ELSEVIER

Journal of Hydrology 242 (2001) 275–301

Journal of Hydrology

www.elsevier.com/locate/jhydrol

Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments

C. Perrin*, C. Michel, V. Andréassian

Water Quality and Hydrology Research Unit, Cemagref, Parc de Tourvoie BP 44, 92163 Antony Cedex, France

Received 21 February 2000; revised 10 July 2000; accepted 30 October 2000

Abstract

Hydrological models must be reliable and robust as these qualities influence all applications based on model output. Previous studies on conceptual rainfall–runoff models have shown that one of the root causes of their output uncertainty is model over-parameterisation. The problem of poorly defined parameters has attracted much attention but has not yet been satisfactorily solved. We believe that the most fruitful way forward is to improve the structures where these parameters act. The main objective of this paper is to examine the role of complexity in hydrological models by studying the relation between the number of optimised parameters and model performance. An extensive comparative performance assessment of the structures of 19 daily lumped models was carried out on 429 catchments, mostly in France but also in the United States, Australia, the Ivory Coast and Brazil. Bulk treatment of the data showed that the complex models outperform the simple ones in calibration mode but not in verification mode. We argue that the main reason why complex models lack stability is that the structure, i.e. the way components are organised, is not suited to extracting information available in hydrological time-series. An inadequate complexity typically results in model over-parameterisation and parameter uncertainty. Although complexity has been used as a response to the challenge of predicting the hydrological effects of environmental changes, this study suggests that such models may have been developed with excessive confidence and that they could face difficulties of parameter estimation and structure validation when confronted with hydro-meteorological time-series. This comparative study indicates that some parsimonious models can yield promising results and should be further developed, although they are not able to tackle all types of problems, which would be the case if their complexity were ideally adapted. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Hydrology; Models; Catchments; Comparison; Complexity; Parsimony

1. Introduction

After more than 30 years of 'classical' hydrological model development, there is an increasing trend in hydrology today to use new tools, that provide distributed information of catchment characteristics. It is tempting to introduce newly available information into increasingly complex catchment models, when one is faced with the task of accurately representing the inherent complexity of real systems. However, although this approach might be useful in terms of knowledge of the processes, it has limitations when applied in an operational context. Conversely, simple catchment models that lump catchment heterogeneities and represent the transformation of precipitation into streamflow, conceptually or empirically,

* Corresponding author. Fax: +33-1-40-96-61-99.
E-mail address: charles.perrin@cemagref.fr (C. Perrin).

0022-1694/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0022-1694(00)00393-0

are generally easy-to-use tools with low data requirements. In spite of the crude approximation resulting from their lumped nature and simple structure, such models have proved efficient in many case studies reported in the literature and they are undoubtedly useful for engineers and water managers. Daily conceptual models (of interest here) have successfully fulfilled most operational requirements, such as flood frequency assessment (Cameron et al., 1999; Uhlenbrook et al., 1999), reservoir management (Yang et al., 1995), flood and drought forecasting (see, e.g. Yang and Michel, 2000). Although these models have, up to now, been unable to predict the change in streamflow caused by land-use changes, they are apt at objectively detecting such changes (see, e.g. Lavabre et al., 1993; Lørup et al., 1998).

Since the early 1960s, many hydrologists have concentrated their efforts on designing rainfall–runoff models. Because they have attracted such widespread interest and so many different ones are being developed, the need for comparative studies was expressed quite early on (WMO, 1975) to evaluate the ability of models to simulate streamflow and to provide guidelines for end-users. Model assessment however is a tricky exercise and the conclusions of such experiments generally depend on the methodology of the comparisons and the characteristics of the test catchments. This is all the more true as models are mostly tested on limited numbers of catchments. Due to increasing computational capacity, it is possible today to extensively test simple models against great many catchments, under a wide range of climate conditions. However it would be naïve to believe that, one might, from a broad-based comparison, identify one single outstanding model according to any assessment criterion that would satisfy all water stakeholders (Leviandier, 1988). Nevertheless, we believe that the robustness and reliability of a model resides primarily in its ability to perform under as varied a set of hydrological conditions as possible. From this standpoint, comparative assessments serve to highlight strengths and weaknesses of modelling approaches of various complexity.

2. Existing comparative assessments

Comparative assessment of the performances of rainfall–runoff models is not a new issue but is rarely the focus of much research, whereas there is a plethora of studies reporting satisfactory results from a single model. Reviews of most comparison exercises carried out so far can be found in Michaud and Sorooshian (1994) or in Refsgaard and Knudsen (1996). The latter state that, in general, no firm conclusion can be drawn regarding differences in model performances. The conclusions of comparisons may be different from one study to another and depend on the objectives, the methodology, the type of model, the test catchments, the optimisation procedure as well as on the criteria used to assess the performances. Still fewer comparisons have focused specifically on continuous conceptual rainfall–runoff models which are the main concern here. Some examples are: studies by WMO (1975, 1986, 1992), Moore and Mein (1975), Kite (1978), Weeks and Hebbert (1980) and Franchini and Pacciani (1991) extended by Franchini et al. (1996), Chiew et al. (1993), Zhang and Lindström (1996), Ye et al. (1997), Gan et al. (1997) and Perrin and Littlewood (2000). The methodologies and results of these studies merit a few general comments.

All the comparisons have involved the application of models to limited sets of catchments (generally less than 10). Some of them used data sets with highly varied hydro-climatic characteristics (WMO, 1975; Chiew et al., 1993), whereas others focused on specific conditions, e.g. dry conditions in Gan et al. (1997) or Ye et al. (1997). The limited data sets generally make the conclusions very dependent on the hydro-climatic conditions, whereas we believe that a model is all the more reliable if it performs well under highly varied conditions.

The comparisons sometimes concern entire modelling approaches where each model designer can choose a specific objective function and/or calibration procedure to run the model (WMO, 1975; WMO, 1986; Perrin and Littlewood, 2000). We believe — and this is one of the cornerstones of our approach — that the quality of a rainfall–runoff modelling methodology resides essentially and primarily in the model structure, i.e. in the core of the link between rainfall and streamflow. Its robustness, reliability or versatility are of prime importance for the quality of the tools derived from it.

Lastly, it is usually difficult to interpret the reasons for comparatively good or bad performances by one particular model and the proposed reasons are

generally assumptions that have not been really demonstrated. Moreover, the results obtained with various models and reported in comparative studies often show differences that are not significant enough to be consistently interpreted as evidence of quality.

3. Objective

The main objective of this study is to test the performances of several structures derived from well-known rainfall–runoff models on a large sample of catchments within a common framework. Our study was conducted in four steps:

1. Collection of a large sample of data from a wide variety of catchments under different climate conditions.
2. Selection of a variety of existing continuous lumped conceptual or empirical rainfall–runoff models working at a daily time-step. Simple versions with a limited number of free parameters have been devised and recoded to obtain structures with, at most, nine free parameters.
3. Implementation of an automatic testing scheme where all model structures can be similarly tested on the whole data sample both in calibration and simulation modes.
4. Choice of assessment criteria to judge the quality of model structures and to analyse results.

This extensive testing of several model structures over many catchments provides a large body of results that can only be analysed statistically. It provides opportunities to both investigate the link between the complexity and the robustness or reliability of models, and to establish possible complementarity between structures.

Restricting the study to a smaller number of catchments, e.g. 20 or 30, would have meant becoming overly dependent on an arbitrary selection of catchments. The main value of this research lies in including as many catchments as were available when the work began. A more in-depth analysis of a few catchments might not have provided a better insight into the relative merits of the tested models. We strongly believe that a really proficient model should go beyond catchments peculiarities.

4. Catchments and data

Except for the studies by Vandewiele et al. (1992) and Xu and Vandewiele (1995), who assessed the performances of monthly water-balance models on, respectively, 79 and 91 basins in Belgium, China and Burma, previous comparisons generally applied the models to a small number of catchments (less than 10). Although computing limitations may have prevented extensive testing in the past, it is today easier to implement testing schemes that can accommodate a large amount of calculations. Here, models are tested on a wide variety of catchments, some of which have already been used in previous work on rainfall–runoff modelling.

4.1. Data collection

One of the main characteristics of the test sample is that daily data were gathered from a large number of catchments. The data requirements of the tested models are low: as inputs they generally only need rainfall and potential evapotranspiration (PE) series and as output streamflow series (used to calibrate the model). Hydro-meteorological data sets were collected for 429 catchments in Australia, Brazil, France, the Ivory Coast and the United States. They consist of daily time-series of areal rainfall and streamflow. PE data are time-series or long-term average values at a daily or 10-day time-step. Between three and 38 calendar years of concomitant data were available on each catchment.

The 26 Australian catchments are taken from the sample used by Chiew and McMahon (1994) to test MODHYDROLOG and their climate conditions vary from tropical humid to semi-arid. The four Brazilian basins are sub-catchments of the São Francisco River basin in the State of Minas Gerais. They are situated in the upper reaches of the basin upstream from the Três Marias Dam and have a humid climate with a mean annual rainfall of about 1500 mm. The bulk of the catchments (307) are located in France. 140 of these catchments were used by Edijatno et al. (1999) to develop the GR3J model. Although France has a mainly temperate climate, its climate conditions are varied and they are all represented in this sample: Mediterranean conditions in the South of France, oceanic influences in the West and some continental

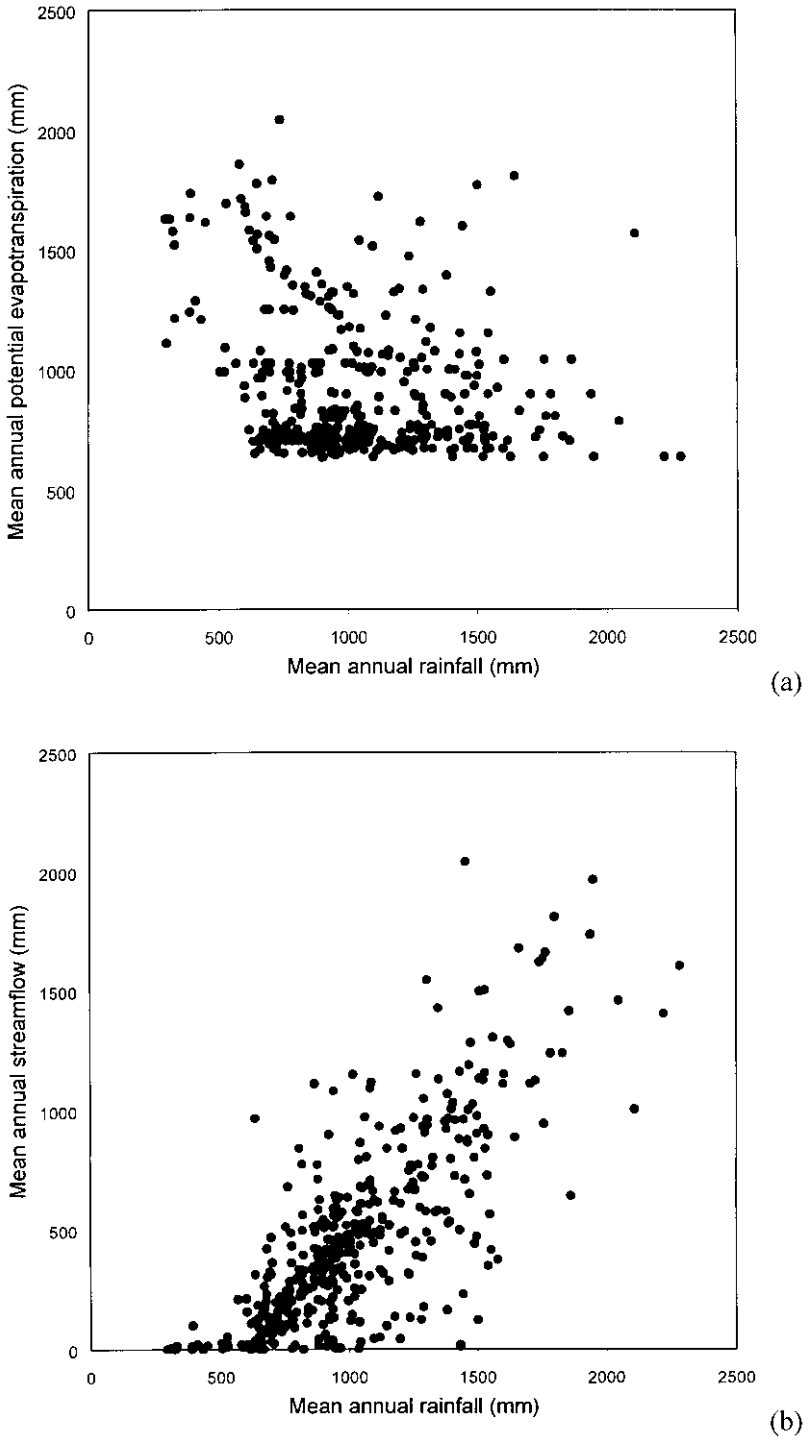


Fig. 1. Plots of: (a) mean annual potential evapotranspiration (mm); and (b) mean annual streamflow against mean annual rainfall (mm).

