Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment–MOPEX. IAHS Publ. 307, 2006.

Compilation of the MOPEX 2004 results

N. CHAHINIAN[1,2], V. ANDRÉASSIAN[1], Q. DUAN[3], V. FORTIN[4,5], H. GUPTA[6], T. HOGUE[7], T. MATHEVET[1,8], A. MONTANARI[9], G. MORETTI[10], R. MOUSSA[11], C. PERRIN[1], J. SCHAAKE[12], T. WAGENER[6,13] & Z. XIE[14]

1 *Hydrology group, Cemagref Antony, Parc Tourvoie BP 44, 92163 Antony cedex, France*
nanee.chahinian@agrocampus-rennes.fr

2 *Now at: Institut de Recherche pour le Développement(IRD), UMR HydroSciences–Université Montpellier II, Case Courrier MSE, Place Eugène Bataillon, 34095 Montpellier cedex 5, France*

3 *Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, California 94550, USA*

4 *Institut de Recherche d'Hydro Qubec 1800, boulevard Lionel-Boulet, Varennes, Québec J3X 1S1, Canada*

5 *Now at: Numerical Prediction Research, Meteorological Research Division, Environment Canada, 2121, North Service Road, Trans-Canada Highway, Dorval, Quebec H9P 1J3, Canada*

6 *University of Arizona, NSF STC SAHRA, Marshall Building Room 532A, 845 North Park Avenue, Tuscon, Arizona 85721, USA*

7 *University of California Los Angeles, Department of Civil and Environmental Engineering, 5732C Boelter Hall, California 90095-1593, USA*

8 *Now at: EDF-DTG, BP 41, 38040 Grenoble cedex, France*

9 *University of Bologna, Viale del Risorgimento 2, 40136 Bologna, Italy*

10 *Ingenieurbüro Winkler und Partner GmbH, Schloßstr. 59A, 70176 Stuttgart, Germany*

11 *Institut National de la Recherche Agronomique, UMR LISAH, 2 place Pierre Viala, 34060 Montpellier cedex, France*

12 *NOAA/NWS/OHD, Hydrology Laboratory, Silver Spring, Maryland 20910, USA*

13 *Penn State University, Department of Civil and Environmental Engineering, 212 Sackett Building, University Park, Pennsylvania 16802, USA*

14 *Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China*

Abstract As part of the MOPEX 2004 workshop, the participants were asked to submit simulations using the common database provided for the workshop (see Chahinian *et al.*, this issue). The simulations were then analysed and the evaluation criteria computed to compare the models' performance for the gauged and ungauged modes using six criteria describing the model's performance in both high and low flow conditions. The comparisons were undertaken for all three participation levels (i.e. 3, 12 and 40 catchments). The results indicate that on the 3-catchment level model ranking may vary according to the tested criterion and catchment. Hence a larger number of catchments are necessary to evaluate the models' performance. Among the 10 models tested on the 12 and 40 catchment samples in gauged mode, GR5H, Mordor and SAC-SMA rank as the top three models. When analysing model results in the ungauged mode, SAC-SMA ranks as the best of the four tested models. The analysis of the submitted files highlights the need for continuing efforts to develop model parameterization strategies for ungagued catchments in order to improve prediction in ungauged basins (PUB).

Key words model comparison; MOPEX; parameter estimation; PUB; rainfall–runoff modelling; ungauged catchments

INTRODUCTION

In view of the focus on hydrological predictions on ungauged catchments as highlighted by the IAHS PUB decade, the MOPEX (Model Parameter Estimation Experiment) extended its scope to the parameterization of hydrological models on ungauged catchments. This paper summarizes the MOPEX 2004 workshop results and provides insights into the simulation capacities of the tested models on both gauged and ungauged catchments.

Model parameter estimation is a bigger problem for ungauged catchments where traditional calibration techniques (Hendrickson *et al*., 1988; Duan & Gupta 1992; Gupta *et al*., 2003) cannot be used because of the lack of data. Hence some authors turn to regionalization methods in order to link model parameters to the catchments' descriptors (Micovic & Quick, 1999; Seibert, 1999; Perrin, 2000; Kokkonen *et al*., 2003; Van der Linden & Woo, 2003; Croke *et al*., 2004; Lee *et al*., 2005; Merz & Blöschl, 2005) or use *a priori* model estimates based on the catchments' physical properties (Koren *et al*., 2003; Leavesley *et al*., 2003). In order to succeed these methods need a strong relationship between the catchment characteristics and the model's parameters as well as a large data set to extract statistically significant results. These two conditions are rarely satisfied. Hence trying to infer post-calibration relationships between calibrated parameters and catchment properties is, in most instances, a perilous exercise because of the inter-dependencies between the various parameters (Johnston & Pilgrim, 1976; Beven, 1993). In addition, even the most "physically based" models need some calibration (Abbott *et al*., 1986) to account for the discrepancies between the parameter measurement and model application scales (Haverkamp *et al*., 1998). Finally, it is often difficult to find long time series of rainfall–runoff data to ensure a sound calibration and large databases containing both hydrological and morphological attribute data to derive these relations. In this setting, model parameter estimation experiments such as MOPEX, involving different models and large databases, provide a creative interaction framework where modellers can share their experiences and benefit from the latest advances in model parameter estimation techniques. In this perspective, the MOPEX 2004 workshop was a clear success marked by high participation and the use of new hourly data.

LIMITS OF THE PRESENT COMPARISON

Even if we believe the Paris MOPEX workshop results to be extremely valuable, we deem it necessary to mention that, for the most part, the models have only been tested on either 3 or 12 catchments, and that this is necessarily too small a data set to provide results of general value. For example, in his comparison of 20 rainfall–runoff models over 429 catchments, Perrin (2000) found several combinations of 3 or 12 catchments showing that 19 of the 20 models were performing best. Nonetheless, it is obvious that the results of this comparison will help participating modellers to improve their models.

METHODS

For the first time since the start of the MOPEX programme, three distributed models took part in the comparison (Afdeff, Hydrotel, Modspa). In total 13 models were tested (Table 1). Four models were used at the daily time step (ModSpa, SAC-SMA,

Table 1 Participant information and general model characteristics.

Participant	Model name	Reference	Runoff/rainfall model	Channel routing function
UCLA University of Arizona	SAC-SMA	Burnash *et al.*, 1973	SAC-SMA	Kinematic wave
Hydro-Quebec	Hydrotel	Fortin *et al.* (1995); (2001)	3 Layer Richards 1D	Kinematic or diffusive wave
INRA Montpellier Cemagref Antony	ModSpa	Moussa (1991) ; (1993)	Diskin-Nazimov (1995)	Diffusive wave
Cemagref Antony	GR4J	Edijatno *et al.* (1999) Perrin *et al.*, (2001)	Non-linear reservoir	Unit hydrograph
	GR5H	Mathevet (2005)	Non-linear reservoir	Unit hydrograph
	HBV0	Perrin *et al.* (2001) Bergström & Forsman (1973)	Linear reservoir/modified HBV	Unit hydrograph
	IHAC	Jakeman *et al.* (1990) Perrin *et al.* (2001)	Linear reservoir/modified IHACRES	Pure delay
	Mordor	Mathevet (2005)	Non-linear reservoirs	Unit hydrograph
	Topmo	Beven & Kirkby (1978) Perrin *et al.* (2001)	Modified TopModel	Pure delay
University of Bologna	Affdef	Brath *et al.* (2002) Moretti & Montanari (2005)	SCS Reservoir model	Maskingham-Cunge
Laurence Livermore Laboratory	NOAH Land Surface Model (NLS)	Chen *et al*, (1996)	Multi-layer soil and energy balance	Linearized St Venant
National Weather Service	SAC-SMA	Burnash *et al.* (1973)	SAC-SMA	Kinematic wave
	SWB	Schaake *et al.* (1996)	Reservoir model	
Institute of Atmospheric Physics, Chinese Academy of Sciences	VIC	Liang *et al.* (1994) Liang & Xi (2001)	Multi-layer soil and energy balance	One parameter simple routing

NOAH, SWB) while the remaining ran hourly simulations. The model structures are presented briefly in this issue (see catalogue of models). It should be noted that although the production stores and transfer functions vary in complexity and nature, all the production functions use reservoir analogies.

Some models needed information that was not provided in the database (Hydrotel) while others did not use any of the morphometric data provided (GR4J; GR5H; HBV0, etc.). The participants were free to seek for any additional data through their own means.

Ideally, the comparison should have been made on the whole data set of 40 catchments. However, in order to encourage participation, the contributors were given the possibility to run their models on either 3, 12 or 40 catchment samples (no research funds are provided by MOPEX). As can be seen in Table 1, the two distributed models and the SAC-SMA model at the hourly time step were run on the 3-catchment (3C) sample. GR4J, GR5H, HBV0, IHAC, Mordor and Topmo were tested on the 40-

catchment (40C) sample but only in the gauged mode. Whereas NOAH, SAC-SMA and SWB were tested on the 12C sample at the daily time step. VIC was the only model tested on the 12C sample at the hourly time step. However, there are disparities in the original submissions, which render the comparisons more difficult. One of the participating teams replaced one of the catchments from the 12C set with another from the 40C. Thus, in order to present consistent results, we carried out the model inter-comparison analysis on the “common” 11 catchments(11C). This slight problem could easily be overcome in the future by providing clearer data submission formats and guidelines.

Hydro-meteorological data consisted of hourly (and daily) measurements of discharge and hourly estimates of rainfall, evapotranspiration, downward solar and infrared radiation, specific air humidity, air temperature and wind speed. For more details on the database, see Chahinian *et al* (this issue).

The participants were free to chose the calibration and validation criteria of their choice. However, they were informed that the model comparison would be carried out on a number of criteria based on the Nash and Sutcliffe efficiency, the bias and the root mean square error. In order to evaluate the importance of low flows the NSE and RMSE criteria were also computed on the square root of the discharge. The RMSE on the log of the discharge was also computed.

$$NS_Q = 1 - \frac{\sum_{t=1}^{N}\left(Q_{obs,t} - Q_{sim,t}\right)^2}{\sum_{t=1}^{N}\left(Q_{obs,t} - \overline{Q}_{obs,t}\right)^2} \tag{1}$$

$$NS_{\sqrt{Q}} = 1 - \frac{\sum_{t-1}^{N}\left(\sqrt{Q_{obs,t}} - \sqrt{Q_{obs,t}}\right)^2}{\sum_{t-1}^{N}\left(\sqrt{Q_{obs,t}} - \overline{\sqrt{Q_{obs,t}}}\right)^2} \tag{2}$$

$$RMSE_Q = \sqrt{\frac{1}{N}\left(\sum_{t=1}^{N}\left(Q_{sim,t} - Q_{obs,t}\right)^2\right)} \tag{3}$$

$$RMSE_{\sqrt{Q}} = \sqrt{\frac{1}{N}\left(\sum_{t=1}^{N}\left(\sqrt{Q_{sim,t}} - \sqrt{Q_{obs,t}}\right)^2\right)} \tag{4}$$

$$RMSE_{\ln Q} = \sqrt{\frac{1}{N}\left[\sum_{t=1}^{N}\ln\left(Q_{sim,t} + \overline{Q}_{obs,t}/10\right) - \ln\left(Q_{obs,t} + \overline{Q}_{obs,t}/10\right)\right]^2} \tag{5}$$

$$Bias = \left[\frac{\sum_{t=1}^{N}\left(Q_{sim,t-}Q_{obs,t}\right)}{\sum_{t=1}^{N}Q_{obs,t}}\right]*100 \tag{6}$$

where *N*, number of time steps used; *t*, time step index; *Q*, streamflow (*sim* and *obs* refer to simulated and observed flows); $\overline{Q}$, mean annual discharge.

