1997. Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* **11**: 341–359.

Tang Y, Reed P, Wagener T. 2006. How effective and efficient are multiobjective evolutionary algorithms at hydrological model calibration?. *Hydrology and Earth System Sciences* **10**: 289–307.

Toth E. 2009. Classification of hydro-meteorological conditions and multiple artificial neural networks for streamflow forecasting. *Hydrology and Earth System Sciences* **13**: 1555–1566.

Vesterstrøm J, Thomsen R, 2004. A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. *Proceedings of the IEEE Congress on Evolutionary Computation*, Portland, OR; 1980–1987.

Vrugt JA, Gupta HV, Bastidas LA, Bouten W, Sorooshian S. 2003a. Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research* **39**(8): 1214.

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

Wagener T, Boyle DP, Lees MJ, Wheater HS, Gupta HV, Sorooshian S. 2001. A framework for development and application of hydrological models. *Hydrology and Earth System Sciences* **5**(1): 13–26.

Wagener T, McIntyre N, Lees MJ, Wheater HS, Gupta HV. 2003. Towards reduced uncertainty in conceptual rainfall-runoff modeling: dynamic identifiability analysis. *Hydrological Processes* **17**: 455–476.

Xie XL, Beni G. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**: 841–847.

Yapo PO, Gupta HV, Sorooshian S. 1998. Multi-objective global optimization for hydrologic models. *Journal of Hydrology* **204**: 83–97.