Table 1
Ranges in size and hydro-climatic characteristics for the sample of 429

Country	Australia	Brazil	France	Ivory coast	United States
Number of catchments	26	4	307	10	82
Catchment area (km^2)	3–2500	11300–50600	1–43800	207–6830	0.1–9890
Mean annual potential evapotrasnpiration (mm)	790–1850	930–1080	630–1250	1330–1770	680–2040
Mean annual precipitation (mm)	300–2100	1450–1580	570–2300	1060–1630	300–1540
Mean annual streamflow (mm)	2–1460	370–590	26–2040	32–460	0.2–840
Catchment yield (%)	0.3–71.3	23.1–38.1	3.7–153	2.6–26.5	0.1–72.6
BFI (%)	0.1–82.3	62.3–79.1	5–98.5	34–74.9	0.1–81.2
Irregularity coefficient of rainfall (%)	158–497	284–290	144–329	235–313	146–447
Irregularity coefficient of streamflow (%)	222–1170	208–260	39–673	173–504	147–1110

features in the eastern part of the country. Three of the 10 catchments in the Ivory Coast are situated in the dry northern part of the country (savannah conditions). Their data have previously been used by Servat and Dezetter (1991, 1992) for rainfall–runoff modelling purposes. The other catchments are located in the centre or the south of the Ivory Coast, with a wetter climate and a denser forest cover. Data from 45 American basins (mainly small agricultural or experimental watersheds) were provided by the Agricultural Research Service (ARS) Water Database (Thurman and Roberts, 1995). They are situated in the states of Arizona, Georgia, Idaho, Iowa, New Mexico, Mississippi, Missouri, Ohio, Oklahoma, Pennsylvania, Texas and Vermont. Data sets for 37 other American catchments from the Model Parameter Estimation Experiment (MOPEX) database were also used. These 37 catchments are located in the states of Arkansas, Kansas, Missouri, New Mexico, Oklahoma and Texas.

The collected precipitation and streamflow data were not re-checked by the authors. The calculation of areal rainfall was dependent on the number and proximity of raingauges. Collected or derived PE data (daily or ten-day values) are of different types: Morton (1983) estimation for the Australian data, pan evaporation in Brazil, Penman (1948) values in France and the Ivory Coast, Hargreaves and Samani (1982) calculations for ARS data and data derived from the atlas by Farnsworth et al. (1982) for American MOPEX basins. In the case of American, Brazilian and French basins, PE data are long-term averages whereas in Australia and the Ivory Coast, data differ from one year to another. However, the importance of high quality is not as crucial in the case of PE data as in rainfall data since models are generally less sensitive to this input variable (see, e.g. Paturel et al., 1995).

Because there are no agreed-upon criteria for data discrimination, there was no preliminary rejection of data sets. In some cases, slight snowmelt, groundwater influence, karstic phenomena or human influences (such as water pumping or diversions, streamflow regulation, land use changes…) are likely to exist. But any selection is inevitably based on some modelling prejudice, which can bias the comparison process. No preliminary data processing such as streamflow re-naturalisation was used. Since we were dealing with a comparative exercise, it was thought necessary to apply the models to all available data, without a priori judgement of hypothetical data quality. If some data are actually of poor quality, all models will suffer equally from this shortcoming.

4.2. Catchment characteristics

The previously cited references provide adequate descriptions of some of the basins used in this study. Given the sample size, we can only present a rough overview of the characteristics of the whole catchment sample. Fig. 1(a) and (b) shows mean annual characteristics (PE, rainfall and streamflow) plotted for all 429 basins. Table 1 summarises the ranges in size and hydro-climatic characteristics of the catchments in each country. An ‘irregularity’ coefficient applied to rainfall (or streamflow) quantifies the

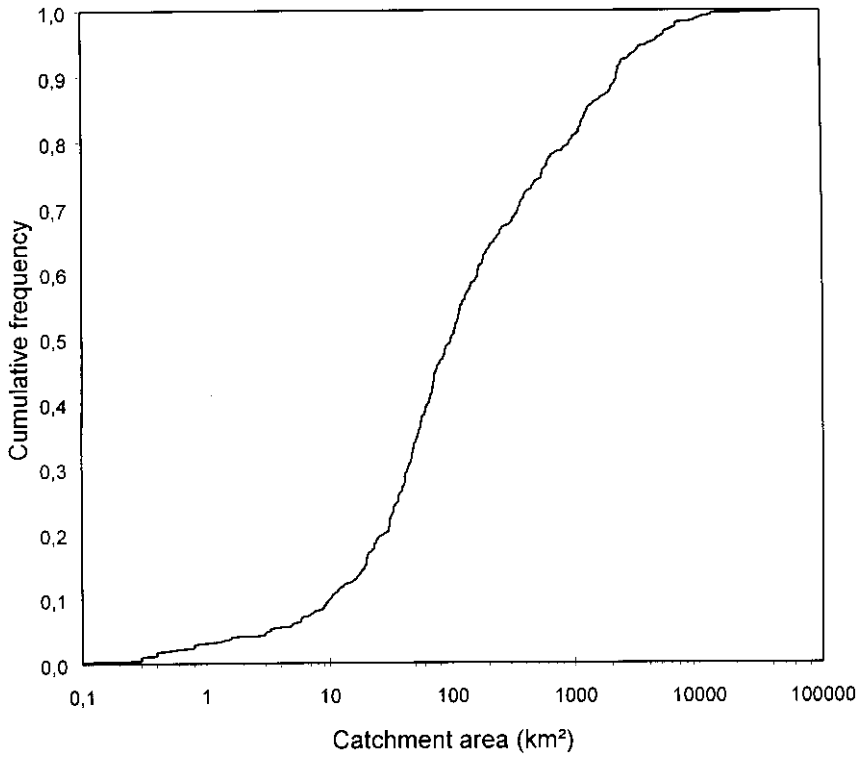


Fig. 2. Distribution of catchment area (in km^2).

seasonal variability of regimes and is defined by:

$$I\,(\%) = 100\frac{V_{\mathrm{mx}} - V_{\mathrm{mn}}}{\overline{V_{\mathrm{m}}}} \quad (1)$$

where V_{mx} is the maximum monthly rainfall (or streamflow); V_{mn} is the minimum monthly rainfall (or streamflow) and $\overline{V_{\mathrm{m}}}$ is the mean monthly rainfall (or streamflow).

Hydro-climatic conditions vary over a wide range from temperate through semi-arid to tropical humid, with a mean annual PE between 630 and 2040 mm, mean annual rainfall between 300 and 2300 mm and mean annual streamflow between 0.2 and 2040 mm. There is also a great diversity of seasonal rainfall and streamflow regimes, from very similar to very contrasting dry and wet seasons. Some catchments are ephemeral, whereas others have substantial groundwater influences. Catchment sizes range from small (0.1 km^2) to more than 10,000 km^2 with a median size of 100 km^2, as shown by the cumulative curve of catchment areas in Fig. 2. The two largest catchments are the Seine River at Paris in France (43,800 km^2) and the São Francisco River at the Três Marias Dam in Brazil (50,600 km^2). These catchments could be first thought to be too large to be adequately modelled with a lumped approach: they end up with having many more rain days with lower rainfall intensities than smaller catchments, making the conceptual representation of processes such as infiltration excess runoff pointless. However our investigation could reveal the extent to which a large size could prevent lumped models from performing satisfactorily.

In addition to various hydro-climatic characteristics, the catchment sample unavoidably includes a great variety of geology, pedology, topography or land-cover conditions.

5. Models

5.1. Model selection

The choice of models used in this comparative

assessment was based on a wide review of the literature. Because this study is oriented towards operational hydrology, all tested models have low data requirements and can be readily used in an operational context. Excluded from the comparison are models that are either spatially distributed, event-based or 'physically based' (in the sense that they require field measurements). The study concentrates on lumped, continuous, empirical or conceptual models. All these models are of the soil moisture accounting storage model type and the distinction between empirical and conceptual models only refers to the way model structures were built, as expressed by Edijatno et al. (1999).

The sample of possible models was restricted by excluding those that are too complex in terms of the number of optimised parameters such as the SSARR model (Speers, 1995) or the Stanford Watershed Model (Crawford and Linsley, 1963). Nineteen models were finally included in the study with a complexity ranging from three to less than 18 parameters in their original versions. They were developed in different contexts, sometimes using very different concepts, but most of them were already successfully tested on many case studies, under various climate and hydrological conditions. For example, the HBV model first developed for a North-European climate was later applied in more than 30 countries worldwide (see, e.g. Bergström, 1995). Here all models were tested on the whole sample of catchments, regardless of their validity range as advised by model designers. Consequently some models were applied out of their a priori application range, which could reveal whether these unusual test conditions could hamper model efficiency.

A simple equation model was also chosen as a baseline reference. The equation proposed by Tsykin (1985) and applied by Chiew et al. (1993) in their comparison of models was selected. It gives runoff as a function of antecedent rainfall by:

$$\mathrm{RUN}_i = a + b\,\mathrm{RAIN}_i^c\,\mathrm{RAIN}_{i-1}^d\,\mathrm{RAIN}_{i-2}^e\ldots \tag{2}$$

where RUN_i is runoff at time i, RAIN_i, RAIN_{i-1}, $\mathrm{RAIN}_{i-2}\ldots$ are rainfalls at times i, $i-1$, $i-2\ldots$, respectively, and a,b,c,d,$e\ldots$ are parameters. A four-parameter modified version is proposed here to introduce seasonal variations contained in PE data, so that runoff is now given by:

$$\mathrm{RUN}_i = \frac{X3}{1+\dfrac{\mathrm{PE}_i}{X2}}\prod_{j=0}^{n}\left(1+\frac{\mathrm{RAIN}_{i-j}}{X1}\right)^{X4^{1-j}} \tag{3}$$

where PE_i is the potential evapotranspiration at time i, n is the memory length of the system, and $X1$, $X2$, $X3$ and $X4$ are parameters. The chosen geometric progression of the exponents applied to rainfall reduces the number of optimised parameters and proved satisfactory. One additional parameter is introduced to simulate a delay, as explained later on.

5.2. Parsimony in conceptual modelling

Nash and Sutcliffe (1970) presented some principles for building models with optimised parameters in their paper. They expressed the need for both simplicity and lack of duplication in model structures. They also added the requirement of versatility, where adding parts to the model is only acceptable if they substantially increase model accuracy and robustness. Conscious of the need to address problems such as impacts of environmental changes, modellers felt compelled to develop more ambitious models, i.e. models with a greater number of parameters. Therefore, many models face the problem of 'equifinality' exposed by Beven (1993), a situation where different parameter sets may yield equivalent model outputs. This means that great uncertainty characterises these poorly defined parameters, with heavy problems of identification during calibration (see, e.g. Gupta and Sorooshian, 1983).

The issue of parameter uncertainty has drawn on a growing interest over the last two decades, partly because it influences the reliability of all further applications of models, e.g. flood frequency estimation (Cameron et al., 1999; Uhlenbrook et al., 1999) or detection of the effects of land-use changes (Mein and Brown, 1978; Nandakumar and Mein, 1997). To reduce this uncertainty, some authors have proposed the use of additional information on catchments, with multiresponse data. Studies by Kuczera and Mroczkowski (1998) with the CATPRO model or by Lamb et al. (1998) in the case of TOPMODEL showed clear improvements in parameter determination but none in streamflow simulation. These results suggest that it is a far more complex task to develop multi-output

Table 2
List of model structures with the retained number of parameters

Original model name and/or reference	Number of optimized parameters in the tested version
Tsykin (1985)	5
GR3J (Edijatno et al., 1999)	3
Model 16 (Bonvoisin and Boorman, 1992)	5
Model 15 (Bonvoisin and Boorman, 1992)	6
PDM (Moore and Clarke, 1981)	6
IHACRES (Jakeman et al., 1990)	7
TANK (Sugawara, 1995)	7
TOPMODEL (Beven and Kirkby, 1979)	7
MODGLO (Servat, 1986)	8
mSFB (Summer et al., 1997)	8
SMAR (Tan and O'Connor, 1996)	8
Wageningen (Warmerdam et al., 1997)	8
Xinanjiang (Zhao and Liu, 1995)	8
Arno (Todini, 1996)	9
Dawdy and O'Donnell (1965)	9
Georgakakos and Baumer (1996)	9
HBV (Bergström, 1995)	9
Institute of Hydrology lumped model (Blackie and Eeles, 1985)	9
MODHYDROLOG (Chiew and McMahon, 1994)	9
NAM (DHI, 1996)	9

models than precipitation-runoff models and that the internal variables of the latter can seldom be considered to reflect reality. The quantification of uncertainty by means of Monte-Carlo based approaches (Beven and Binley, 1992; Kuczera and Parent, 1998) only circumvents the problem.

Although the over-parameterisation issue is well known, only a few hydrologists adhere to the principle of parsimony advocated by Nash and Sutcliffe (1970) and later by Jakeman and Hornberger (1993) or Wheater et al. (1993) in conceptual rainfall–runoff modelling. However several examples in the literature seem to converge toward this point of view. Mein and Brown (1978) in the case of the modified 13-parameter SFB model showed that a drastic reduction in the number of optimised parameters only caused a slight reduction of the model performances. Chiew and McMahon (1994) indicated that, in the case of the MODHYDROLOG model, all 19 parameters are not necessary and that, in most cases, the calibration of only nine of them is sufficient to give adequate estimates of streamflow. Zhao and Liu (1995) noted that the output of the Xinanjiang model is generally sensitive to only seven of the 15 parameters in the model. In the case of the SMAR model family, Tan and O'Connor (1996) showed that the eight-parameter SMARY version is more versatile than the nine-parameter SMARG version. Recently Abdulla et al. (1999) observed that, in the case of the four baseflow parameters of the ARNO model, one or more of the parameters may not be useful and that a reparameterised model involving fewer parameters might perform equally well. Uhlenbrook et al. (1999) also reported that good simulations could be achieved with the HBV model over a wide range of parameter values even for sensitive parameters and that the increase in simulation quality was quite small when more complex versions of the model were used.