High flows and floods will influence the criteria calculated on the discharge values, whereas those calculated on the ln(Q) will be more influenced by low flows. The criteria calculated on $\sqrt{Q}$ will represent all the discharge values, giving equal weight to both high and low flows. To include the greatest number of models, in the analysis the criteria were computed both at the hourly and daily time steps. However, given the redundancy in the results, only the daily time step will be presented and discussed in this paper.

This analysis is based on the digital files submitted by the participants from June 2004 until March 2005 only. As some modellers chose to improve their results by carrying additional work after the deadline, some discrepancies might occur between their results (companion papers, this issue) and the ones presented herein.

No analysis will be carried out on the evolution of parameter values between the ungauged and gauged modes as companion papers in this issue cover this aspect.

RESULTS AND DISCUSSION

The results are organized by mode and participation level. First, model results will be analysed in the gauged mode for increasing participation levels and then in the ungauged mode.

Table 2 Participation level.

Author	Model	Participation level	Gauged Mode	Ungauged Mode	Parameterization method
Fortin	Hydrotel	3 Catchments	X	X	Pedotransfer+ vegetation classification
Hogue *et al.*,	SAC-SMA	3 Catchments	X	X	Sister catchment
Montanari & Moretti	Affdeff	3 Catchments	X	X	Single flood event duration < 600 hrs
Moussa (*)	ModSpa*	3 Catchments	X	X	Pedotransfer+ vegetation classification
Xie	VIC	12 Catchments	X	X	Vegetation classification
Schaake & Duan	SAC-SMA*	12 Catchments	X	X	Ensemble simulations
	NOAH-LSM*	12 Catchments	X	X	
	SWB*	12 Catchments	X	X	
Mathevet *et al.*	GR4J	40 Catchments	X		
	GR5H	40 Catchments	X		
	HBV0	40 Catchments	X		
	IHAC	40 Catchments	X		
	Mordor	40 Catchments	X		
	Topmo	40 Catchments	X		

(*) Daily time step.

Table 3 Main hydrological characteristics of the selected catchments based on the “Banque Hydro” database.

Participation level	Code	Name	Area (km^2)	Instantaneous peak discharge (m^3s^{-1})	Mean annual discharge (m^3s^{-1})
3C	J3024010	Le Guillec à Trézilidé	43.0	12.40	0.67
	V6035010	Le Toulourenc à Malaucène	150.0	81.00	1.32
	Y5615030	Le Loup à Villeneuve-Loubet	279.0	228	4.47
12C	A1522020	La Lauch à Guebwiller	68.1	41	1.65
	H2001020	L’Yonne à Corancy	98.0	45.90	2.89
	H3613020	Le Lunain à Épisy	252.0	12.10	0.74
	H5723011	L’Orgeval à Boissy-le-Châtel	104.0	33.10	0.63
	J2034010	Le Guindy à Plouguiel	125.0	27.60	1.21
	J4124420	La Rivière de Pont-l'Abbé à Plonéour-Lanvern	32.1	4.52	0.53
	K0744010	L’Anzon à Débats-Rivière-d'Orpra	181.0	72.30	2.54
	K0753210	Le Lignon du Forez à Boën	371.0	285	5.70
	Y3514020	Le Vistre à Bernis	291.0	43.30	2.11
40C	A5723010	L’Ingressin à Toul	54.7	9.930	0.44
	H2513110	Le Tholon à Champvallon	131.0	17.90	0.86
	H3613010	Le Lunain à Paley	163.0	18.30	0.56
	H3923010	Le ru d'Ancoeur à Blandy	181.0	23.90	0.59
	H4252010	L’Orge à Morsang-sur-Orge	922.0	41.20	3.96
	H7853010	Le Sausseron à Nesles-la-Vallée	101.0	3.320	0.56
	H7913030	La Mauldre à Aulnay-sur-Mauldre	369.0	28.50	2.15
	J4712010	L’Éllé au Faouët	142.0	59.20	2.75
	K0100020	La Loire à Goudet	432.0	1600	5.70
	K0253020	La Borne occidentale à Espaly-Saint-Marcel	375.0	261	3.66
	K0550010	La Loire à Bas-en-Basset	3234.0	3500	38.60
	K0614010	Le Furan à Andrézieux-Bouthéon	178.0	142	2.54
	K0813020	L’Aix à Saint-Germain-Laval	193.0	195	3.02
	K0974010	Le Gand à Neaux	85.0	58.40	0.90
	K1173210	L’Arconce à Montceaux-l'Étoile	599.0	147	5.77
	K2724210	L’Artière à Clermont-Ferrand	49.0	8.660	0.27
	K2783010	La Morge à Maringues	713.0	103	4.29
	K5623010	L’Auron au Pondy	199.0	29.80	0.98
	K5653010	L’Auron à Bourges	585.0	83.80	3.77
	P3245010	Le Mayne à Saint-Cyr-la-Roche	49.0	22.90	0.70
	U4305410	La Denante à Davayé	11.1	8.300	0.13
	U4525210	Le Morgon à Villefranche-sur-Saône	68.0	17.90	0.49
	V3315010	La Valencize à Chavanay	36.0	17.30	0.36
	V3517010	Le Ternay à Savas	25.5	16.00	0.34
	V6052010	L’Ouvèze à Vaison-la-Romaine	585.0	1000	6.07
	X2414030	L’Artuby à la Bastide	91.0	104	1.04
	Y5615010	Le Loup à Tourrettes-sur-Loup	206.0	147	3.67
	Y5625020	La Cagne à Cagnes-sur-Mer	95.0	160	0.82

Validation mode

All 14 models were tested on the 3 catchments in validation mode, 10 were further tested on the 11 catchments and 6 on the complete set of 40 catchments.

Validation mode 3C level

The results for the validation period of the gauged mode are presented in Table 4(a)–(e). All the models produce high NS values for the first and last catchments. The lowest ranking model is SAC-SMA calibrated by Hogue *et al.*, with a NS criterion of 67%, and the highest is SWB with 87%. With the exception of GR4J, Topmo and Mordor, the results are poorer for the second catchment. In some instances the difference in performance ranges 40–50% (Afdeff, NOAH, SAC-SMA and VIC). It is interesting to note that results of Schaake & Duan on the SAC-SMA model are overall more stable than those of Hogue *et al*. As the model structure and data are identical, the difference in performance is likely to be caused by the differences in the calibration procedure.

The NS values obtained on the third catchment are of the same order of magnitude as the first one except for Afdeff and HBV0. However, the two catchments are in

Table 4 Daily Nash and Sutcliff, *Bias* and RMSE on ln(Q) $\sqrt{Q}$ and *Q* criteria per model and per catchment for the gauged mode (validation period; 01 August 1999–31 July 2002; daily time step).

Model (a)	Nash and Sutcliffe efficiency (%) J3024010	V6035010	Y5615030
Affdeff	73	10	43
GR4J	75	77	79
GR5H	82	76	81
HBV0	76.5	66	53
Hydrotel	83	82	NA
IHAC	74	70	76
ModSpa	82	77	84
Mordor	84	88	88
NOAH–LSM	72	28	78
SAC–SMA[1]	67	17	81
SAC–SMA	85	82	88
SWB	87	75	82.5
Topmo	70	76	84
VIC	79	38	84
(b)	*Bias* (%)		
Affdeff	–23.71	56.31	56.46
GR4J	–13.37	–6.63	–5.14
GR5H	–9.18	–7.64	–3.70
HBV0	–6.69	5.07	–1.43
Hydrotel	–0.11	1.06	NA
IHAC	–9.78	–2.71	–3.61
ModSpa	–8.41	–6.62	7.61
Mordor	–14.44	–1.37	–0.98
NOAH–LSM	–14.95	50.43	0.74
SAC–SMA[1]	–7.75	–0.25	0.01
SAC–SMA	–11.02	12.52	1.35
SWB	–13.13	4.95	–6.76
Topmo	13.30	6.12	–2.13
VIC	22.68	87.45	16.44

Table 4 (*cont.*)

(c)	RMSE $ln(Q)$ (mm day^{-1})		
Affdeff	0.16	0.87	0.41
GR4J	0.25	0.59	0.38
GR5H	0.26	0.45	0.50
HBV0	0.28	0.60	0.42
Hydrotel	0.14	0.20	NA
IHAC	0.30	0.56	0.41
ModSpa	0.09	0.44	0.18
Mordor	0.24	0.43	0.37
NOAH–LSM	0.41	0.55	0.27
SAC–SMA[1]	0.91	1.46	1.43
SAC–SMA	0.05	0.27	0.14
SWB	0.06	0.64	0.23
Topmo	0.30	0.49	0.40
VIC	0.10	0.92	0.28
(d)	RMSE $\sqrt{(Q)}$ (mm day^{-1})		
Affdeff	0.27	0.50	0.58
GR4J	*0.17*	0.24	0.25
GR5H	0.15	0.23	0.24
HBV0	0.17	0.24	0.28
Hydrotel	0.22	0.22	NA
IHAC	0.19	0.25	0.24
ModSpa	0.21	0.32	0.22
Mordor	0.15	0.20	0.21
NOAH–LSM	0.35	0.42	0.35
SAC–SMA[1]	0.21	0.35	0.30
SAC–SMA	0.11	0.25	0.25
SWB	0.17	0.36	0.32
Topmo	0.24	0.20	0.20
VIC	0.24	0.49	0.34
(e)	RMSE(Q) (mm day^{-1})		
Affdeff	0.84	1.52	2.35
GR4J	*0.69*	0.63	1.12
GR5H	0.55	0.60	1.03
HBV0	0.66	0.75	1.78
Hydrotel	0.66	0.72	NA
IHAC	0.71	0.71	1.26
ModSpa	0.68	0.80	0.75
Mordor	0.60	0.50	0.99
NOAH–LSM	0.86	1.36	1.55
SAC–SMA[1]	0.91	0.68	1.43
SAC–SMA	0.62	0.81	1.13
SWB	0.59	0.61	1.38
Topmo	0.75	0.61	1.01
VIC	0.45	1.27	1.30

NA, not available.

different climatic zones and are not prone to the same overland flow processes, i.e. Dunn type mechanism for the first one and Hortonian overland flow for the second and third catchments. Therefore, model performance cannot be related to the process type, nor can it be linked to the model structure.

A comparison between NS_Q and $NS_{\sqrt{Q}}$ indicates that model ranking is not affected by high flows (Table 4(e)). Table 4(b) further indicates that the bias values are higher than the NS values;. $|Bias| \leq 10\%$ for six models for the first catchment and 10 for the second and third catchments. Afdeff and VIC produce the highest bias values. In the case of the latter, the associated NS criteria are high. This indicates that the model can accurately reproduce the response time of the catchments but tends to overestimate the peak discharges, i.e. the rising and falling limbs of the simulated and observed hydrographs are well aligned, but the peaks are over estimated (Fig. 1). In comparison Table 4(c) indicates that low flows and recession curves are well reproduced.

In conclusion, based on the 3C sample, SAC-SMA (Schaake & Duan) has the best performance, followed by Hydrotel and Mordor. Afdeff and SAC-SMA (Hogue *et al.*) have the poorest performance. At this stage, with the exception of SAC-SMA, we cannot establish whether the difference is due to the model structure or the calibration strategy and effort. Indeed, some models were calibrated by trial and error (Afdeff, Modspa) while others used automatic search algorithms (GR5H, SAC-SMA, NOAH); some authors specified the time spent on the calibration process while others did not.

No clear difference can be seen between the distributed and lumped models. However, in comparison with Hydrotel and Modspa, Afdeff seems to have a poorer performance. This can be explained by the difference between the model's grid size and the coarse data resolution (for further details see Fortin *et al.*, this issue).

Validation mode 11C level

The validation results of the 11C models are summarized in Figures 2 and 3. They represent the cumulative Nash & Sutcliffe efficiencies on respectively discharge and its square root value. Figure 3 indicates that Noah and VIC yield the poorest results: 50% of the catchments are simulated with a Nash & Sutcliffe efficiency <30% and < 60% respectively. In comparison GR5H, SAC-SMA and Mordor have the best performances as 50% of the catchments have Nash and Sutcliffe efficiencies >75%. The five models calibrated by Mathevet *et al.*, exhibit very similar performances over the validation period, highlighting the robustness of their simple calibration procedure (Mathevet, 2005). The rankings, with regard to low flows, show that NOAH and VIC perform comparatively less well than the other models, while GR5H and Mordor yield the best results, followed closely by SAC-SMA.

With regard to the bias criterion (Table 5(b)), NOAH has the overall highest bias values for seven of the catchments with VIC having the other four. No clear ranking can be seen for the lowest bias values (i.e. the best model with regard to bias) as with the exception of GR5H and NOAH, all the models rank first on at least one catchment. In addition, GR4H, GR5H and Mordor tend to underestimate discharge values for all but one catchment (K0753210). The latter's flows are overestimated by all the models. This is not surprising as it has the highest mean annual and instantaneous peak discharge values of the sample (Table 3).

The model rankings regarding $RMSE_Q$, $RMSE_{\sqrt{Q}}$, $RMSE_{\ln Q}$ are similar to those of the NS_Q and $NS_{\sqrt{Q}}$ (Table 5). Mordor, GR5H and SAC-SMA rank in the top three with

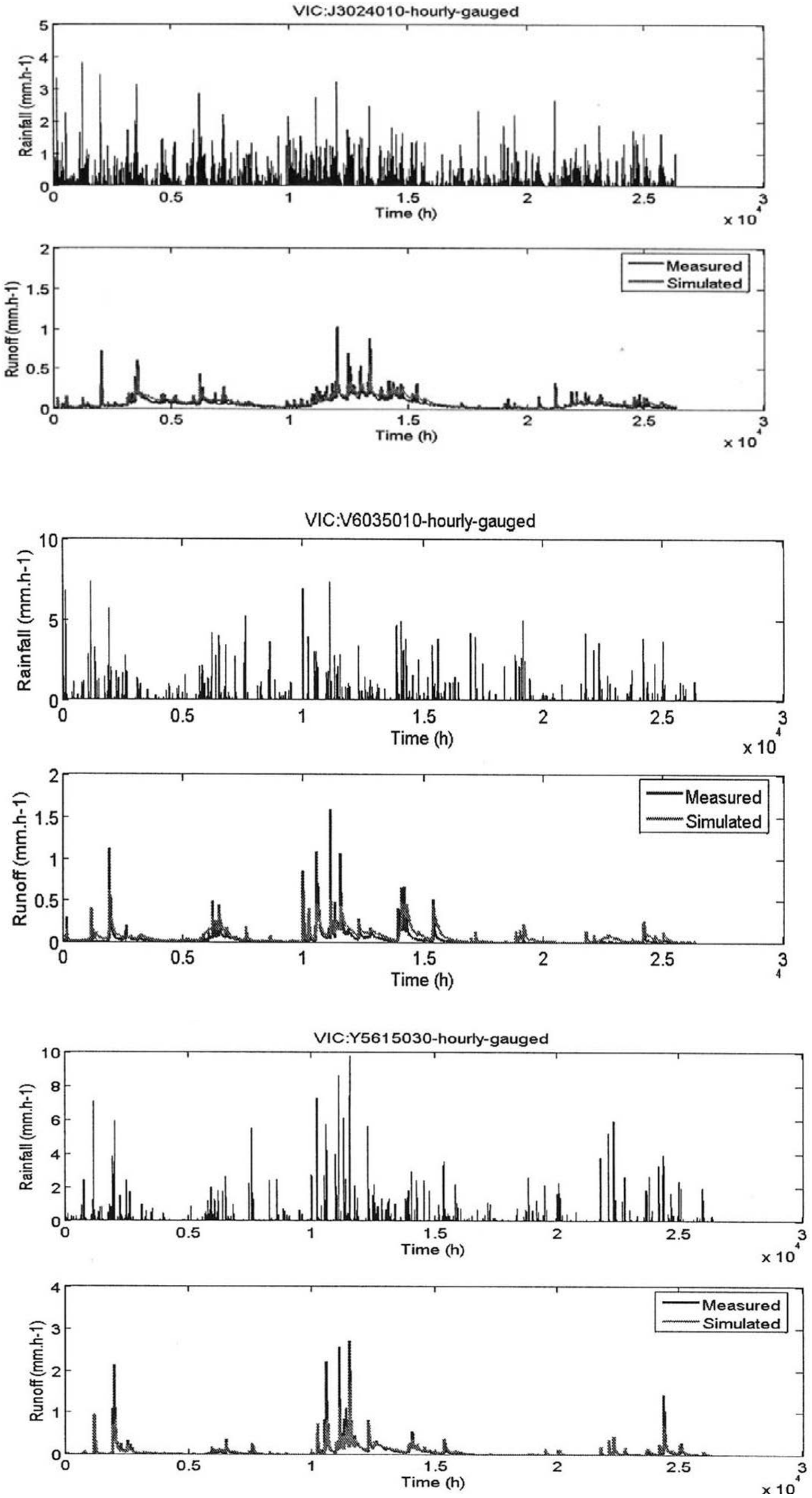


Fig. 1 Simulated and observed hydrographs for VIC model.

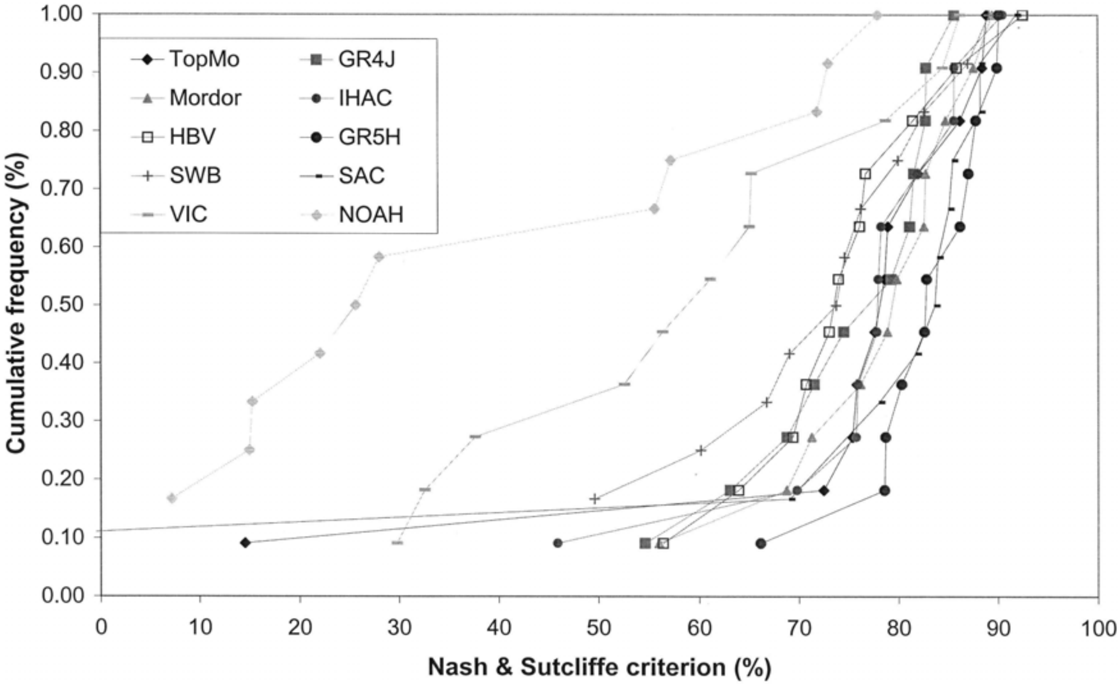


Fig. 2 Nash and Sutcliffe criteria distribution for the GR4J, GR5H, HBV0, IHAC, Mordor NOAH, SAC-SMA, SWB, Topmo and VIC (validation).

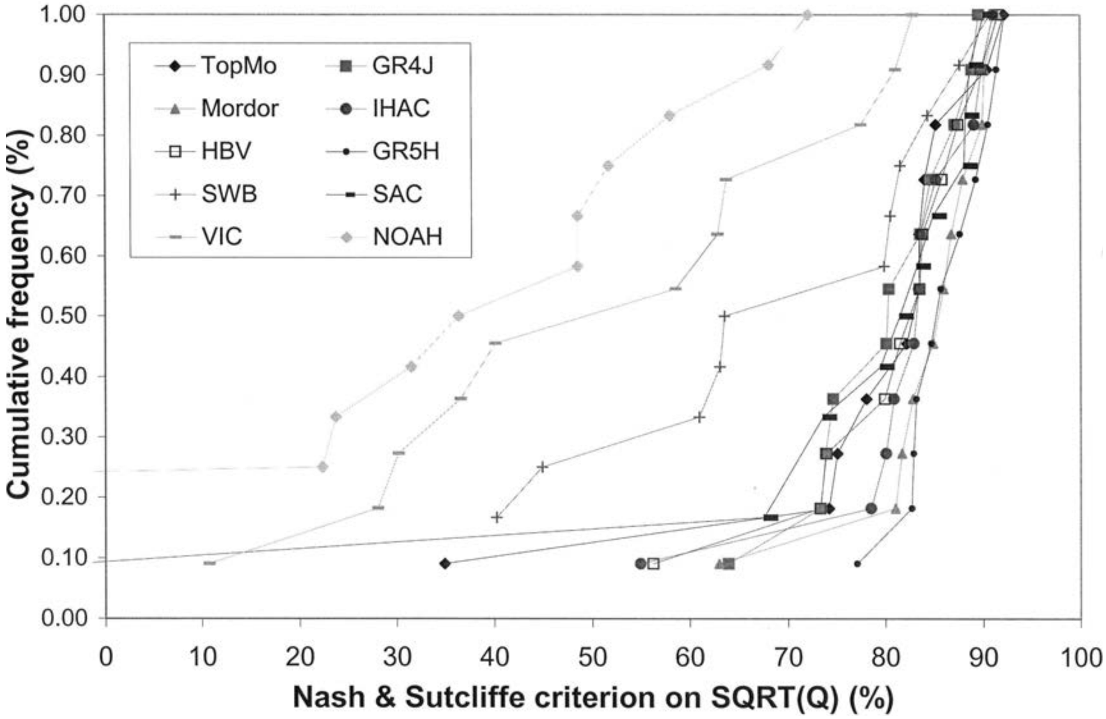


Fig. 3 Nash and Sutcliffe on SQRT (Q) criteria distribution for the GR4J, GR5H, HBV0, IHAC, Mordor NOAH, SAC-SMA, SWB, Topmo and VIC (validation).