5.3. Structure simplifications and modifications

The appraisal of complete original models is not the aim of this paper. The aim is to detect whether some structures are more efficient than others, and more importantly, whether performance is connected to complexity, i.e. to the number of free parameters. Therefore, the number of optimised parameters of the tested model structures was arbitrarily limited to nine. Given this upper limit, model structures with a larger number of parameters were simplified. To this end, some parameters of low sensitivity were fixed or parts of the structure (loss or transfer functions) were simplified taking into account previous sensitivity analyses or modeller recommendations, when available. In some cases, modifications were introduced because they significantly improved the performances of tested structures on our data sample. To account for the time-lag in daily rainfall–runoff transformation, especially in large basins, a non-integer pure time delay was introduced to all model structures that did not originally include such a delaying function. With this parameter, all models can be used on the large catchments involved in the tests.

Table 2 lists model structures selected in the study with the number of optimised parameters of the tested version. The subsequent results may be different from those obtained with the original modelling methodologies. Therefore, each structure has been given a code name, M*ij*, where *i* is the number of optimised parameters and *j* is an arbi-

trary rank in the list of models with i parameters. These codes will be used in the following sections to display results. Since our study focuses on the issue of model complexity, it was not thought necessary to unveil the exact correspondence between model codes and original models. It was out of the scope of this paper to assess the more complex original models. The number of parameters was the only characteristic deemed of interest in the current comparison. Only the tested version (see Eq. (3)) of the Tsykin model was given the distinct value zero for j.

In order to provide the models with equal amount of information about the catchments, those that originally required specific catchment descriptors or were designed for specific climate input data had to be modified. For example, the distribution curve of the topographic index in TOPMODEL was parameterised with a logistic function using two additional parameters to be optimised, and temperature input data were replaced by PE data in IHACRES model.

A limitation of this research is that a thorough evaluation of each original model is beyond the scope of this wide and unified comparative study. Because of the large number of models, their structures are not described in detail (as carried out by Franchini and Pacciani, 1991; Franchini et al., 1996) but this information is available from the authors. A few general comments can be made. Model complexity ranges from three to nine optimised parameters (however, no four-parameter model was included). Structures involve from two to five storages. All model structures have a soil moisture accounting component that accounts for the evolution in time of catchment moisture state. Loss modules, determining effective rainfall, depend on one to five parameters. Procedures responsible for flow routing depend on two to seven parameters. The routing process considers at least two flow components and is either linear or non-linear. The number of parameters used for the transfer module is generally higher than that involved in the loss module.

6. Testing methodology

As mentioned above, the aim of this study was not to assess original models and their attached modelling methodologies such as those proposed in ready-to-use hydrological packages. These methodologies are not only based on a rainfall–runoff model structure, but also include specific procedures, chosen by the modeller, to select calibration periods, determine the optimum values of parameters, assess model fit quality or to determine uncertainties. Here, our sole objective was to examine crude model structures built on many different concepts. All tested models were given the same treatment in calibration and assessment procedures. This contrasts with other comparative studies (WMO, 1975; Perrin and Littlewood, 2000) where the parameter estimation technique and the modelling methodology were chosen by each model designer. The comparison methodology is described below.

6.1. Split sample test procedure

The evaluation of a model must take account of the primary objective of the model. In rainfall–runoff modelling, streamflow is the only model output. Indeed this variable is of crucial importance for many model applications (engineering design, water resources management, flood forecasting, etc.). The better the model simulations, the more reliable the applications. Therefore, this study (like others such as those by Chiew et al., 1993 or Ye et al., 1997) has focused on the ability of models to simulate streamflow.

Klemeš (1986) proposed a hierarchical assessment methodology to test model performances in calibration–simulation mode (split sample test) or in transposition mode (proxy-basin test). These tests can also include non-stationary conditions in the catchment, in which case they are called ‘differential’ tests. This scheme gives a key importance to model verification by assessing the transposability of models in time, space or under changing environmental conditions. This whole verification approach is powerful and desirable but quite cumbersome and is therefore seldom fully applied. Note, however, two studies which implemented this testing framework: Refsgaard and Knudsen (1996) applied it to three models of different types (NAM, MIKE SHE and WATBAL) on three Zimbabwean catchments; Donelly-Makowecki and Moore (1999) tested

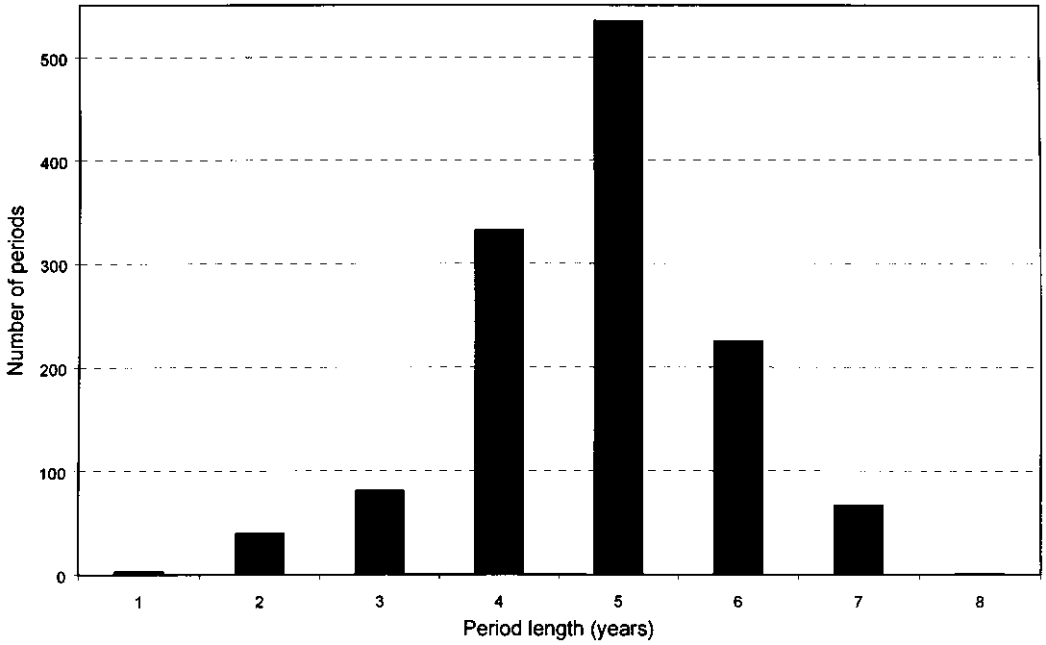


Fig. 3. Distribution of the length of test periods (in years) on the sample of 429 catchments.

TOPMODEL and simple lumped reservoir models with this scheme on two Canadian catchments in British Columbia. Because of the size of our sample, the framework adopted here is based only on the split-sample test, in line with most of the comparative studies reported in the literature. For each catchment, data time-series were split into between two and six independent (non-overlapping) sub-periods, depending on the length of the total available record. The periods vary between one and eight years but are mostly four- to six-year long, as shown in Fig. 3. In total, 1284 test periods were identified on the 429 catchments, i.e. an average of three periods per catchment.

For each catchment, the models were successively calibrated on each sub-period and then tested in verification mode on all the remaining periods. For example, on a catchment with six test periods, six calibration and 30 verification tests were performed. This represents a total of 3204 verification tests for the 429 catchments. To our knowledge, this is the most wide-ranging model evaluation reported in the literature so far. The distribution of calibration periods and verification tests between different data origins is given in Table 3. French catchments represent about 72% of the test sample and 56% of the verification tests. Hence the comparison results are strongly influenced by the French hydrological context, while basins from other countries (44% of the verification tests) provide a valuable broadening of the spectrum of hydro-climatic conditions, thereby improving the reliability of the assessment.

In this study, only verification test results are used to illustrate model performances, since in an operational context, models are always used in verification mode.

6.2. Warm-up period (state initialisation)

The quality of model simulations depends partly on how well unknown values are determined. Parameters are the most crucial unknowns. However, the contents in model storages at the beginning of simulation periods are also unknown values which must be determined so as not to jeopardise the simulations. Special care was taken to prevent problems linked to inappropriate initial conditions. First, the initial level in model storages was set to average seasonal values for the corresponding time of year. Second, a one-year warm-up period was inserted at the

Table 3
Distribution of calibration periods and verification tests by catchment origin (with proportion of the total sample in brackets)

Country	Australia	Brazil	France	Ivory Coast	United States	Total
Number of catchments	26 (6.1%)	4 (0.9%)	307 (71.6%)	10 (2.3%)	82 (19.1%)	429
Number of calibration periods	63 (4.9%)	12 (0.9%)	856 (66.7%)	41 (3.2%)	312 (24.3%)	1284
Number of verification tests	98 (3.1%)	24 (0.7%)	1780 (55.5%)	156 (4.9%)	1146 (35.8%)	3204

beginning of each period to attenuate the effect of the storage initialisation. Model results for the first year were ignored in the computation of goodness-of-fit criteria, as done by Chiew and McMahon (1994) or Edijatno et al. (1999). The division of data records into sub-periods was based on calendar years. Fig. 4 shows the splitting methodology in the case of a 23-year record, which is split into four sub-periods including the one-year warm-up period.

6.3. Optimisation technique

Many authors agree that the quality of model parameters partly depends on the efficiency and robustness of the optimisation algorithm (Duan et al., 1992). Considerable efforts have been made over the past two decades to develop and implement more complex and more efficient calibration techniques able to cope with many parameters and highly non-linear structures. The reasons for this are to be found in problems inherent to conceptual models, notably parameter interaction, low sensitivity of some parameters or presence of local optima (Johnston and Pilgrim, 1976; Duan et al., 1992), where classical local search methods are theoretically of limited efficiency. With enhanced computer power came proposals of global search techniques intended to explore a large part of the response surface (e.g. Genetic Algorithm applied by Wang, 1991, or Shuffled Complex Evolution — University of Arizona method developed by Duan et al., 1992). Although these methods have a theoretical advantage over local search techniques and have proved more efficient in many cases (see, e.g.

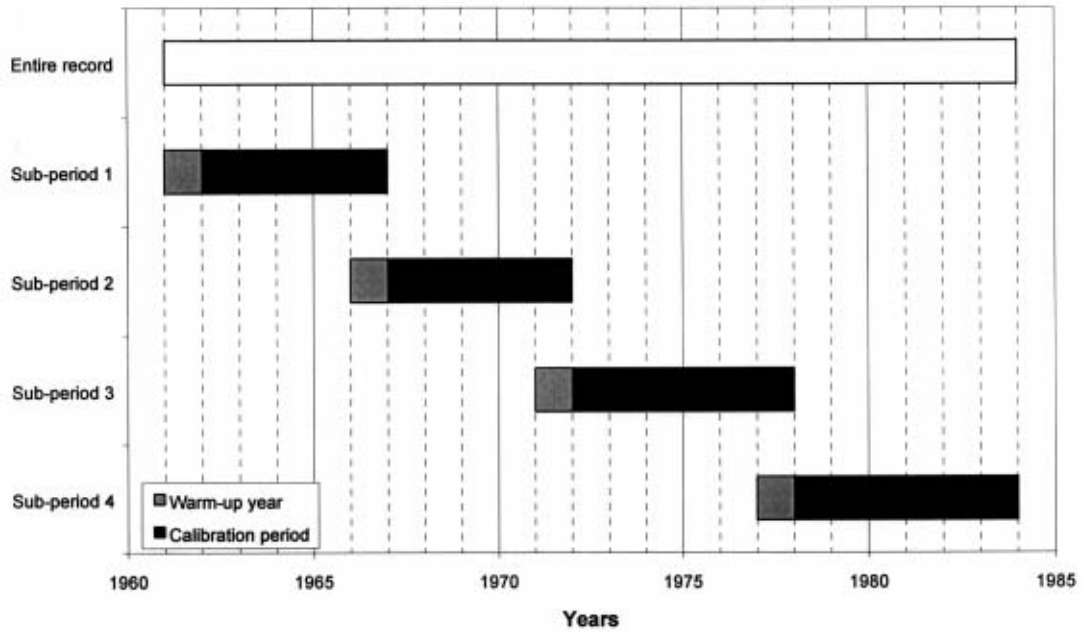


Fig. 4. Example of data record splitting into two sub-periods of six years and two sub-periods of seven years including a warm-up year.

Tanakamaru, 1995), they, like local search methods, are confronted with problems of model structures, especially the non-uniqueness of optimum, which makes different parameters sets equivalent in terms of the value of the objective function. Results obtained by Wang (1991) showed for example that different optimum parameter sets with equal objective function value could be found for a seven-parameter version of the Xinanjiang model when calibrated with a Genetic Algorithm technique. In an interesting study, Gan and Biftu (1996) used real data to test local and global search techniques on four conceptual models. They showed that, in most cases, global convergence cannot be achieved because it is impossible to identify one single optimum. Global search techniques have a slight advantage in calibration which generally does not extend beyond the calibration period. For the evaluation of the 19 model structures, a simple local search technique was deemed sufficient. It was also less demanding in terms of computation considering the size of our test sample (1284 calibration runs for 19 models).