Mordor having the best performance. NOAH ranks last once again. Table 6 summarizes the overall model performance for all 11 catchments and six criteria.

Table 5 *Bias*, RMSE on *Q*, $\sqrt{Q}$ and ln(*Q*) values per model and catchment for the gauged mode (validation period) for the 11 catchment sample.

Catchment	GR4H	GR5H	HBV	IHAC	Mordor	NOAH	SAC–SMA	SWB	Topmo	VIC
(a) *Bias* (%)										
a1522020	–5.61	–2.08	29.25	0.63	–1.98	67.91	35.42	5.65	29.44	22.68
h2001020	–4.86	–4.29	–11.07	–3.79	–3.14	14.07	48.21	–11.45	–10.00	–6.52
h3613020	–19.64	–8.40	11.74	–25.40	–20.84	50.50	–8.40	21.03	26.50	–4.91
j2034010	–14.07	–9.22	3.21	–9.97	–11.61	–8.54	4.07	–5.35	4.07	30.58
j3024010	–13.38	–9.19	–6.70	–9.79	–12.44	16.73	–11.02	–13.13	–13.31	6.46
j4124420	–13.39	–6.94	5.18	–5.70	–10.23	26.30	3.25	–19.26	5.16	16.44
k0744010	–5.67	–8.68	4.31	–3.84	–3.07	63.44	7.54	2.50	3.63	34.54
k0753210	0.03	2.18	0.33	3.10	2.08	51.88	5.88	9.70	2.68	35.51
v6035010	–6.66	–7.67	5.07	–2.74	–1.38	46.28	12.52	4.95	6.12	65.39
y3514020	–5.46	–5.02	–4.01	–5.48	–4.63	4.61	–4.19	–2.80	–2.36	70.01
y5615030	–5.15	–3.70	–1.43	3.62	–0.99	1.25	1.35	–6.76	–2.13	87.45
(b) RMSE (*Q*) (mm day^{-1})										
a1522020	1.23	1.10	1.60	1.38	1.10	2.81	1.37	1.31	1.46	2.01
h2001020	1.39	1.21	1.45	1.28	1.09	1.21	0.94	1.30	1.37	0.87
h3613020	0.18	0.12	0.18	0.20	0.17	1.66	0.32	1.28	0.25	0.23
j2034010	0.43	0.33	0.39	0.40	0.35	0.72	0.97	0.36	0.36	0.92
j3024010	0.69	0.55	0.66	0.71	0.60	0.86	0.62	0.59	0.75	0.75
j4124420	0.73	0.54	0.47	0.53	0.51	1.27	0.54	0.93	0.79	1.26
k0744010	0.79	0.76	0.66	0.71	0.67	1.06	0.68	0.71	0.61	0.76
k0753210	0.71	0.59	0.62	0.60	0.58	1.00	0.46	0.63	0.58	0.67
v6035010	0.63	0.60	0.75	0.71	0.50	1.36	0.68	0.81	0.61	1.27
y3514020	0.37	0.31	0.36	0.28	0.30	0.75	0.28	0.51	0.28	0.48
y5615030	1.12	1.04	1.78	1.26	1.10	1.55	1.13	1.38	0.99	1.30
(c) RMSE on $\sqrt{Q}$ (mm day^{-1})										
a1522020	0.11	0.26	0.38	0.30	0.26	0.61	0.42	0.32	0.38	0.57
h2001020	0.25	0.25	0.28	0.25	0.23	0.34	0.25	0.29	0.31	0.28
h3613020	0.16	0.07	0.12	0.12	0.11	0.32	0.18	0.43	0.14	0.14
j2034010	0.23	0.12	0.14	0.13	0.12	0.49	0.27	0.13	0.13	0.42
j3024010	0.24	0.15	0.17	0.20	0.15	0.29	0.15	0.17	0.24	0.24
j4124420	0.29	0.16	0.16	0.18	0.16	0.35	0.21	0.27	0.24	0.51
k0744010	0.30	0.23	0.20	0.22	0.20	0.34	0.24	0.29	0.19	0.30
k0753210	0.22	0.18	0.18	0.18	0.18	0.41	0.17	0.26	0.18	0.26
v6035010	0.17	0.23	0.24	0.25	0.20	0.37	0.25	0.36	0.21	0.49
y3514020	0.25	0.13	0.16	0.14	0.12	0.42	0.15	0.26	0.13	0.20
y5615030	0.14	0.24	0.28	0.24	0.26	0.35	0.25	0.32	0.21	0.34
(d) RMSE on ln(*Q*) (mm day^{-1})										
a1522020	0.41	0.38	0.44	0.36	0.37	0.47	0.33	0.15	0.43	0.50
h2001020	0.43	0.34	0.34	0.36	0.31	0.18	0.09	0.11	0.39	0.17
h3613020	0.43	0.30	0.50	0.47	0.44	0.61	0.19	0.11	0.48	0.13
j2034010	0.31	0.34	0.39	0.36	0.39	0.55	0.19	0.71	0.33	0.59
j3024010	0.59	0.26	0.28	0.30	0.24	0.37	0.08	0.07	0.30	0.10
j4124420	0.37	0.25	0.29	0.29	0.25	0.41	0.05	0.10	0.33	0.49
k0744010	0.41	0.46	0.37	0.44	0.40	0.55	0.08	0.33	0.39	0.33
k0753210	0.31	0.31	0.31	0.29	0.30	0.35	0.18	0.22	0.31	0.19
v6035010	0.25	0.45	0.60	0.56	0.43	0.55	0.27	0.64	0.49	0.92
y3514020	0.38	0.39	0.48	0.46	0.40	0.44	0.16	0.52	0.38	0.21
y5615030	0.36	0.51	0.42	0.41	0.37	0.27	0.14	0.23	0.40	0.28

Table 6 Synthesis of overall model classification for the gauged mode (validation period) for the 11 catchment sample.

Criteria	Best Model	Worst Model
Nash(Q)	– (SAC–SMA on 4 catchments)	NOAH
Nash($\sqrt{Q}$)	– (GR5H on 5 catchments)	NOAH
Bias	–	NOAH
RMSE	Mordor	NOAH
RMSE($\sqrt{Q}$)	Mordor	NOAH
RMSE(ln(Q))	SAC_SAM	–

(–) No model scores best on at least 6 catchments.

Table 7 *Bias*, RMSE values for the gauged mode (validation period) for the 40 catchment sample (continues next page).

Catchment	GR4J	GR5H	HBV0	IHAC	Mordor	Topmo
(a) *Bias*						
A1522020	–5.61	–2.08	29.25	0.63	–1.98	29.44
A5723010	0.76	1.82	18.40	4.59	6.01	19.02
H2001020	–4.86	–4.29	–11.07	–3.79	–3.14	–10.00
H2513110	–6.96	–4.21	–4.71	–16.81	–14.03	–1.41
H3613010	–27.01	–21.25	14.33	–32.87	–27.97	24.53
H3613020	–19.64	–8.40	11.74	–25.40	–20.84	26.50
H3923010	–33.14	–29.54	19.25	–34.24	–34.61	22.66
H4252010	–0.01	5.51	8.39	–5.20	–8.25	28.23
H5723011	–37.06	–32.84	–12.26	–35.45	–28.42	–8.80
H7853010	16.10	27.58	8.96	–4.15	3.72	43.37
H7913030	1.74	14.34	8.23	–0.10	–5.79	19.99
J2034010	–14.07	–9.22	3.21	–9.97	–11.61	4.07
J3024010	–13.38	–9.19	–6.69	–9.78	–12.44	–13.30
J4124420	–13.46	–7.00	5.16	–5.66	–10.23	5.04
J4712010	–13.61	–12.85	–11.46	–9.65	–13.39	–10.15
K0100020	–5.64	–7.90	–10.81	–7.20	–5.61	62.32
K0253020	–4.94	–4.06	10.07	–7.45	–1.27	8.93
K0550010	–11.78	–12.33	–2.92	–12.51	–9.15	–2.33
K0614010	–8.23	–7.51	–7.21	–12.90	–6.34	–2.32
K0744010	–5.67	–8.68	4.31	–3.84	–3.07	3.63
K0753210	0.03	2.18	0.33	3.11	2.08	2.68
K0813020	–3.25	–8.05	6.66	–2.11	–1.50	5.02
K0974010	–18.87	–21.29	–13.32	–23.37	–19.17	–8.65
K1173210	–17.69	–18.79	0.95	–19.99	–16.48	–4.72
K2724210	–8.68	–2.23	22.45	–9.54	–5.82	21.95
K2783010	0.92	4.06	5.17	–4.53	–0.29	17.18
K5623010	–13.65	–14.55	7.07	–17.84	–15.28	6.01
K5653010	–12.17	–15.29	–1.74	–18.34	–15.38	–0.71
P3245010	–16.02	–15.78	2.50	–12.55	–12.71	–5.88
U4305410	–6.90	–7.44	–7.03	–11.52	–7.25	–9.17
U4525210	–9.43	–7.51	–3.48	–14.28	–10.43	–1.57
V3315010	–7.59	–7.18	5.98	–6.54	–5.35	3.62
V3517010	3.69	2.47	–2.11	8.07	3.67	1.17

Table 7 (*cont.*)

Catchment	GR4J	GR5H	HBV0	IHAC	Mordor	Topmo
V6035010	–6.64	–7.65	5.08	–2.71	–1.38	6.13
V6052010	17.01	11.75	12.18	9.00	16.48	13.28
X2414030	10.29	11.80	18.09	19.47	9.97	13.77
Y3514020	–5.45	–5.02	–4.00	–5.48	–4.63	–2.36
Y5615010	0.98	9.78	5.26	14.65	7.11	9.52
Y5615030	–5.15	–3.71	–1.44	3.61	–0.99	–2.13
Y5625020	–0.11	–0.17	9.33	7.43	4.94	10.96
(b) RMSE(Q)						
A1522020	1.23	1.10	1.60	1.38	1.10	1.46
A5723010	0.41	0.34	0.41	0.36	0.28	0.45
H2001020	1.39	1.21	1.45	1.28	1.09	1.37
H2513110	0.22	0.18	0.25	0.27	0.23	0.21
H3613010	0.32	0.25	0.30	0.32	0.29	0.32
H3613020	0.18	0.12	0.18	0.20	0.17	0.25
H3923010	0.60	0.49	0.48	0.58	0.57	0.42
H4252010	0.27	0.22	0.27	0.19	0.19	0.30
H5723011	1.32	1.18	1.08	1.28	1.25	0.88
H7853010	0.18	0.22	0.19	0.14	0.15	0.45
H7913030	0.26	0.23	0.28	0.20	0.23	0.34
J2034010	0.43	0.33	0.39	0.40	0.35	0.36
J3024010	0.69	0.55	0.66	0.71	0.60	0.75
J4124420	0.73	0.54	0.47	0.53	0.51	0.79
J4712010	1.29	0.90	1.15	1.16	1.19	0.83
K0100020	1.14	1.12	1.13	1.22	1.03	1.35
K0253020	0.44	0.38	0.40	0.43	0.37	0.38
K0550010	0.64	0.59	0.61	0.66	0.56	0.55
K0614010	0.75	0.69	0.69	0.69	0.64	0.62
K0744010	0.79	0.76	0.66	0.71	0.67	0.61
K0753210	0.71	0.59	0.62	0.60	0.58	0.58
K0813020	0.88	0.85	0.75	0.76	0.77	0.73
K0974010	0.84	0.81	0.76	0.88	0.80	0.66
K1173210	0.70	0.63	0.53	0.68	0.57	0.49
K2724210	0.33	0.31	0.35	0.33	0.32	0.71
K2783010	0.17	0.14	0.17	0.17	0.13	0.20
K5623010	0.56	0.48	0.41	0.59	0.46	0.33
K5653010	0.52	0.46	0.43	0.56	0.45	0.35
P3245010	0.82	0.71	0.73	0.76	0.64	0.68
U4305410	0.80	0.71	0.74	0.73	0.63	0.73
U4525210	0.42	0.37	0.43	0.44	0.38	0.44
V3315010	0.50	0.43	0.39	0.46	0.41	0.36
V3517010	0.63	0.56	0.55	0.63	0.57	0.55
V6035010	0.63	0.60	0.75	0.71	0.50	0.61
V6052010	0.89	0.81	0.92	0.99	0.84	0.84
X2414030	1.20	1.07	1.23	1.26	0.91	1.23
Y3514020	0.37	0.31	0.36	0.28	0.30	0.28
Y5615010	1.16	1.48	1.21	1.21	0.87	1.35
Y5615030	1.12	1.04	1.78	1.26	1.01	0.99
Y5625020	0.90	0.83	1.12	1.09	0.97	0.92
(c) RMSE $\sqrt{Q}$						
A1522020	0.30	0.26	0.38	0.30	0.26	0.38

Table 7 (*cont*.)

Catchment	GR4J	GR5H	HBV0	IHAC	Mordor	Topmo
A5723010	0.17	0.14	0.17	0.14	0.11	0.19
H2001020	0.29	0.25	0.28	0.25	0.23	0.31
H2513110	0.11	0.09	0.14	0.14	0.12	0.13
H3613010	0.18	0.14	0.21	0.19	0.17	0.20
H3613020	0.11	0.07	0.12	0.12	0.11	0.14
H3923010	0.25	0.20	0.23	0.24	0.22	0.18
H4252010	0.15	0.12	0.17	0.11	0.10	0.17
H5723011	0.35	0.31	0.29	0.34	0.34	0.23
H7853010	0.11	0.14	0.12	0.08	0.09	0.22
H7913030	0.12	0.11	0.14	0.09	0.10	0.17
J2034010	0.14	0.12	0.14	0.13	0.12	0.13
J3024010	0.17	0.15	0.17	0.20	0.15	0.24
J4124420	0.22	0.16	0.16	0.18	0.16	0.24
J4712010	0.25	0.19	0.21	0.23	0.21	0.18
K0100020	0.26	0.23	0.25	0.26	0.23	0.39
K0253020	0.19	0.16	0.18	0.18	0.16	0.17
K0550010	0.21	0.18	0.19	0.20	0.17	0.17
K0614010	0.24	0.23	0.23	0.23	0.21	0.22
K0744010	0.25	0.23	0.20	0.22	0.20	0.19
K0753210	0.23	0.18	0.18	0.18	0.18	0.18
K0813020	0.27	0.25	0.23	0.23	0.23	0.23
K0974010	0.31	0.29	0.25	0.32	0.28	0.24
K1173210	0.23	0.20	0.19	0.23	0.19	0.20
K2724210	0.18	0.18	0.21	0.18	0.18	0.27
K2783010	0.11	0.09	0.12	0.11	0.09	0.14
K5623010	0.22	0.18	0.17	0.22	0.17	0.15
K5653010	0.17	0.14	0.16	0.19	0.14	0.14
P3245010	0.26	0.23	0.23	0.23	0.20	0.22
U4305410	0.28	0.26	0.25	0.25	0.23	0.26
U4525210	0.16	0.14	0.18	0.17	0.15	0.18
V3315010	0.20	0.17	0.17	0.19	0.16	0.16
V3517010	0.25	0.22	0.20	0.24	0.22	0.20
V6035010	0.24	0.23	0.24	0.25	0.20	0.21
V6052010	0.28	0.24	0.27	0.30	0.25	0.25
X2414030	0.30	0.22	0.32	0.31	0.24	0.28
Y3514020	0.16	0.13	0.16	0.14	0.12	0.13
Y5615010	0.25	0.23	0.26	0.25	0.19	0.22
Y5615030	0.25	0.24	0.28	0.24	0.21	0.21
Y5625020	0.26	0.23	0.27	0.26	0.22	0.23
d) RMSE ln(Q)						
A1522020	0.41	0.38	0.44	0.36	0.37	0.43
A5723010	0.39	0.37	0.38	0.37	0.37	0.40
H2001020	0.37	0.34	0.34	0.36	0.31	0.39
H2513110	0.31	0.27	0.30	0.28	0.28	0.28
H3613010	0.55	0.52	0.74	0.58	0.54	0.62
H3613020	0.41	0.30	0.50	0.47	0.44	0.48
H3923010	0.53	0.48	0.54	0.59	0.53	0.48
H4252010	0.50	0.46	0.55	0.42	0.37	0.50
H5723011	0.52	0.44	0.49	0.52	0.62	0.48
H7853010	0.29	0.29	0.34	0.33	0.29	0.30
H7913030	0.23	0.20	0.23	0.23	0.23	0.24

Table 7 (*cont.*)

Catchment	GR4J	GR5H	HBV0	IHAC	Mordor	Topmo
J2034010	0.36	0.34	0.39	0.36	0.39	0.33
J3024010	0.25	0.26	0.28	0.30	0.24	0.30
J4124420	0.31	0.25	0.29	0.29	0.25	0.33
J4712010	0.33	0.29	0.35	0.33	0.26	0.31
K0100020	0.48	0.42	0.34	0.45	0.41	0.46
K0253020	0.43	0.40	0.45	0.47	0.41	0.43
K0550010	0.38	0.37	0.39	0.40	0.39	0.40
K0614010	0.21	0.19	0.20	0.20	0.19	0.21
K0744010	0.43	0.46	0.37	0.44	0.40	0.39
K0753210	0.31	0.31	0.31	0.29	0.30	0.31
K0813020	0.44	0.46	0.39	0.44	0.42	0.39
K0974010	0.61	0.61	0.54	0.68	0.59	0.54
K1173210	0.44	0.39	0.49	0.48	0.37	0.45
K2724210	0.54	0.55	0.63	0.51	0.49	0.58
K2783010	0.47	0.40	0.52	0.47	0.44	0.54
K5623010	0.67	0.60	0.50	0.70	0.57	0.55
K5653010	0.49	0.40	0.50	0.50	0.43	0.46
P3245010	0.42	0.37	0.44	0.38	0.39	0.40
U4305410	0.46	0.46	0.44	0.43	0.43	0.47
U4525210	0.38	0.38	0.50	0.44	0.37	0.49
V3315010	0.44	0.42	0.45	0.46	0.44	0.44
V3517010	0.44	0.41	0.37	0.45	0.42	0.38
V6035010	0.59	0.45	0.60	0.56	0.43	0.49
V6052010	0.58	0.46	0.54	0.54	0.49	0.51
X2414030	0.48	0.48	0.50	0.52	0.48	0.51
Y3514020	0.43	0.39	0.48	0.46	0.40	0.38
Y5615010	0.48	0.44	0.49	0.46	0.42	0.46
Y5615030	0.38	0.51	0.42	0.41	0.37	0.40
Y5625020	0.68	0.55	0.64	0.60	0.52	0.65

Validation mode 40C level

Although no clear ranking can be established on average regarding the NS_Q and $NS_{\sqrt{Q}}$ criteria, Mordor, GR5H and SAC-SMA perform better than the other models. NOAH seems to have the poorest performance. These findings highlight the fact that model performance in validation cannot be linked to the number of calibrated parameters as GR5H has 5 free parameters, Mordor 10 and SAC-SMA 13.

The results for the validation mode on the 40C sample are summarized in Figs 4, 5 and 6 and Table 8(a)–(d).

Analysis of Figs 4 and 5 indicates that similar to the 11C sample, Mordor and GR5H have the best overall performance with regard to the Nash and Sutcliffe criterion on both (Q) and $\sqrt{Q}$. In comparison Topmo has the worst performance. However, all five models yield Nash and Sutcliffe efficiencies >70% for 50% of the catchments, therefore their performances should be considered as good.

Figure 6 shows a general similarity in the models' behaviour regarding bias. However, for five catchments (H3613010, H3613020, H3923010, J4124420,

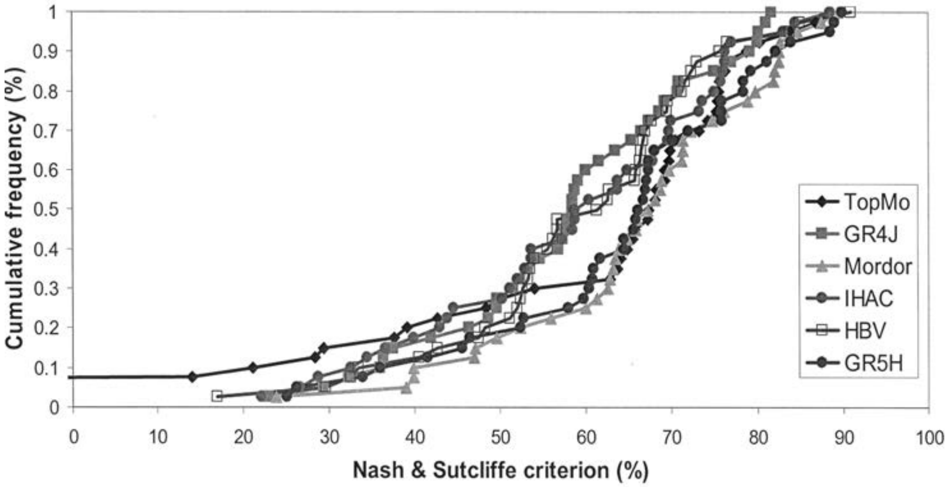


Fig. 4 Nash and Sutcliffe on (*Q*) criteria distribution for the GR4J, GR5H, HBV0, IHAC, Mordor and Topmo models on the 40 catchments in gauged mode (validation).