The selected optimisation technique is the steepest descent method summarised in Edijatno et al. (1999). In this method each optimisation run starts with an initial parameter set identified for each model as the one yielding the best results on the whole sample of catchments. Then the algorithm evolves step-by-step in the parameter space toward the 'optimum' parameter values. Outside the scope of this paper, the application of the algorithm to four model structures with different numbers of parameters, showed that the combined use in calibration of two different initial parameter sets for each structure did not produce significant improvements in model performances in verification mode.

6.4. Objective function

It must be reminded first that the choice of an objective function depends primarily on the intended applications of the model and that it is therefore essentially user-dependent (Diskin and Simon, 1977). The objective function selected to calibrate the models is of the least-square type. It is based on the formulation proposed by Nash and Sutcliffe (1970) and given by:

$$\mathrm{CR1}(\%) = 100\left(1 - \frac{\sum_{i=1}^{n}(Q_{\mathrm{obs},i} - Q_{\mathrm{cal},i})^2}{\sum_{i=1}^{n}(Q_{\mathrm{obs},i} - \overline{Q_{\mathrm{obs}}})^2}\right) \quad (4)$$

where $Q_{\mathrm{obs},i}$ and $Q_{\mathrm{cal},i}$ are the observed and calculated streamflows at time step i, $\overline{Q_{\mathrm{obs}}}$ is the mean observed streamflow over the calibration period and n the number of time steps. Because of non-constant variance of model errors, this criterion tends to emphasise large errors, i.e. those generally occurring during flood events. A more all-purpose criterion is obtained by using root-square transformed streamflow, as carried out by Chiew and McMahon (1994). The objective function is therefore given by:

$$\mathrm{CR2}(\%) = 100\left(1 - \frac{\sum_{i=1}^{n}(\sqrt{Q_{\mathrm{obs},i}} - \sqrt{Q_{\mathrm{cal},i}})^2}{\sum_{i=1}^{n}(\sqrt{Q_{\mathrm{obs},i}} - \overline{\sqrt{Q_{\mathrm{obs}}}})^2}\right) \quad (5)$$

This is the criterion used here to calibrate all models. Hence, the model structures are required to satisfy this objective in calibration as well as in verification mode. However, as this is a comparative exercise, it was thought worthwhile to add other criteria in order to judge the simulation quality in validation.

6.5. Assessment criteria

In the complex operation of evaluating model performances in terms of streamflow simulation quality, assessment criteria must be selected. The interpretation of graphical (qualitative) criteria is quite subjective, as discussed by Houghton-Carr (1999). Therefore, numerical criteria were preferred here. Weglarczyk (1998) noticed however that there is no best statistical quality criterion for hydrological simulation models. Hence if a single criterion is chosen, model verification becomes a partial exercise. Following recommendations by WMO (1975, 1986) or The ASCE task committee (1993), a panel of several (four) numerical criteria was chosen to assess model performances, including the objective function

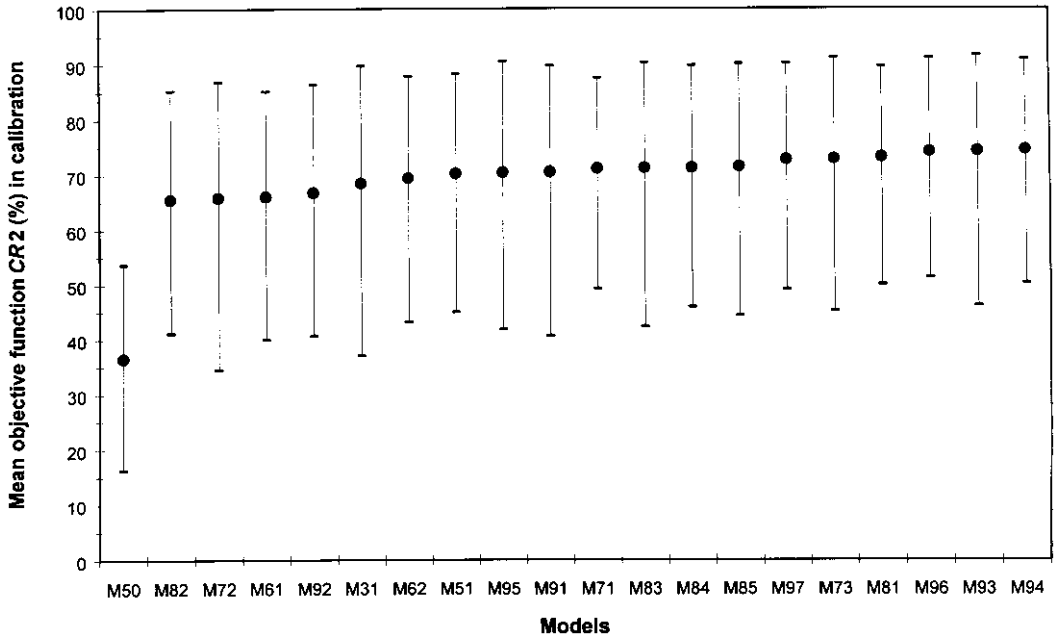


Fig. 5. Mean results with 0.1 and 0.9 percentiles of CR2 obtained by M50 and 19 conceptual models on the 1284 calibrations.

used in calibration. The assessment criteria presented below are built on three analytical formulations of model error, namely quadratic, absolute and cumulative errors, given in Eqs. (6), (7) and (9), respectively.

Two numerical measurements are based on the mean square model error SE defined by:

$$\mathrm{SE} = \frac{1}{n}\sum_{i=1}^{n}(Q_{\mathrm{obs},i} - Q_{\mathrm{cal},i})^2 \tag{6}$$

Based on SE, the first two assessment criteria CR1 and CR2 defined in Eq. (4) and (5) were used. These measures vary between $-\infty$ and 100% for perfect agreement and are easy to interpret. They quantify the ability of the model to explain streamflow variance, i.e. the improvement achieved by any model in simulating streamflow compared to a basic reference model simulating a constant streamflow equal to the mean observed one. CR1 puts more emphasis on flood simulations. The Nash–Sutcliffe criterion has been extensively used in hydrology, although Garrick et al. (1978) or Martinec and Rango (1989) have shown that it has its weaknesses.

The third criterion is built on the mean absolute model error AE defined by:

$$\mathrm{AE} = \frac{1}{n}\sum_{i=1}^{n}|Q_{\mathrm{obs},i} - Q_{\mathrm{cal},i}| \tag{7}$$

A transformation similar to the one adopted for CR1 or CR2 allowed the criterion to vary from $-\infty$ to 100%. This criterion (CR3) is defined by:

$$\mathrm{CR3}(\%) = 100\left(1 - \frac{\sum_{i=1}^{n}|Q_{\mathrm{obs},i} - Q_{\mathrm{cal},i}|}{\sum_{i=1}^{n}|Q_{\mathrm{obs},i} - \overline{Q_{\mathrm{obs}}}|}\right) \tag{8}$$

This criterion is potentially useful in a forecasting context for example, where the simulations must be as close as possible to the observed values at every time step (Ye et al., 1997).

The fourth criterion is based on the mean cumulative error CE of the model defined by:

$$\mathrm{CE} = \frac{1}{n}\sum_{i=1}^{n}(Q_{\mathrm{obs},i} - Q_{\mathrm{cal},i}) \tag{9}$$

This error can also be written in relative terms (balance error) by:

$$CE^*(\%) = 100.\left(\frac{\sum_{i=1}^{n}(Q_{obs,i} - Q_{cal,i})}{\sum_{i=1}^{n} Q_{obs,i}}\right) \quad (10)$$

The zero value indicates perfect agreement for CE^*. As CE^* can take positive or negative values, CE^* values obtained over different periods can yield an average value close to zero, i.e. a good model balance although this is not the case. To circumvent this problem and make all four criteria vary in the same interval $]-\infty; 1]$ with perfect agreement indicated by 100%, the following formulation is proposed for criterion CR4:

$$CR4(\%) = 100\left[1 - \left|\sqrt{\frac{\sum_{i=1}^{n} Q_{cal,i}}{\sum_{i=1}^{n} Q_{obs,i}}} - \sqrt{\frac{\sum_{i=1}^{n} Q_{obs,i}}{\sum_{i=1}^{n} Q_{cal,i}}}\right|\right] \quad (11)$$

CR4 measures the ability of the model to correctly reproduce streamflow volumes over the studied period. Criterion CR4 is different from the three other criteria (CR1–CR3) in that it does not measure a departure from observed values at each time step of the simulation (for this reason, CR4 cannot be used alone as the calibration criterion).

With all four formulations, values can be averaged over the sample of catchments, thus facilitating statistical comparisons of model performances given our large test sample. These four quality measurements were thought sufficient to assess basic qualities of model simulations. They are not, however, completely independent of one another.

7. Results and discussion

The 19 daily rainfall–runoff model structures, as well as the baseline M50 model, were successively applied to the 429 catchments of the sample. All 1284 calibration runs and their corresponding 3204 verification tests were performed for each structure. In the discussion of the results, we concentrate mainly on performances obtained in verification mode, since this is the most common mode of model use in an operational context. In the following, the term 'model' sometimes used alone stands for 'model structure', as we only assess the performances of model structures, not full modelling methodologies.

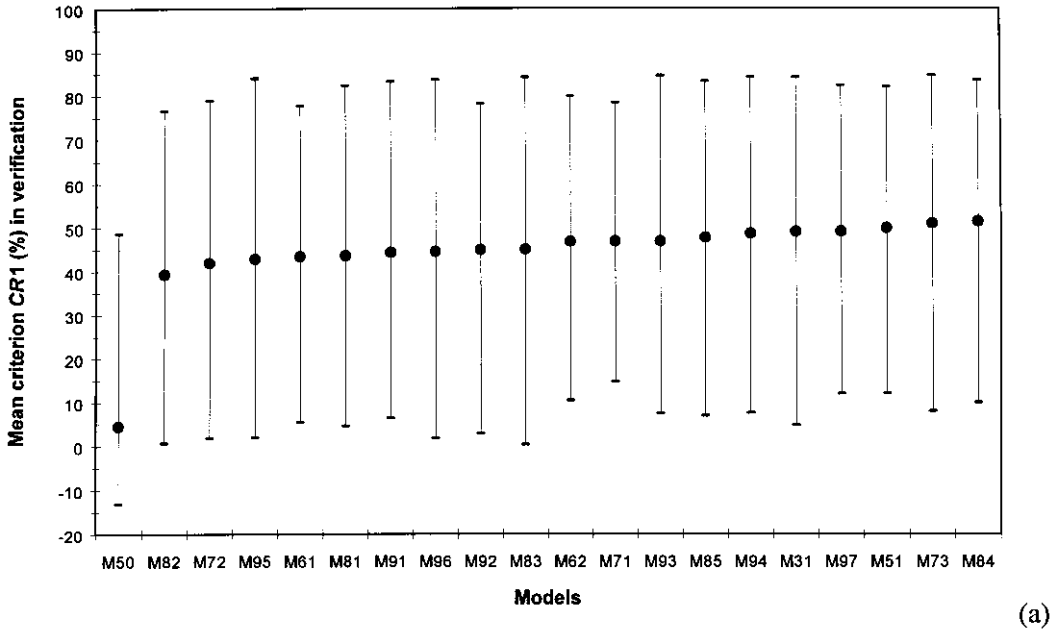
In the following sections, we analyse the results of this comparative assessment and attempt to answer a few questions that we believe to be some of the most important in rainfall–runoff modelling.

7.1. How well do soil moisture accounting models actually perform?

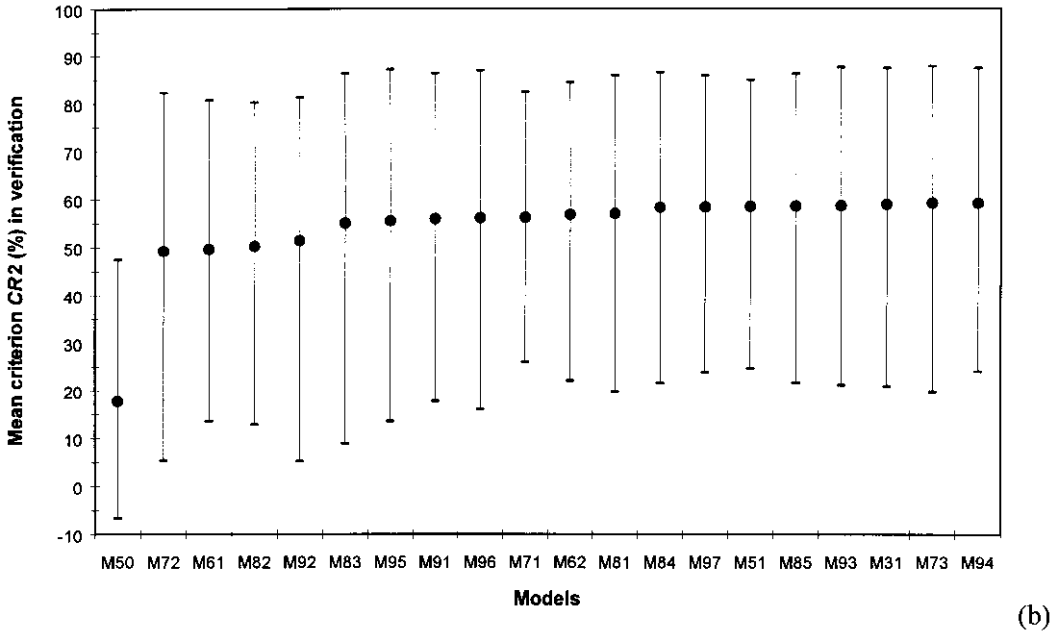
The baseline M50 model used in this study represents a simple alternative approach to soil moisture accounting modelling methods. It was already used by Chiew et al. (1993) in a slightly different form. Thus, it was interesting to begin by comparing its results with those obtained by the 19 SMA models studied here.