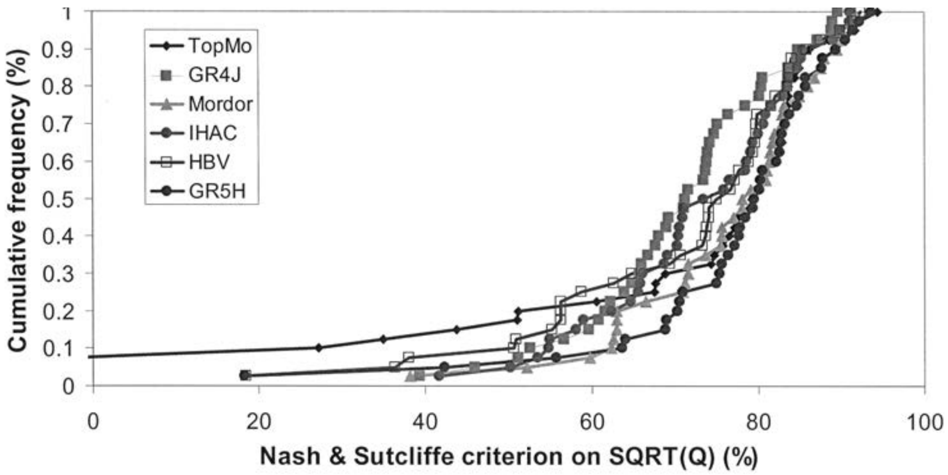


Fig. 5 Nash and Sutcliffe on SQRT(*Q*) criteria distribution for the GR4J, GR5H, HBV0, IHAC, Mordor and Topmo models on the 40 catchments in gauged mode (validation).

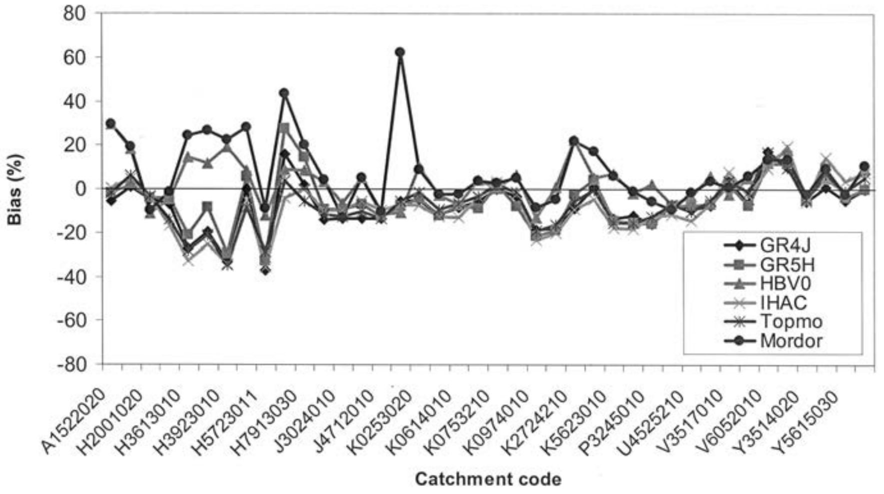


Fig. 6 Bias values for the GR4J, GR5H, HBV0, IHAC, Mordor and Topmo models on the 40 catchments in gauged mode (validation).

Table 8 Nash and Sutcliffe and Bias criteria per model and per catchment for the ungauged mode (01 August 1996–31 July 2002; daily time step).

Model	Nash and Sutcliffe efficiency (%)		
	J3024010	V6035010	Y5615030
Affdeff	76	–188	–88
GR4J	*NA*	*NA*	*NA*
GR5H	*NA*	*NA*	*NA*
HBV0	*NA*	*NA*	*NA*
Hydrotel	–1	53	50
IHAC	*NA*	*NA*	*NA*
ModSpa	65	64	69
Mordor	*NA*	*NA*	*NA*
NOAH–LSM	37	16	77
SAC–SMA^1	64	43	39
SAC–SMA	71	78	78
SWB	66	71	80
Topmo	NA	NA	NA
VIC	72	45	82
	Bias (%)		
Affdeff	–16.44	104.87	52.88
GR4J	NA	NA	NA
GR5H	NA	NA	NA
HBV0	NA	NA	NA
Hydrotel	2.11	30.50	–17.05
IHAC	NA	NA	NA
ModSpa	1.80	10.20	12.38
Mordor	NA	NA	NA
NOAH–LSM	–48.20	93.53	7.55
SAC–SMA1	–24.85	–18.92	–46.90
SAC–SMA	–11.79	10.91	–22.96
SWB	–4.88	26.84	–15.58
Topmo	NA	NA	NA
VIC	26.24	94.73	21.91
	RMSE ln(Q) (mm day^{-1})		
Affdeff	0.16	1.14	0.58
GR4J	NA	NA	NA
GR5H	NA	NA	NA
HBV0	NA	NA	NA
Hydrotel	0.44	0.30	0.59
IHAC	NA	NA	NA
ModSpa	0.11	0.58	0.34
Mordor	NA	NA	NA
NOAH–LSM	0.51	0.79	0.33
SAC–SMA1	0.35	0.55	0.52
SAC–SMA	0.21	0.30	0.14
SWB	0.47	0.42	0.29
Topmo	NA	NA	NA
VIC	0.12	1.20	0.32
	RMSE $\sqrt{Q}$ (mm day^{-1})		
Affdeff	0.24	0.63	0.61
GR4J	NA	NA	NA
GR5H	NA	NA	NA

Table 8 (*cont.*)

	J3024010	V6035010	Y5615030
HBV0	NA	NA	NA
Hydrotel	0.39	0.29	0.48
IHAC	NA	NA	NA
ModSpa	0.24	0.35	0.39
Mordor	NA	NA	NA
NOAH–LSM	0.38	0.48	0.36
SAC–SMA1	0.32	0.38	0.52
SAC–SMA	0.27	0.25	0.29
SWB	0.35	0.32	0.34
Topmo	NA	NA	NA
VIC	0.24	0.51	0.36
	RMSE(*Q*) (mm day^{-1})		
Affdeff	0.66	2.59	4.16
GR4J	NA	NA	NA
GR5H	NA	NA	NA
HBV0	NA	NA	NA
Hydrotel	1.34	1.15	2.25
IHAC	NA	NA	NA
ModSpa	0.80	0.92	1.68
Mordor	NA	NA	NA
NOAH–LSM	1.07	1.40	1.46
SAC–SMA1	0.81	1.15	2.37
SAC–SMA	0.73	0.72	1.43
SWB	0.79	0.82	1.34
Topmo	NA	NA	NA
VIC	0.72	1.14	1.28

NA, not available.

K2724210) Mordor and HBV stand out from the rest of the tested models. No physical or hydrometric characteristic can be found to explain this behaviour as the database contains other catchments having similar morphometric and hydrological properties for which the models give similar trends.

With regard to RMSE on *Q*, Mordor and Topmo rank first and second, respectively. GR4H is the only model that does not rank first at all; all remaining four models rank first at least once. For RMSE on SQRT(*Q*) Mordor and GR5H rank first and second and Topmo ranks third. However, the RMSE values on both (*Q*) and SQRT(*Q*) calculated for all five models are very similar i.e. the greatest difference between the highest and the lowest value for a given catchment is 0.16 mm day^{-1}. Hence one cannot conclude on a difference in the models' performances based on these criteria. This is also the case for the RMSE values on ln(*Q*).

Ungauged mode

Eight models were used to submit simulations for the ungauged mode on the 3C database and four models were tested on the 11C sample. Table 2 synthesizes the parameterization strategies used by the various participants; the outcomes are presented in detail in the companion papers (this issue).

Ungauged mode 3C Level

Table 8 indicates that all models performed less well in the ungauged mode than in the gauged mode, all catchments and criteria considered. This finding is not surprising in itself and stresses the extreme difficulty of the exercise.

With regard to the Nash and Sutlciffe criterion on (Q) SAC-SMA, SWB and Modspa have the most stable overall behaviour as the criteria vary within a maximum bound of 10% on all three catchments. VIC seems to perform rather well in the ungauged mode; it ranks first in terms of NS on the third catchment and second on the first one.

The most unstable behaviour, in terms of Nash and RMSE variation, is reported for Afdeff. It is interesting to note that the NS value obtained using Afdeff on the first catchment (76%) is very high both with regard to the other models and to the model's own performance on the other two catchments. It is within close range of the NS criterion calculated for the ungauged mode. The model was parameterized using a single flood event of duration <600 hours, hence the good result achieved on this catchment could be linked to the quality and representativeness of the selected flood event. In contrast, the method seems poorly adapted to the two other catchments (V6035010 and Y5615030), which are subject to more Mediterranean semiarid conditions highlighted by a high temporal variability of both precipitation and runoff.

None of the selected models clearly out-performs the rest with regard to both low and high flows, respectively, RMSE on ln(Q) and $\sqrt{Q}$ as the model ranking changes according to the test catchments. However, SAC-SMA (Schaake & Duan) ranks first twice for both criteria and shows an overall more stable behaviour than the other models. The best performance in terms of bias is reported for Modspa, respectively, first twice and second once, while the worst performances are reported for Afdeff, NOAH and VIC.

Ungauged mode 11C Level

Figures 7 and 8 present the Nash and Sutcliffe criterion distributions, both using Q and SQRT(Q), for the 11C samples. The results are once again poorer than for the gauged mode, and clearly worse for NOAH than for the other three models as NOAH scores a Nash criterion <–4% on 50% of the catchments. In comparison VIC, SWB and SAC-SMA have Nash criteria that are respectively >55%, >66% and >70%. SAC-SMA's performance is the best.

The models' ranking is not modified with regard to the NS on SQRT(Q). However, NOAH's performance improves drastically as 50% of the catchments have now a Nash criterion >27%. This would mean that NOAH is more prone to errors on high flows. The difference in performance is less clear for the other three models.

Table 9 indicates that the bias values are high for all four models. Furthermore, the models' behaviour is quite similar when considering their variation per catchment,

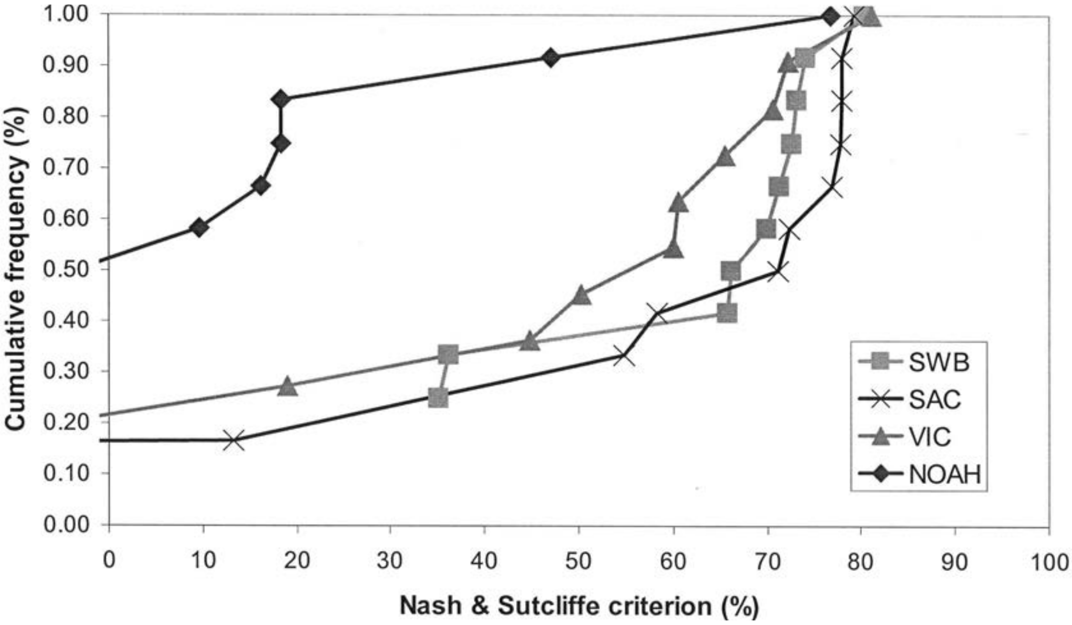


Fig. 7 Nash and Sutcliffe criteria distribution for the NOAH, SAC-SMA, SWB, and VIC models in ungauged mode.

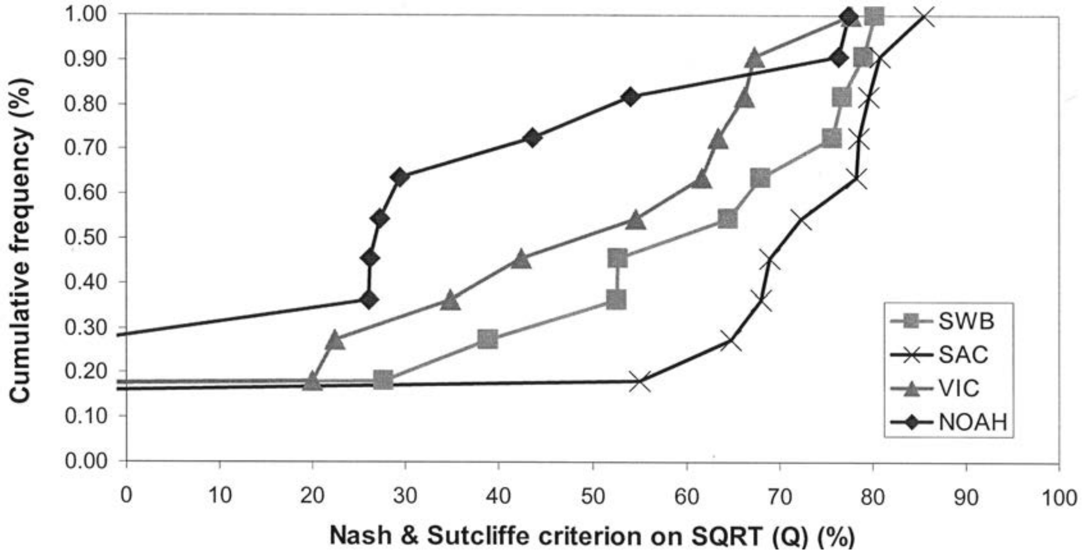


Fig. 8 Nash and Sutcliffe on SQRT(Q) criteria distribution for the NOAH, SAC-SMA, SWB, and VIC models in ungauged mode.

with the exception of catchments V6035010 and J4124420 for VIC. The bias values for the ungauged mode are higher than those of the gauged mode, when considering the same length of record (Fig. 9). This is also the case of the RMSE criteria both on Q, SQRT(Q) and ln(Q) (Fig. 10).

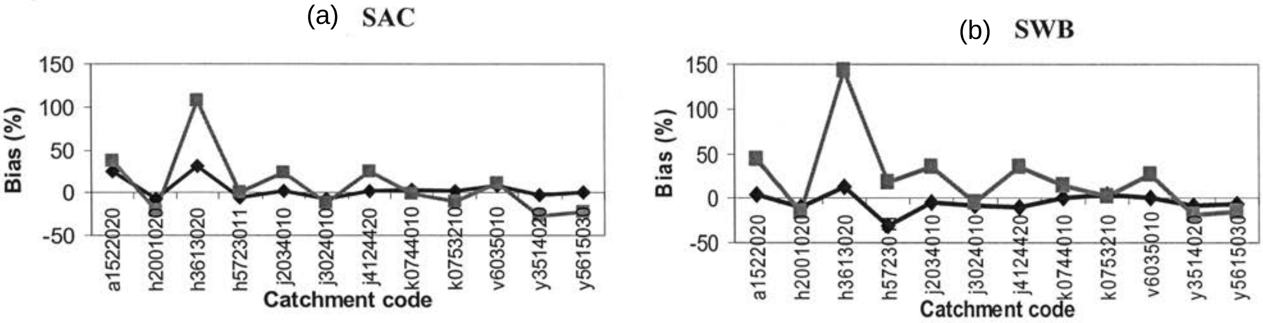
Table 10 synthesizes the models' overall performance, taking into account all the catchments and criteria. It indicates that the best results are obtained by SAC-SMA and the poorest by NOAH. It is not clear however, whether the difference is due to the models' performance or the modellers' longer experience with the model.

Table 9 *Bias*, RMSE on Q, $\sqrt{Q}$ and ln(Q) values per model and catchment for the ungauged mode for the 11 catchment sample.

Catchment (a) *Bias* (%)	NOAH	SAC–SMA	SWB	VIC
a1522020	68.39	37.19	43.71	63.08
h2001020	17.52	–21.46	–15.91	–10.59
h3613020	129.96	108.00	143.42	181.81
j2034010	38.51	1.02	18.04	77.09
j3024010	34.69	23.63	34.85	26.24
j4124420	–4.82	–11.79	–4.88	65.35
k0744010	46.11	25.28	34.58	24.72
k0753210	90.02	–0.91	13.80	19.38
v6035010	68.51	–10.82	1.65	94.73
y3514020	93.53	10.91	26.84	30.56
y5615030	19.86	–26.78	–18.87	21.91
(b) RMSE (Q) (mm day^{-1})				
a1522020	4.11	1.80	1.62	1.96
h2001020	1.78	1.29	1.27	1.18
h3613020	0.84	0.68	0.81	0.78
j2034010	1.37	0.91	1.10	1.01
j3024010	0.85	0.63	0.75	0.72
j4124420	1.07	0.73	0.79	1.40
k0744010	1.79	1.42	1.23	0.78
k0753210	1.29	0.58	0.70	0.73
v6035010	1.25	0.62	0.67	1.14
y3514020	1.40	0.72	0.82	0.60
y5615030	0.95	0.44	0.48	1.28
(c) RMSE $\sqrt{Q}$ (mm day^{-1})				
a1522020	0.73	0.40	0.43	0.55
h2001020	0.43	0.33	0.31	0.29
h3613020	0.41	0.32	0.42	0.44
j2034010	0.48	0.27	0.37	0.37
j3024010	0.31	0.19	0.28	0.24
j4124420	0.38	0.27	0.35	0.45
k0744010	0.47	0.36	0.37	0.31
k0753210	0.53	0.21	0.23	0.26
v6035010	0.47	0.21	0.21	0.51
y3514020	0.48	0.25	0.32	0.23
y5615030	0.31	0.21	0.32	0.36
(d) RMSE ln(Q) (mm day^{-1})				
a1522020	0.53	0.19	0.27	0.45
h2001020	0.23	0.16	0.17	0.13
h3613020	0.77	0.46	0.91	1.17
j2034010	0.78	0.31	0.49	0.41
j3024010	0.38	0.08	0.28	0.12
j4124420	0.51	0.21	0.47	0.35
k0744010	0.35	0.17	0.25	0.38
k0753210	0.94	0.16	0.15	0.18
v6035010	0.58	0.12	0.11	1.20
y3514020	0.79	0.30	0.42	0.28
y5615030	0.59	0.33	0.93	0.32

Table 10 Synthesis of overall model classification for the ungauged mode for the 11 catchment sample.

Criteria	Best Model	Worst Model
Nash(Q)	SAC-SMA	NOAH
Nash($\sqrt{Q}$)	SAC-SMA	NOAH
Bias	SAC-SMA	NOAH
RMSE	SAC-SMA	NOAH
RMSE($\sqrt{Q}$)	SAC-SMA	NOAH
RMSE(ln(*Q*))	SAC_SAM	NOAH



(c) **NOAH**

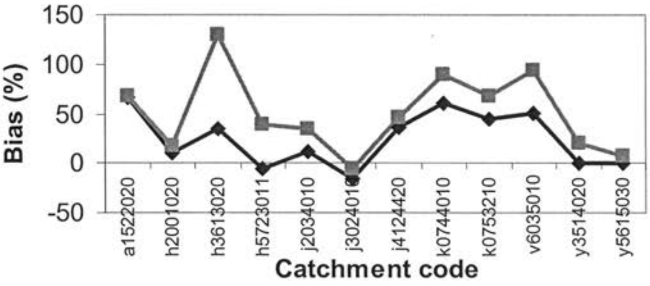


Fig. 9 Comparison between the bias values of the ungauged and gauged modes for NOAH, SAC-SMA and SWB.

The failure of the models in the ungauged mode highlights the need to investigate new parameterization strategies and methods in order to find alternatives for model calibration. Currently three mainstream strategies are followed by the MOPEX community to estimate model parameters on ungauged catchments. The first believes in the physical meaning of the model parameters and hence tries to infer these from soil and vegetation data (Modspa, VIC, Hydrotel, …). In this instance the results depend directly on the quality and spatial resolution of the data, two conditions that are not easily met in operational hydrology. The second group tries to investigate new methods to transfer parameter values from other model applications (see paper on the sister catchment approach by Hogue *et al.*, and papers on ensemble simulations by Schaake *et al.*, and Andréassian *et al.*, this issue) of course these methods then need to