Fig. 5 shows mean performances averaged on the 1284 calibrations together with 0.1 and 0.9 percentiles for the objective function (CR2). Fig. 6(a)–(d) shows mean performances averaged on the 3204 verifications, along with 0.1 and 0.9 percentiles for all four quality measurements (CR1–CR4). It is immediately clear that the performances of the 19 models are systematically much better (in calibration or verification) than those produced by the baseline model M50. This model is unable to provide satisfactory simulations on average, which corroborates the results obtained by Chiew et al. (1993) at a daily time step. We believe that this lack of accuracy stems both from the absence of a reservoir routing function and from too rudimentary a simulation of antecedent conditions. In contrast, the 19 conceptual model structures all include a procedure that (explicitly or implicitly) follows moisture conditions over time. Thus, the modelling of antecedent catchment moisture conditions appears to be an efficient approach on many of the test catchments.

Strikingly, the performances of the 19 models appear relatively similar. This is further illustrated in Table 4, which shows the ranges covered by the four assessment criteria in calibration and simulation modes: mean performances are generally less than 10% apart. The value of the percentiles 0.9 are even

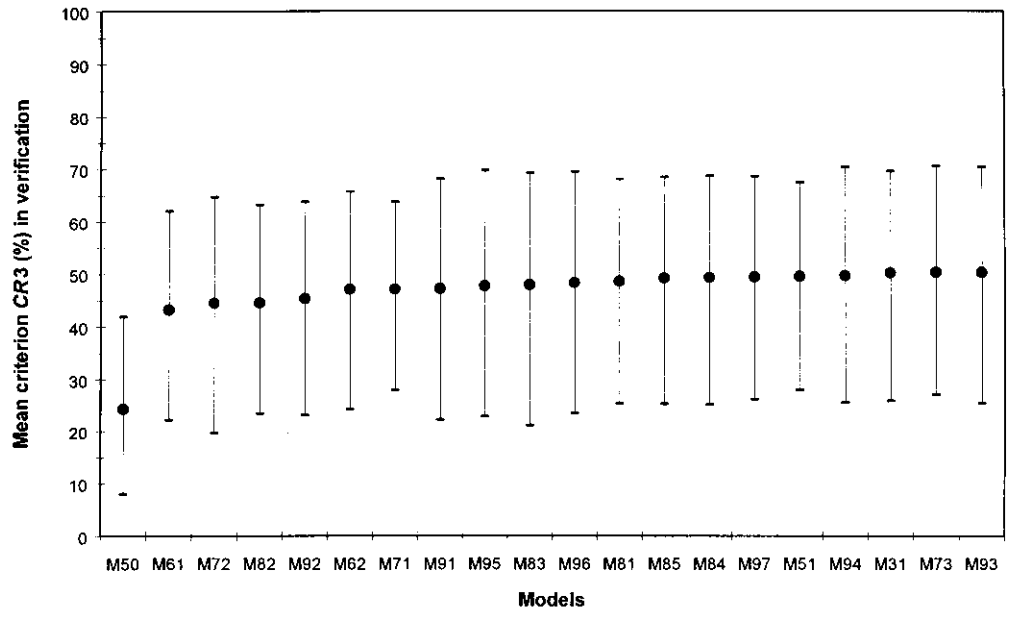


(a)

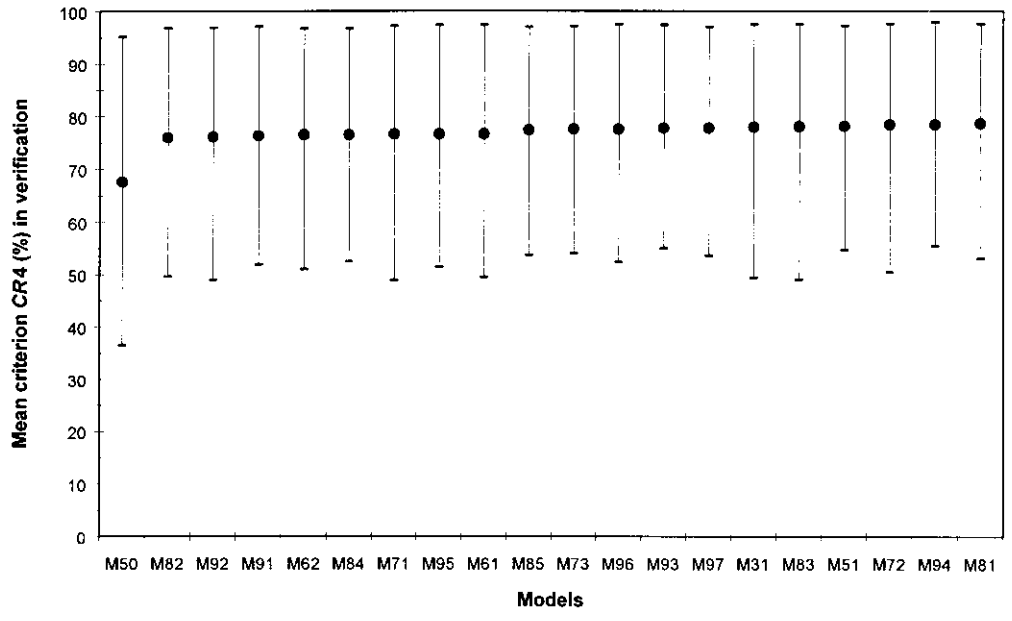


(b)

Fig. 6. Mean results with 0.1 and 0.9 percentiles obtained by M50 and 19 conceptual models on the 3204 verification tests for: (a) CR1; (b) CR2; (c) CR3; and (d) CR4.



(c)



(d)

Fig. 6. (*continued*)

closer, which suggests that all models are able to reach similarly high performances on a number of catchments, albeit not on the same catchments. Comparatively, the range of percentiles 0.1 is greater.

Fig. 7(a) and (b) summarise the ranks obtained by each model in calibration and verification for the four assessment criteria. It shows that no single model was superior to the others according

Table 4
Ranges of performances (means, 0.1 and 0.9 percentiles) of the 19 model structures for the four assessment criteria CR1 to CR4 in calibration and verification

		Percentile 0.1 (%)	Mean (%)	Percentile 0.9 (%)
CR1	Calibration	20.4–39.4	57.3–68.2	82.0–89.0
	Verification	0.3–14.7	39.3–51.2	76.6–84.5
CR2	Calibration	34.4–51.2	65.5–74.3	85.1–91.5
	Verification	5.2–26.0	49.2–59.2	80.2–87.8
CR3	Calibration	34.0–40.4	52.2–59.5	67.5–75.2
	Verification	19.7–27.9	43.2–50.3	62.0–70.5
CR4	Calibration	64.4–73.8	83.9–89.3	97.3–99.3
	Verification	48.9–55.5	76.0–78.7	96.7–97.9

to all quality criteria either in calibration or in verification. This also underlines the advantage of using several criteria to assess validation performances. It reveals that some models prove complementary qualities by providing good performances according to criteria different from the objective function CR2 used in calibration.

The following comments complement the above results. All models obtain a mean CR1 rating above 60% for 120 catchments (28% of the test sample). Conversely, all models obtain performances below 60% for 149 catchments (35% of the test sample). This means that for about one third of the catchments, rainfall–runoff models do not produce satisfactory answers. What are the reasons for model failure? The inadequacy of model structure is likely to be the main one. Errors in data, impaired streamflow, groundwater, snowmelt or human influences may also be responsible for the inability of the model to simulate streamflow correctly on these catchments. All these factors explain the quite low mean performances of the 19 models (less than 60% in verification mode for the first three criteria). It can also be noticed that the application of models outside their a priori validity range was neither the root cause of model failure nor a factor limiting significantly their numerical efficiency. Besides results indicate that a large catchment size does not prevent the tested lumped models from performing satisfactorily, which indicates that the mathematical formulation of some processes in conceptual models, considered as valid only on small catchments, remains also valid on larger ones.

7.2. How is model robustness reflected by calibration/verification results?

Together with reliability, discussed in the next section, robustness is one of the most important qualities of any model. We propose to quantify robustness by the decrease of average performance between calibration and verification. This decrease is a well known but undesirable feature of hydrological models.

It is clear from the comparison between Figs. 5 and 6(a) and from Table 5, which gives the mean, maximum and minimum differences of mean performances from calibration to verification, that there is a significant drop in the mean values of criteria from calibration to verification. However, this drop is not the same for all 19 model structures, as indicated by the minimum and maximum differences in Table 5. In the case of CR2, for example, the drop ranges from 9.4 to 17.9%. This shows that some models are more robust, i.e. that they have better stability, and are thus more likely to yield simulations of the same quality level in verification as in calibration.

7.3. Can some structures guarantee better reliability?

The previous results are now analysed in terms of model reliability. We propose to quantify model reliability by the number of catchments where a model ranks among the best performing structures. A model is said to be in class 1 when it ranks among the first three and in class 2 when it ranks from fourth to sixth.

The average performance of each model on all periods is used to sort the models on each catchment by

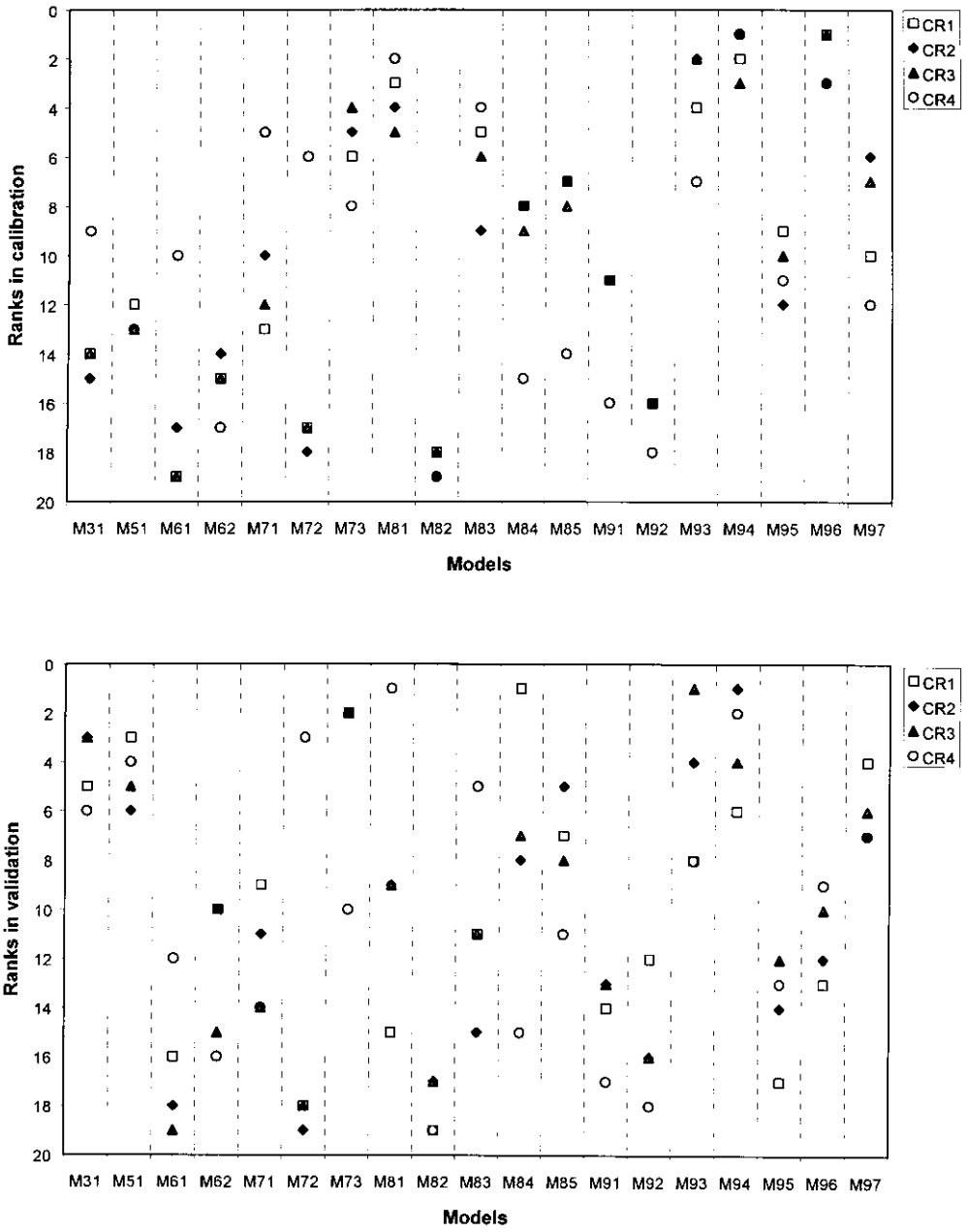


Fig. 7. Model ranking for the four assessment criteria: (a) in calibration mode; and (b) in simulation mode.

order of increasing performance. In Fig. 8, each model is shown with the number of catchments on which it ranked in class 1 or 2 according to the CR1 criterion (Nash–Sutcliffe criterion) in verification, i.e. the number of basins where it was among the six best performing structures. The number ranges from 31 to 240 (between 6 and 162 for class 1 only). This disparity indicates that some models reach comparatively better performances on a greater number of catchments than others and therefore are more

Table 5
Mean, minimum and maximum differences between mean performances of criteria CR1 to CR4 from calibration to verification for the 19 models

	CR1	CR2	CR3	CR4
Mean difference (%)	−16.7	−14.5	−8.5	−8.9
Minimum difference (%)	−11.9	−9.4	−5.2	−7.1
Maximum difference (%)	−23.8	−17.9	−11.2	−11.3

reliable. Most amazingly, we noticed that it was always possible to find at least three catchments in the sample for which any one of the 19 structures ranked as the most satisfactory. This indicates that the ranks of the models, shown in Figs. 5 and 6, still depend on the characteristics of the test sample. It means that by taking sub-samples of the 429 catchments, we could have ranked the models differently, as illustrated by Fig. 9 where the catchment sample was split into two sub-samples (sub-sample 1 with the 307 French basins and sub-sample 2 with the remaining 122 basins). It shows different model rankings and different levels of performance. Therefore, we believe that the reliability of a model can only be assessed on large test samples with varied catchments.

7.4. Can we talk of 'equifinality' between model structures?

Here, we investigate the possibility of extending the concept of 'equifinality' (see, e.g. Beven, 1993) to model structures, by checking whether different model structures are able to provide similar results on a catchment. We base the following analysis on mean CR1 values per catchment obtained on verification:

- For 41 catchments, there is virtually no difference between the results obtained by the first two models.
- For 66% of catchments the difference between the first and the second model is less than 2% and it is greater than 10% of the CR1 rating for only 43 catchments (out of 429). For only 15 of these 43 catchments the CR1 performance of the top model is greater than 60%. This indicates that the major differences between the best models generally occur when all models have difficulties in simulating the behaviour of a catchment.
- Lastly, the difference between the first and the

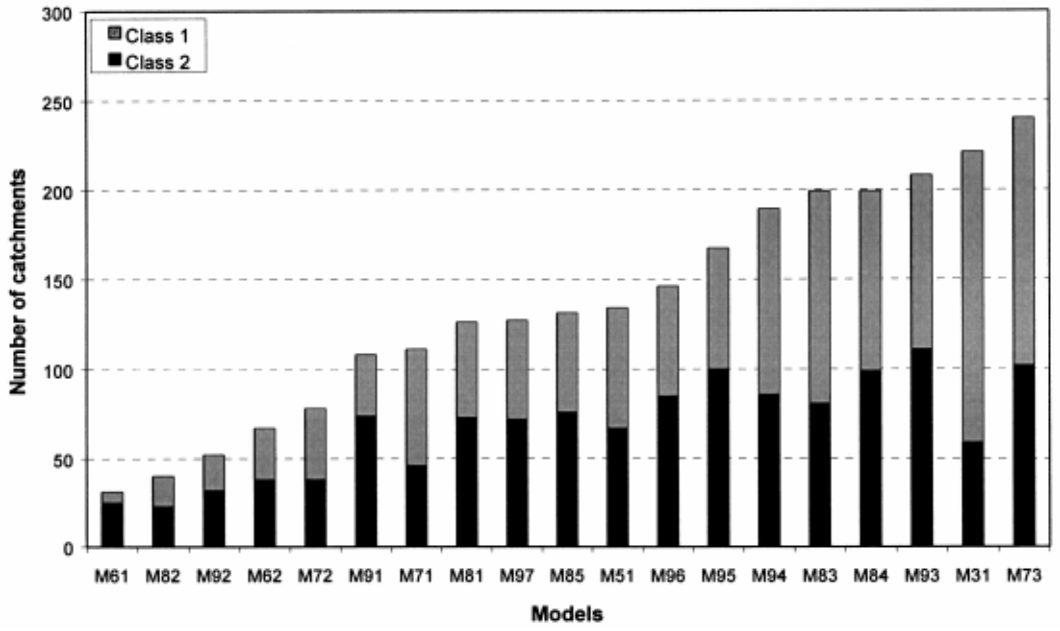
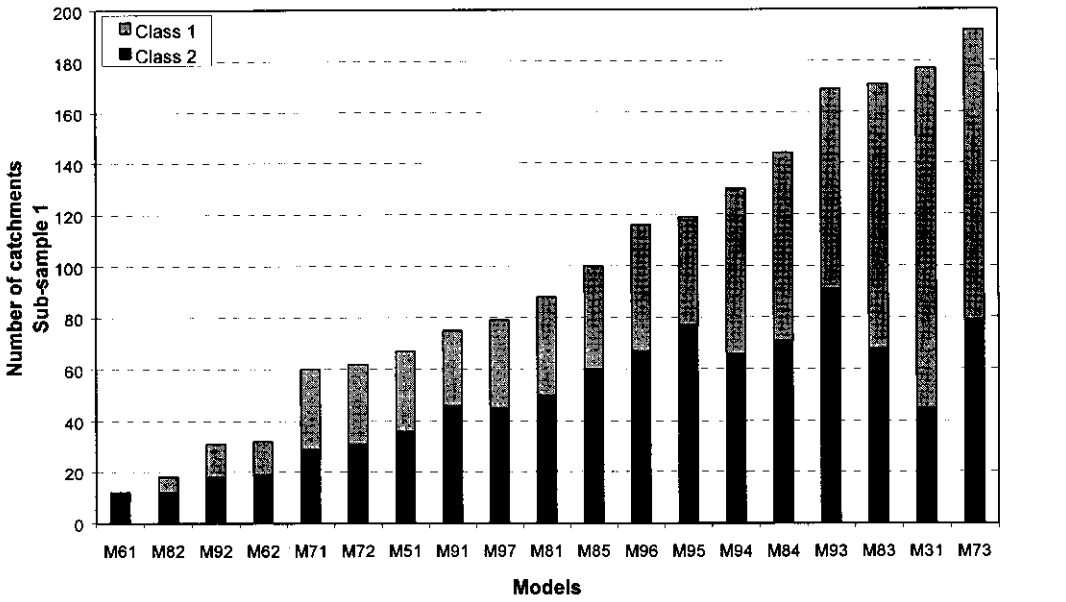
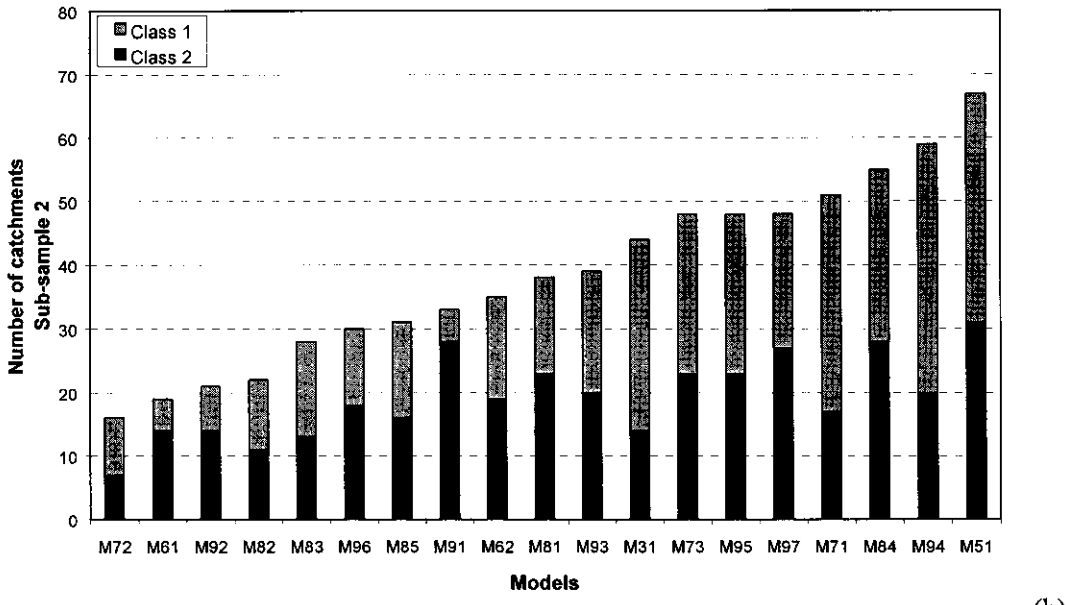


Fig. 8. Number of catchments (in the whole sample of basins) for which a model ranks in class 1 or in class 2 according to CR1 performances.



(a)



(b)

Fig. 9. Number of catchments for which a model ranks in class 1 or in class 2 according to CR1 performances on: (a) sub-sample 1; and (b) sub-sample 2.

third model is less than 2% for 44% of catchments and the difference between the first and the tenth model is less than 10% in the case of 271 catchments (63% of test basins).

This analysis tends to support a kind of 'equifinality' principle, i.e. on a large proportion of the catchments (modelling cases), different models can give similar results. The results presented by Franchini and Pacciani (1991) and complemented by Franchini et al. (1996) in the case of the Sieve basin also show that eight different models yield very similar results in simulation mode.

7.5. Does the number of free parameters increase model performances?

During the past 30 years, many modellers of the rainfall–runoff relationship have considered that increasing model complexity is the paramount solution for improving model performances. Here, we quantify model complexity by the number of parameters optimised during the calibration phase. Fig. 10(a) and (b) show the performances of models as measured by CR2, the Nash–Sutcliffe criterion on square root transformed streamflow, plotted against the number of model parameters. In calibration mode, models with a larger number of parameters generally benefit from this increase in degrees of freedom and yield a better fit of observed data. But this trend disappears at the verification stage where models with a limited number of parameters achieve results as good as those of more complex models.

Ye et al. (1997) and Gan et al. (1997) already noticed this lack of overall superiority of complex models over simpler ones. This can be partly explained by the stability of performances from calibration to simulation, as shown in Fig. 10(c). As discussed previously, it seems that the complex models tend to be less robust, i.e. they tend to have less stable performances than simpler ones.

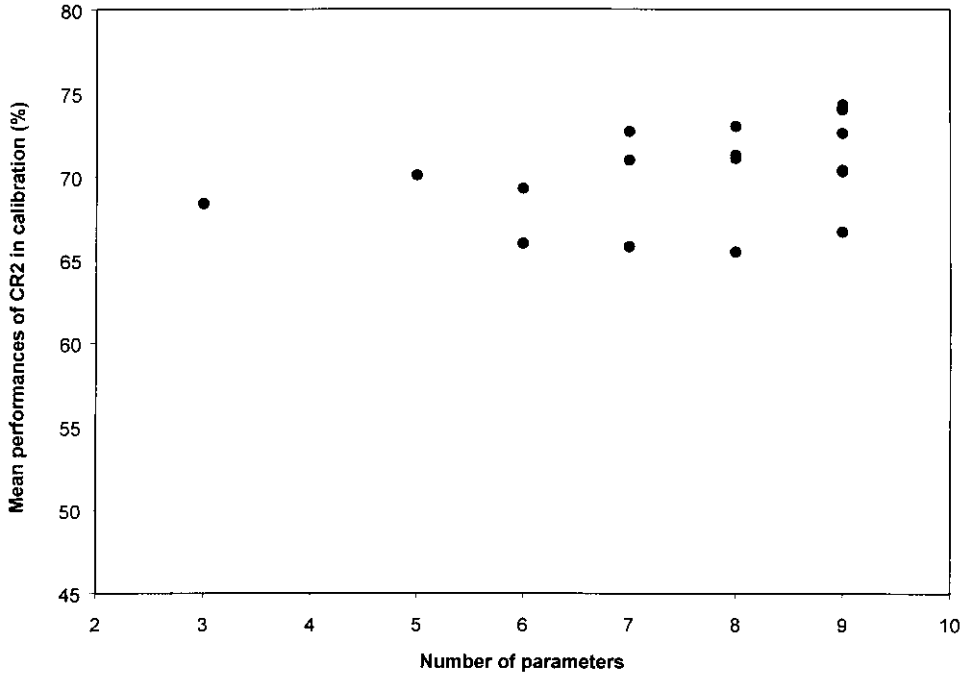
It is also worth noting that, in calibration or simulation, models with the same number of parameters may produce quite different results. This strengthens the argument by Gan et al. (1997) that the structure of the model (i.e. the type of storages, the nature of mathematical functions, the way elements are interrelated in the structure, the parameterisation of loss or routing modules) is of crucial importance for the success in modelling. Complexity alone cannot guarantee good and reliable performances. To date, model structures do not seem to be accurate enough to support a high complexity.

Fig. 10(b) shows that different structures with different levels of complexity can attain similar quality levels of performances. A possible conclusion, also proposed by Nash and Sutcliffe (1970), could be to advise the use of the simpler among these, almost equivalent, models. This idea is discussed further in the following paragraph.