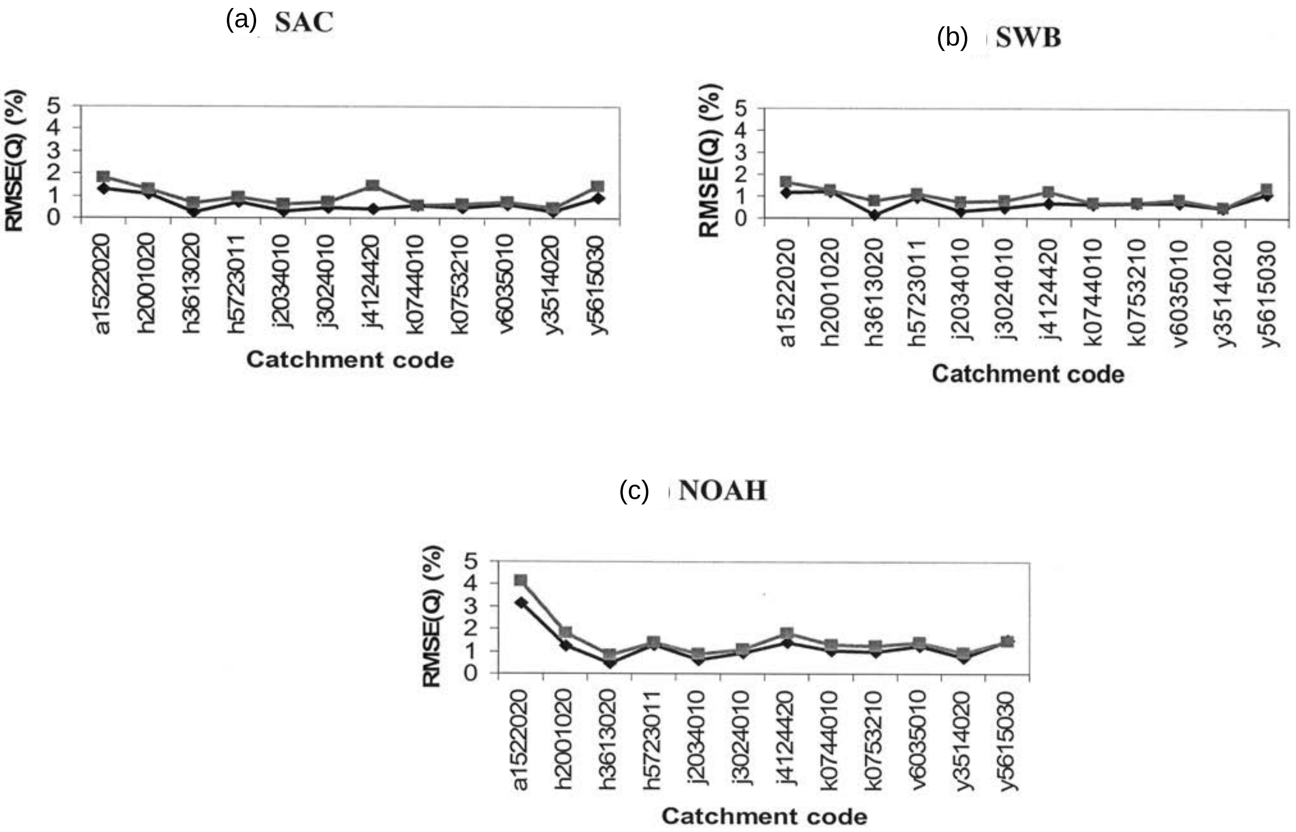


Fig. 10 Comparison between the RMSE values of the ungauged and gauged modes for NOAH, SAC-SMA and SWB.

determine a similarity index to quantify the resemblance between catchments. The third alternative consists of calibrating the model on fewer data points such as a single flood event and using the parameters for the rest of the time series (see papers by Fortin *et al.*, and Rojas-Serna *et al.*, this issue). This approach assumes a temporal stability of the models' parameters and still needs some minimal hydrological data to ensure model calibration.

In light of the simulation files that were submitted it is not possible to choose a single method as being the most efficient. Further tests should be undertaken to investigate these issues. The participation level and enthusiasm generated by the MOPEX 2004 workshop is a clear indication of the necessity and collective will to pursue these efforts both as individual scientists and as a research group. Further collaboration with the distributed modellers' community, through programmes such as DMIP, could provide an enriching forum in this sense.

CONCLUSION AND PERSPECTIVES

As part of the MOPEX 2004 workshop, the participants were asked to submit simulations using the database provided by Météo-France and Cemagref (see Chahinian *et al.*, this issue). For the first time since the launch of the MOPEX programme, the data set contained hourly hydrometeorological data. Fourteen models were tested on 3 catchments, 10 on 11 catchments and 6 on 40 catchments in gauged mode. The participation was slightly lower for the ungauged mode: 8 models were tested on 3 catchments and 4 on 11 catchments.

The objective of this paper was to analyse and compare the results produced by various models in order to see the advantages and shortcomings of their parameter estimation strategies. The results indicate that all the models succeeded in yielding acceptable results when calibrated using discharge data. The differences in calibration critera did not alter these results, although some models had high bias values despite good Nash & Sutcliffe efficiencies, this may be related to the individual effort modellers put in to the calibration process, although the latter is difficult to quantify. However, all the models' performance dropped in the ungauged mode, all criteria and catchments considered. This highlights a need to improve alternative parameter estimation techniques that do not rely heavily on calibration. Interesting approaches based on similarity indices between catchments, ensemble simulations and limited discharge data mining were presented. In order to achieve good results these techniques need large data samples to establish statistically significant relationships. Hence more experiments should be carried out on a sample of a large number of catchments, assuming that data providers are willing to share their data. Of course modellers will also have to agree to put their models through "blind tests". With its open access databases and experience in knowledge transfer, MOPEX can be an important forum between modellers, experimentalists, and water agencies. Its growing interest in predictions in ungauged catchments renders it an essential part of the IAHS's PUB decade.

Acknowledgements The MOPEX workshop organization funds were provided by the Cemagref's International Relations' Department. The authors would like to thank Météo-France, Cemagref and the National Weather Service for providing the comparison data and ENGREF for hosting the 2004 workshop.

REFERENCES

Abbott, M., Bathurst, J., Cunge, J., O'Connel, P. & Rasmussen, J. (1986) An introduction to the European Hydrological System-Système Hydrologique Européen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system. *J. Hydrol.* **87**, 45–59.

Bergström, S. & Forsman, A. (1973) Development of a conceptual deterministic rainfall–runoff model. *Nordic Hydrol.* **4**, 147–170.

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

Beven, K. & Kirkby, M. (1978) A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci.* **24**(1), 43–69.

Brath, A., Montanari, A & Moretti, G. (2002). On the use of simulation techniques for the estimation of peak river flows. In: *Proc. Int. Conf. on Flood Estimation, Berna, 6-8 March 2002, Rep II-17*), 587–599. CHR/KHR, International Commission for the Hydrology of the Rhine basin, Lelystad, The Netherlands.

Burnash, R., Ferral, R. & McGuire, R. (1973) *A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers*. Joint Federal-State River Forecast Center, Sacramento, Calicornia, USA.

Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H., Koren, V., Duan, Q., Ek, M. & Betts, A. (1996) Modeling of land--surface evaporation by four schemes and comparison with FIFE observations. *J. Geophys. Res*. **101**, 7251–7268.

Croke, B., Merritt, W. & Jakeman, A. (2004) A dynamic model for predicting hydrologic response to land cover changes in gauged and ungauged catchments. *J. Hydrol.* **291**(1–2), 115–131.

Duan, Q. & Gupta, V. (1992) Effective and efficient global optimization for conceptual rainfall–runoff models. *Water Resour. Res*. **28**(4), 1015–1031.

Edijatno (1991) Mise au point d'un modèle élémentaire pluie-débit au pas de temps journalier. PhD Thesis, University Louis Pasteur/ENGEES, Strasbourg, France.

Fortin, J. P., Moussa, R., Bocquillon, R. & Villeneuve, J. P. (1995) Hydrotel un modèle hydrologique distribué pouvant bénéficier des données fournies par la télédétection et les systèmes d'informations géographiques. *Rev. Sci. Eau* **8**(1): 97–124.

Fortin, J. P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J. & Villeneuve, J. P. (2001) Distributed watershed model compatible with remote sensing and GIS data. I: Description of model. *J. Hydrol. Engng* **6**(2), 91–99.

Gupta, H., Soorooshian, S., Hogue, T. & Boyle, D. (2003) Advances in automatic calibration of watershed models. In *Calibration of Watershed Models* (ed. by Q. Duan, H. Gupta, S. Sorooshian, A. Rousseau & R. Turcotte), 9–28. American Geophysical Union**,** Washington, DC, USA.

Haverkamp, R., Parlange, J. Y., Cuenca, R. & Ross, P. (1998) Scaling of the Richards equation and its application to watershed modelling. In *Scale Dependence and Scale Invariance in Hydrology* (ed. by G. Sposito), 190–243, Cambridge University Press, New York, USA.

Hendrickson, J., Sorooshian, S. & Brazil, L. (1988) Comparison of Newton-type and direct search algorithms for calibration of conceptual rainfall–runoff models. *Water Resour. Res*. **24**(5), 691–700.

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. *J. Hydrol.* **117**, 275–300.

Johnston, P. & Pilgrim, D. (1976) Parameter optimization for watershed models. *Water Resour. Res*. **12**(3), 477–186.

Kokkonen, T., Jakeman, A., Young, P., Koivusalo, H. (2003) Predicting daily flows in ungauged catchments: model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. *Hydrol. Processes* **17**(11), 2219–2238.

Koren, V., Smith, M. & Duan, Q. (2003) Use of a priori parameter estimates in the derivation of spatially consistent parameter sets of rainfall–runoff models In: *Calibration of Watershed Models* (ed. by Q. Duan, H. Gupta, S. Sorooshian, A. Rousseau & R. Turcotte), 239–254. American Geophysical Union**,** Washington, DC, USA.

Liang, X. & Xie, Z. (2001) A new surface runoff parameterisation with subgrid-scale soil heterogeneity for land surface models. *Adv. Water Resour*. **24**, 1173–1193.

Liang, X., Lettenmaier, D., Wood, E. & Burges, S. (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res*. **99**(D7), **14**, 415–414, 428.

Leavesley, G., Hay, L., Viger, R. & Markstorm, S. (2003) Use of *a priori* parameter-estimation methods to constrain calibration of distributed-parameter models. In: *Calibration of Watershed Models* (ed. by Q. Duan, H. Gupta, S. Sorooshian, A. Rousseau & R. Turcotte), 255–266. American Geophysical Union**,** Washington, DC, USA

Lee, H., McIntyre, N., Wheater, H. & Young, A. (2005) Selection of conceptual models for regionalization of the rainfall–runoff relationship. *J. Hydrol*. **315**(1–4), 71–92.

Mathevet T. (2005) Quels modèles pluie-débit au pas de temps horaire? Développements empiriques et intercomparaison de modèles sur un large échantillon de bassins versants. PhD Thesis, ENGREF-Cemagref, Paris, France.

Mertz, R. & Bloschel, G. (2005) Flood frequency regionalization—spatial proximity vs. catchment attributes. *J. Hydrol*. **302**(1–4), 283–306.

Micovic, J. & Quick, M. (1999) A rainfall and snowmelt runoff modelling approach to flow estimation at ungauged sites in British Columbia. *J. Hydrol*. **226**(1–2), 101–120.

Moretti, G. & Montanari, A. (2005) AFFDEF: a spatially distributed grid based rainfall–runoff model for continuous time simulations of river discharge. *Environ. Modelling & Software* (submitted).

Moussa, R. (1991) Variabilité spatio-temporelle et modélisation hydrologique. Application au bassin du Gardon d'Anduze. PhD Thesis, University of Montpellier II, France.

Moussa, R. (1993) Modélisation hydrologique spatialisée et système d'information géographique. *La Houille Blanche* **5**, 293–301.

Perrin, C. (2000) Vers une amélioration d'un modèle global pluie-débit au travers d'une approche comparative. PhD thesis, INPG (Grenoble) / Cemagref (Antony), France.

Perrin, C., Michel, C. & Andréassian, V. (2001) Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *J. Hydrol*. **242**(3–4), 275–301.

Schaake, J., Koren, V., Duan, Q., Mitchell, K. & Chen, F. (1996) Simple water balance model for estimating runoff at different spatial and temporal scales. *J. Geophys. Res*. **101**(D3), 7461–7475.

Seibert, J. (1999) Regionalization of parameters for a conceptual rainfall–runoff model. *Agr. For. Met*. **98–99**, 279–293.

Van der Linden, S. & Woo, M. (2003) Transferability of hydrological model parameters between basins in data-sparse areas, subarctic Canada. *J. Hydrol*. **270**(3–4), 182–194.