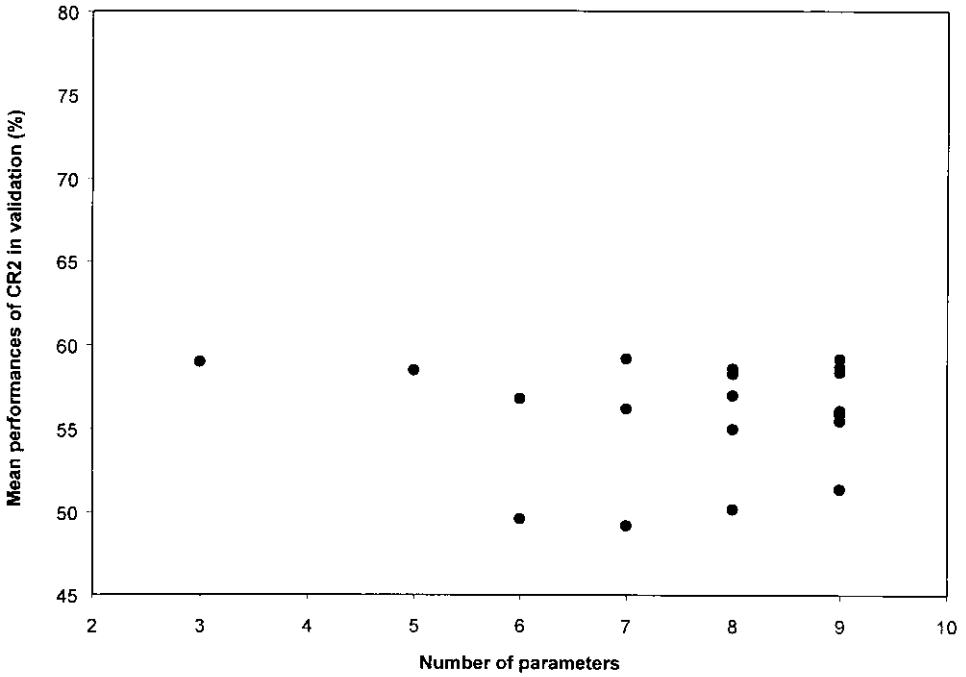
7.6. Complementarity of model structures: can we reach the 'ideal' model?

We have mentioned previously that, among the 19 tested models, all do not achieve top performances on the same catchments, and that it is always possible to find a catchment where one model outperforms all the others. How could this complementarity between different model structures be used? Let us consider the situation in which a modeller would have the choice between the 19 model structures to simulate rainfall–runoff relationship on each catchment. In each case study, testing of all structures could determine which is the most suitable model for that particular case. This was done for all 429 catchments, by retaining for each one the best performance in simulation mode provided by the 19 structures. We considered this 'best performance set' as one that could be derived from an 'ideal model', which from now on will be referred to as the $M_{\alpha\omega}$ model.

Fig. 11 shows the distributions of CR1 values obtained in verification by M50, $M_{\alpha\omega}$ and the 19 conceptual models. There is a significant gap between $M_{\alpha\omega}$ performances and those of the best conceptual model. On average, the $M_{\alpha\omega}$ model reaches 64.4% for criterion CR1 whereas the mean performance of the best conceptual model is only 54.9%. This seems to indicate possible complementarity between structures. To test this idea we have looked for the pairs of structures whose association was the most successful. Table 6 shows the symmetric matrix, for eight of the 19 models, featuring the number of catchments for which at least the model in row or the model in



(a)



(b)

Fig. 10. Mean performances for criterion CR2: (a) in calibration; (b) in verification; and (c) difference between verification and calibration, versus the number of model parameters.

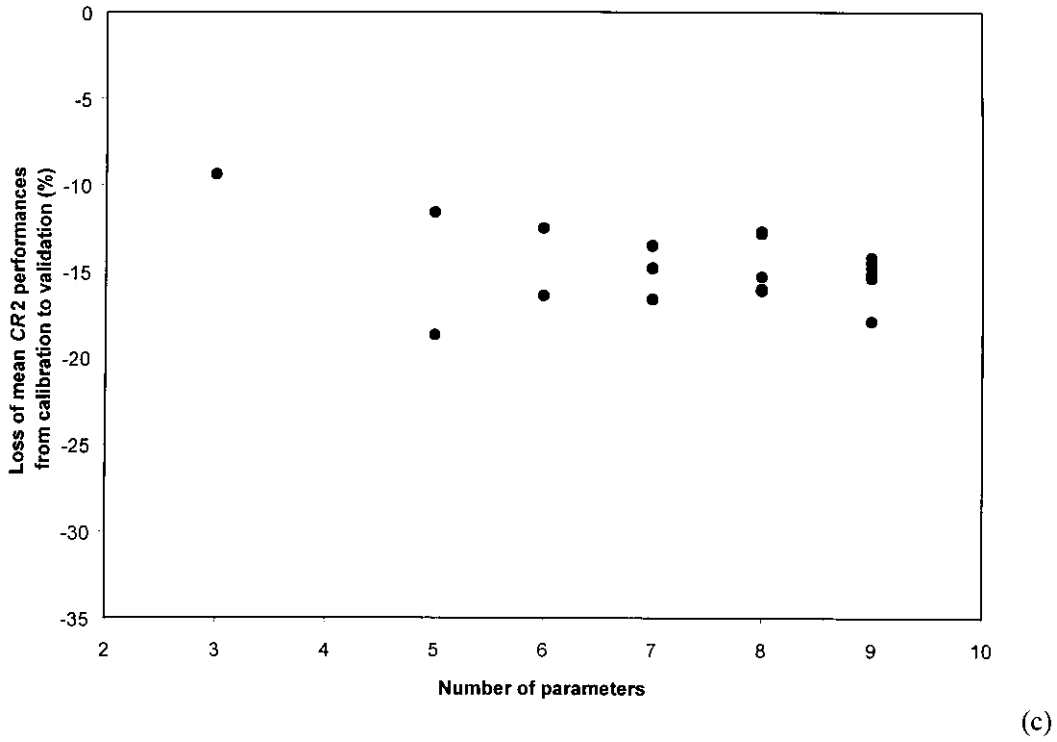


Fig. 10. (*continued*)

column ranked among the best four structures. On the diagonal are the numbers of catchments for which one single model ranked among the four best structures. If two models are associated, the resulting combined model is clearly more reliable. The improvement may be almost 100 catchments as compared with the best performing one of the two individual models.

This shows that each model contains specific components whose efficiency can be satisfactorily complemented by others. Hence the association of two models can alleviate the limitations of both. Finding the models that together give the best performance could indicate which components from one could be successfully introduced into the other. Ideally associating the 19 models gives the results of the $M_{\alpha\omega}$ model, with very substantial improvements. In this sense Shamseldin et al. (1997) or more recently Shamseldin and O'Connor (1999) discussing forecasting, demonstrated that model association can be used successfully by combining outputs of different models a posteriori. The use of several different models can, however, make several model applications (such as those using regionalisation) more difficult to implement. We believe that prior attempts to include the best components of different models into a single one are worth considering. The resulting model would be more efficient and more reliable, although the distribution curve of the $M_{\alpha\omega}$ model in Fig. 11 probably sets the upper (and unreachable) limit for possible improvements using components that are parts of the 19 models.

8. Conclusion and future work

This paper discusses the degree of model complexity (as reflected by the number of optimised parameters) required to simulate rainfall–runoff relationships on a wide variety of catchments. To assess the actual value of complexity in a model, an extensive testing scheme was carried out on 19 daily lumped model structures with three to nine optimised

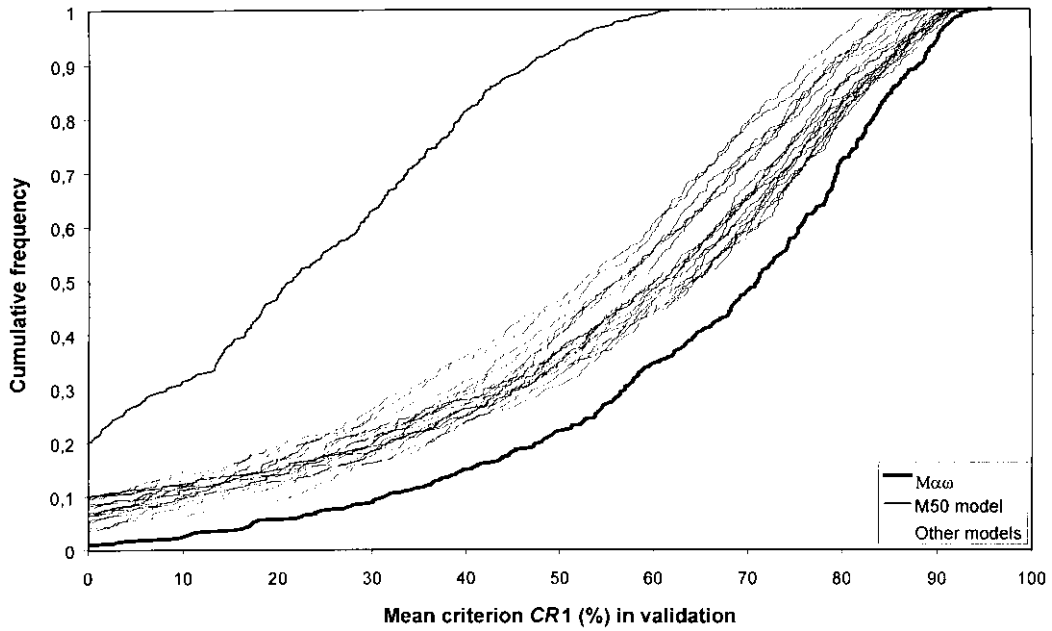


Fig. 11. Distributions of mean performances per catchment obtained in verification with criterion CR1.

parameters. Model structures were tested for their ability to simulate streamflow, i.e. the target variable of such models in an operational hydrological context. Their performances were compared on a sample of 429 catchments located mainly in France but also in the United States, Australia, the Ivory Coast and Brazil.

It was demonstrated that very simple models can achieve a level of performance almost as high as models with more parameters. These more complex models are subject to over-parameterisation, which prevents them from reaching their potential performance level. It seems therefore that the number of free parameters might be restricted to between three and five in lumped rainfall–runoff models. Steefel and Van Cappellen (1998) commenting on a paper by Oreskes et al. (1994), stated that in many respects, the ultimate test of a particular model is its simplicity relative to its performance on a given number of observations. From this point of view, it can be argued that, for equivalent performances, the simplest models should be preferred. They do effectively cause less problems of parameter uncertainty, which is of utmost importance, e.g. in regionalization where parameter identifiability is a key issue (Wheater et al., 1993). Steefel and Van Cappellen (1998) add that there is however a limit to model simplicity and this limit is reached when the model fails to adequately explain the observations (i.e. simplicity alone cannot be used as a valid criterion for how good a model is). We share this opinion and believe that one can strive for model parsimony as long as this does not impair the ability of the model to simulate streamflow values.

This study also provides insights into the important issue of model complementarity. In our test, none of

Table 6
Number of catchments for which at least one of the two models (row or column) is ranked in the first four best structures (results based on mean CR1 criterion per catchment in simulation)

First model	Second model							
	M31	M51	M73	M83	M84	M93	M94	M95
M31	185							
M51	241	89						
M73	284	238	170					
M83	270	228	263	149				
M84	270	198	229	257	132			
M93	270	201	262	242	230	130		
M94	260	207	263	228	239	224	134	
M95	248	181	222	223	214	202	204	105

the models was found to be the best one in all cases. In contrast, associating complementary model structures was shown to improve the results of single models applied independently. This opens up new possibilities of merging efficient components from different models into a single simple one, where structure and well-chosen parameterisation should be the main concerns. Research is under way to explore the possibility of building such a model, starting from the simplest structure and gradually and accurately increasing the complexity as needed to improve model performances. This approach, advised by Nash and Sutcliffe (1970) also follows recommendations by Bergström (1991), who states that it avoids 'the frustration of abandoning seemingly elegant concepts and theories when going from complex to simpler model structures'. We hope that this paper will stimulate further work on model testing and on development of new, simpler and more efficient models that would benefit operational hydrology.

Acknowledgements

ENGEES (National engineering school for water and environment in Strasbourg, France) is thanked for its support of this study. The authors thank Dr Francis H.S. Chiew at the Department of Civil and Environmental Engineering of the University of Melbourne, Australia, for providing data sets of the Australian catchments and for his fruitful review of this paper; Dr Eric Servat at the Research Institute for Development (IRD, formerly ORSTOM) in Montpellier, France, for providing data sets for catchments in the Ivory Coast; Jane L. Thurman from the Water Data Center at the US Department of Agriculture, Beltsville, United States, for providing ARS data; Dr Nilo de Oliveira Nascimento at the University of Minas Gerais, Belo Horizonte, Brazil for providing data for Brazilian catchments. Data sets of American catchments from the MOPEX database were made available for the Workshop on regionalization of parameters of hydrological and atmospheric land surface models, 27–28 July 1999, Birmingham, Great Britain (see Web site www.nws.noaa.gov/oh/mopex). For French catchments, streamflow data were provided by the HYDRO database of the French Ministry for Environment, and Météo France provided potential evapotranspiration and precipitation data. The anonymous reviewer is also thanked for his comments which helped to improve the text.

References

Abdulla, F.A., Lettenmaier, D.P., Liang, X., 1999. Estimation of the ARNO model baseflow parameters using daily streamflow data. Journal of Hydrology 222, 37–54.

The ASCE Task Committee, 1993. The ASCE task committee on definition of criteria for evaluation of watershed models of the watershed management committee, Irrigation and Drainage division, Criteria for evaluation of watershed models. Journal of Irrigation and Drainage Engineering 119 (3), 429–442.

Bergström, S., 1991. Principles and confidence in hydrological modelling. Nordic Hydrology 22, 123–136.

Bergström, S., 1995. The HBV model. In: Singh, V.P. (Ed.), Computer Models in Watershed Modeling, Water Resources Publications, pp. 443–476.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrological Processes 6, 279–298.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Bulletin 24 (1), 43–69.

Beven, K.J., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources 16, 41–51.

Blackie, J.R., Eeles, C.W.C., 1985. Lumped catchment models. In: Anderson, M.G., Burt, T.P. (Eds.). Hydrological Forecasting. Wiley, New York, pp. 311–345 (chap. 11).

Bonvoisin, N.J., Boorman, D.B., 1992. Daily rainfall–runoff modelling as an aid to the transfer of hydrological parameters. Report to MAFF, Institute of Hydrology, Wallingford, UK.

Cameron, D.S., Beven, K.J., Tawn, J., Blazkova, S., Naden, P., 1999. Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty). Journal of Hydrology 219, 169–187.

Chiew, F., McMahon, T., 1994. Application of the daily rainfall–runoff model MODHYDROLOG to 28 Australian catchments. Journal of Hydrology 153, 383–416.

Chiew, F.H.S., Stewardson, M.J., McMahon, T.A., 1993. Comparison of six rainfall–runoff modelling approaches. Journal of Hydrology 147, 1–36.

Crawford, N.H., Linsley, R.K., 1963. A conceptual model of the hydrologic cycle. IAHS Publication No. 63, pp. 573–587.

Danish Hydraulic Institute — DHI, 1996. NAM — MIKE 11. Documentation and User's Guide. Danish Hydraulic Institute, Hørsholm, Danemark.

Dawdy, D.R., O'Donnell, T., 1965. Mathematical models of catchment behavior. American Society of Civil Engineers Proceedings 91 (HY4), 123–137.

Diskin, M.H., Simon, E., 1977. A procedure for the selection of objective functions for hydrologic simulation models. Journal of Hydrology 34, 129–149.

Donelly-Makowecki, L.M., Hydrological forecasting, R.D., 1999.

Hierarchical testing of three rainfall–runoff models in small forested catchments. Journal of Hydrology 219 (3–4), 136–152.

Duan, Q., Sorooshian, S., Gupta, V.K., 1992. Effective and efficient global optimization for conceptual rainfall–runoff models. Water Resources Research 28 (4), 1015–1031.

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

Farnsworth, R.K., Thompson, E.S., Peck, E.L., 1982. Evaporation atlas for the contiguous 48 United States. National Oceanic and Atmospheric Administration, National Weather Service, NOAA technical report NWS No. 33, Washington, DC.

Franchini, M., Pacciani, M., 1991. Comparative analysis of several conceptual rainfall–runoff models. Journal of Hydrology 122, 161–219.

Franchini, M., Wendling, J., Obled, C., Todini, E., 1996. Physical interpretation and sensitivity analysis of the TOPMODEL. Journal of Hydrology 175, 293–338.

Gan, T.Y., Biftu, G.F., 1996. Automatic calibration of conceptual rainfall–runoff models: optimization algorithms, catchment conditions, and model structure. Water Resources Research 32 (12), 3513–3524.

Gan, T.Y., Dlamini, E.M., Biftu, G.F., 1997. Effects of model complexity and structure, data quality and objective function on hydrologic modeling. Journal of Hydrology 192, 81–103.

Garrick, M., Cunnane, C., Nash, J.E., 1978. A criterion of efficiency for rainfall–runoff models. Journal of Hydrology 38, 375–381.

Georgakakos, K.P., Baumer, O.W., 1996. Measurement and utilization of on-site soil moisture data. Journal of Hydrology 184, 131–152.

Gupta, V.K., Sorooshian, S., 1983. Uniqueness and observability of conceptual rainfall–runoff model parameters: the percolation process examined. Water Resources Research 19 (1), 269–276.

Hargreaves, G.H., Samani, Z.A., 1982. Estmating potential evapotranspiration. Journal of Irrigation and Drainage Engineering, Technical Note 108 (3), 225–230.

Houghton-Carr, H.A., 1999. Assessment criteria for simple conceptual daily rainfall–runoff models. Hydrological Sciences Journal 44 (2), 237–261.

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

Jakeman, A.J., Littlewood, I.G., Whitehead, P.G., 1990. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. Journal of Hydrology 117, 275–300.

Johnston, P.R., Pilgrim, D.H., 1976. Parameter optimization for watershed models. Water Resources Research 12 (3), 477–486.

Kite, G.W., 1978. Development of a hydrologic model for a Canadian watershed. Canadian Journal of Civil Engineering 5, 126–134.

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrological Sciences Journal 31 (1), 13–24.

Kuczera, G., Mroczkowski, M., 1998. Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. Water Resources Research 34 (6), 1481–1489.

Kuczera, G., Parent, E., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. Journal of Hydrology 211, 69–85.

Lamb, R., Beven, K., Myrabø, S., 1998. Use of spatially distributed water table observations to constrain uncertainty in a rainfall–runoff model. Advances in Water Resources 22 (4), 305–317.

Lavabre, J., Sempere Torres, D., et Cernesson, F., 1993. Changes in the hydrological response of a small Mediterranean basin a year after a wildfire. Journal of Hydrology 142, 273–299.

Leviandier, T., 1988. Mise en oeuvre et interprétation de la comparaison de modèles (Implementation and interpretation of model comparison) (in French). La Houille Blanche (5–6), 395–398.

Lørup, J.K., Refsgaard, J.C., et Mazvimavi, D., 1998. Assessing the effects of land use change on catchment runoff by combined use of statistical tests and hydrological modelling: case studies from Zimbabwe. Journal of Hydrology 205, 147–163.

Martinec, J., Rango, A., 1989. Merits of statistical criteria for the performance of hydrological models. Water Resources Bulletin 25 (2), 421–432.

Michaud, J., Sorooshian, S., 1994. Comparison of simple versus complex distributed runoff models on a midsized semiarid watershed. Water Resources Research 30 (3), 593–605.

Mein, R.G., Brown, B.M., 1978. Sensitivity of optimized parameters in watershed models. Water Resources Research 14 (2), 299–303.

Moore, I.D., Mein, R.G., 1975. An evaluation of three rainfall–runoff models. Proceedings of the Hydrological Symposium, Sydney, May 1975. Inst. Eng. Aust., Nat. Conf. Publ., vol. 75, no. 3, pp. 122–126.

Moore, R.J., Clarke, R.T., 1981. A distribution function approach to rainfall–runoff modeling. Water Resources Research 17 (5), 1367–1382.

Morton, F.I., 1983. Operational estimates of actual evapotranspiration and their significance to the science and practice of hydrology. Journal of Hydrology 66, 1–76.

Nandakumar, N., Mein, R.G., 1997. Uncertainty in rainfall–runoff model simulations and the implications for predicting the hydrologic effects of land-use change. Journal of Hydrology 192, 211–232.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part I — a discussion of principles. Journal of Hydrology 27 (3), 282–290.

Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation and confirmation of numerical models in the Earth Sciences. Science 263, 644–646.

Paturel, J.E., Servat, E., Vassiliadis, A., 1995. Sensitivity of conceptual rainfall–runoff algorithms to errors in input data — case of the GR2M model. Journal of Hydrology 168, 11–125.

Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London A 193, 120–145.

Perrin, C., Littlewood, I.G., 2000. A comparative assessment of two rainfall–runoff modelling approaches: GR4J and IHACRES. In: Elias, V., Littlewood, I.G. (Eds.). Proceedings of the Liblice Conference (22–24 September 1998). IHP-V, Technical Documents in Hydrology No. 37UNESCO, Paris, pp. 191–201.

Refsgaard, J.C., Knudsen, J., 1996. Operational validation and inter-

comparison of different types of hydrological models. Water Resources Research 32 (7), 2189–2202.

Servat, E., 1986. Présentation de trois modèles globaux conceptuels déterministes: CREC 5. MODGLO, MODIBI. ORSTOM, Département F, Unité de Recherche 604.

Servat, E., Dezetter, A., 1991. Selection of calibration objective functions in the context of rainfall–runoff modelling in a Sudanese savannah area. Hydrological Sciences Journal 36 (4), 307–331.

Servat, E., Dezetter, A., 1992. Modélisation de la relation pluie-débit et estimation des apports en eau dans le nord-ouest de la Côte d'ivoire (in French). Hydrologie Continentale 7 (2), 129–142.

Shamseldin, A.Y., O'Connor, K.M., Liang, G.C., 1997. Methods for combining the outputs of different rainfall–runoff models. Journal of Hydrology 197, 203–229.

Shamseldin, A.Y., O'Connor, K.M., 1999. A real-time combination method for the outputs of different rainfall–runoff models. Hydrological Sciences Journal 44 (6), 895–912.

Speers, D.D., 1995. SSARR Model. In: Singh, V.P. (Ed.), Computer Models in Watershed Hydrology, chap. 11. Water Resources Publications, pp. 367–394.

Steefel, C.I., Van Cappellen, P., 1998. Reactive transport modeling of natural systems. Journal of Hydrology 209, 1–7.

Sugawara, M., 1995. The development of a hydrological model — tank. In: Kite, G. (Ed.), Time and the River. Essays by Eminent Hydrologists, chap. 7. Water Resources Publications, pp. 201–258.

Summer, N.R., Fleming, P.M., Bates, B.C., 1997. Calibration of a modified SFB model for twenty-five Australian catchments using simulated annealing. Journal of Hydrology 197, 166–188.

Tan, B.Q., O'Connor, K.M., 1996. Application of an empirical infiltration equation in the SMAR conceptual model. Journal of Hydrology 185, 275–295.

Tanakamaru, H., 1995. Parameter estimation for the Tank Model using global optimisation. Transactions of JSIDRE 178, 103–112.

Thurman, J.L., Roberts, R.T., 1995. New strategies for the Water Data Center. Journal of Soil and Water Conservation 50 (5), 530–531.

Todini, E., 1996. The ARNO rainfall–runoff model. Journal of Hydrology 175, 339–382.

Tsykin, E.N., 1985. Multiple nonlinear statistical models for runoff simulation and prediction. Journal of Hydrology 77, 209–226.

Uhlenbrook, S., Seibert, J., Leibundgut, C., Rodhe, A., 1999. Prediction uncertainty of conceptual rainfall–runoff models caused by problems in identifying model parameters and structure. Hydrological Sciences Journal 44 (5), 779–797.

Vandewiele, G.L., Xu, C.Y., Win, N.L., 1992. Methodology and comparative study of monthly models in Belgium, China and Burma. Journal of Hydrology 134, 315–347.

Wang, Q.J., 1991. The genetic algorithm and its application to calibrating conceptual rainfall–runoff models. Water Resources Research 27 (9), 2467–2471.

Warmerdam, P.M.M., Kole, J., Chormanski, J., 1997. Modelling rainfall–runoff processes in the Hupselse Beek research basin. Ecohydrological processes in small basins. Proceedings of the Strasbourg Conference (24–26 September 1996), IHP-V, Technical Documents in Hydrology No.14, pp. 155–160.

Weeks, W.D., Hebbert, R.H.B., 1980. A comparison of rainfall–runoff models. Nordic Hydrology 11, 7–24.

Weglarczyk, S., 1998. The interdependence and applicability of some statistical quality measures for hydrological models. Journal of Hydrology 206, 98–103.

Wheater, H.S., Jakeman, A.J., 1993. Progress and directions in rainfall–runoff modelling. In: Jakeman, A.J., et Beck, M.B., McAleer, M.J. (Eds.). Modelling Change in Environmental Systems, 5. Wiley, New York, pp. 101–132 (chap. 5).

World Meteorological Organization — WMO, 1975. Intercomparison of conceptual models used in operational hydrological forecasting. Operational Hydrology Report No. 7, World Meteorological Organization, Geneva, Switzerland.

World Meteorological Organization — WMO, 1986. Intercomparison of models of snowmelt runoff. Operational Hydrology Report No. 23, World Meteorological Organization, Geneva, Switzerland.

World Meteorological Organization — WMO, 1992. Simulated real-time intercomparison of hydrological models. Operational Hydrology Report No. 38, World Meteorological Organization, Geneva, Switzerland.

Xu, C.Y., Vandewiele, G.L., 1995. Parsimonious monthly rainfall–runoff models for humid basins with different input requirements. Advances in Water Resources 18, 39–48.

Yang, X., et Michel, C., 2000. Flood forecasting with a watershed model: a new method of parameter updating. Hydrological Sciences Journal 45 (4), 537–546.

Yang, X., Parent, E., Michel, C., et Roche, P.A., 1995. Comparison of real-time reservoir-operation techniques. Journal of Water Resources Planning and Management 121 (5), 345–351.

Ye, W., Bates, B.C., Viney, N.R., Silvapan, M., Jakeman, A.J., 1997. Performance of conceptual rainfall–runoff models in low-yielding ephemeral catchments. Water Resources Research 33 (1), 153–166.

Zhang, X., Lindström, G., 1996. A comparative study of a Swedish and a Chinese hydrological model. Water Resources Bulletin 32 (5), 985–994.

Zhao, R.J., Liu, X.R., 1995. The Xinanjiang model. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology, chap. 7. Water Resources Publications, pp. 215–232.