Hydrol. Earth Syst. Sci., 17, 4227–4239, 2013
www.hydrol-earth-syst-sci.net/17/4227/2013/
doi:10.5194/hess-17-4227-2013

Hydrology and Earth System Sciences
Open Access

The influence of conceptual model structure on model performance: a comparative study for 237 French catchments

W. R. van Esse[1,*], C. Perrin[2], M. J. Booij[1], D. C. M. Augustijn[1], F. Fenicia[3,4], D. Kavetski[5], and F. Lobligeois[2]

[1]University of Twente, Faculty of Engineering Technology, Department of Water Engineering and Management, Enschede, the Netherlands
[2]Irstea, Hydrosystems and Bioprocesses Research Unit (HBAN), Antony, France
[3]Public Research Centre-Gabriel Lippmann, Belvaux, Luxembourg
[4]Delft University of Technology, Faculty of Civil Engineering and Geosciences, Water Resources Section, Delft, the Netherlands
[5]University of Adelaide, School of Civil Environmental and Mining Engineering, Adelaide, Australia
[*]currently at: Nelen & Schuurmans, Utrecht, the Netherlands

Correspondence to: M. J. Booij (m.j.booij@utwente.nl)

Received: 17 April 2013 – Published in Hydrol. Earth Syst. Sci. Discuss.: 29 April 2013
Revised: 29 August 2013 – Accepted: 27 September 2013 – Published: 29 October 2013

Abstract. Models with a fixed structure are widely used in hydrological studies and operational applications. For various reasons, these models do not always perform well. As an alternative, flexible modelling approaches allow the identification and refinement of the model structure as part of the modelling process. In this study, twelve different conceptual model structures from the SUPERFLEX framework are compared with the fixed model structure GR4H, using a large set of 237 French catchments and discharge-based performance metrics. The results show that, in general, the flexible approach performs better than the fixed approach. However, the flexible approach has a higher chance of inconsistent results when calibrated on two different periods. When analysing the subset of 116 catchments where the two approaches produce consistent performance over multiple time periods, their average performance relative to each other is almost equivalent. From the point of view of developing a well-performing fixed model structure, the findings favour models with parallel reservoirs and a power function to describe the reservoir outflow. In general, conceptual hydrological models perform better on larger and/or wetter catchments than on smaller and/or drier catchments. The model structures performed poorly when there were large climatic differences between the calibration and validation periods, in catchments with flashy flows, and in catchments with unexplained variations in low flow measurements.

1 Introduction

Building accurate and computationally efficient hydrological models remains a challenging issue, despite the huge efforts by the community since the pioneering work of Linsley and Crawford (1960). The challenges in representing hydrological processes have resulted in a large variety of models and modelling approaches, ranging from lumped conceptual models to distributed physically based models (we refer the reader to the reviews by Singh and Frevert (2002a, b), Pechlivanidis et al. (2011) and others). In this study, we will focus on two types of modelling approaches, namely the *fixed* and *flexible* modelling approaches, and limit the analysis to the case of lumped conceptual models.

1.1 The fixed modelling approach

The fixed modelling approach assumes that a single model structure can be developed to apply in the majority of contexts and conditions. As examples, one may mention TOPMODEL (Beven et al., 1995), HBV (Lindström et al., 1997), Xinanjiang (Zhao and Liu, 1995), NAM (DHI, 2008) or GR4J (Perrin et al., 2003). Although these models have been continuously improved and adapted over the years, the core of their structure remained more or less similar, and it was assumed to be general enough to be applicable in a wide variety of basins. For example, Bergström (1995) provides

a detailed review of the applications of the HBV model on catchments over the five continents. Hence, end-users would take the model as it is proposed by the developers and apply it on their case study. This approach has several advantages. It saves time to the end-user who can take the model "off-the-shelf", avoiding the time-consuming process of developing a model, and go straight to the calibration step (e.g. Refsgaard et al., 2005; Clark et al., 2011a, for a description of the modelling steps). This is particularly convenient when applying the model on many catchments, as is often the case in operational conditions. A "default" general model is also useful in catchments where data may be too limited to develop a full model from scratch.

However, a weakness of the fixed modelling approach is that the processes included in the model, and their mathematical representation, may not correspond to the actual dominant processes in particular catchments of interest. This results in large structural errors and poor model performance. Several examples of fixed model failures have been described in the literature (e.g. Le Moine et al., 2008; Goswami and O'Connor, 2010; Refsgaard and Hansen, 2010). Several sources of errors may explain model failures, including errors in model structure, data, numerical errors, and mismatches between data availability and requirements. The diagnosis of these failures is often difficult without extra information (e.g. Tromp-van Meerveld and McDonnell, 2006; Krueger et al., 2010; Clark et al., 2011a; Renard et al., 2011, to mention a few). In addition, aggregate performance criteria such as the Nash–Sutcliffe efficiency index may hide internal model inconsistencies. More detailed evaluation frameworks and field observations of internal variables (e.g. soil moisture) should hence be used to more thoroughly evaluate structural limitations (e.g. Euser et al., 2013; McMillan et al., 2011).

1.2 The flexible modelling approach

The flexible modelling approach does not attempt to develop a "one-size-fits-all" model structure. Instead, in any modelling application, it calls for multiple working hypotheses to be considered (Clark et al., 2011b). We expect that, in many catchments, model structural errors represent one of the dominant sources of errors (e.g. Perrin et al., 2001; Renard et al., 2011; and others). Thus, the flexible approach offers the possibility to search among multiple structures or model components for the one that best approximates the relevant aspects of a catchment's behaviour. The main advantage of this approach is to reduce structural uncertainty, which should result in more robust model applications. Modelling approaches such as RRMT (Wagener et al., 2001), MMS (Leavesley et al., 2002), FLEX (Fenicia et al., 2008), FUSE (Clark et al., 2008) and SUPERFLEX (Fenicia et al., 2011; Kavetski and Fenicia, 2011) are examples of flexible modelling frameworks. However, the flexible approach is generally more time consuming for the end-user. It may also end up with several model structures performing equivalently, in which case an "ensemble" approach may be needed to make predictions (see e.g. Velazquez et al., 2010).

1.3 Previous model comparisons

Previous studies have already provided useful insights into the performance of different model structures. For instance, Chiew et al. (1993) indicated that relatively simple model structures can be used for larger timescales (months, years), whereas Refsgaard and Knudsen (1996) reported that models of different complexity performed equally well in their case study. In the study of Perrin et al. (2001), complex lumped conceptual models outperformed simple models in calibration but not in validation. Reed et al. (2004) found that a lumped model, used as benchmark, generally showed equivalent or better overall performance than distributed models. More recent studies of Breuer et al. (2009) and Seiller et al. (2012) showed only slight differences between models of differing complexity. In most of these intercomparisons and in hydrological modelling studies in general, there is a common aim to find an appropriate model for a particular purpose or condition, considering hydrological signature, catchment type and spatial and temporal scale (Rogers, 1978; Moussa and Bocquillon, 1996; Booij, 2003).

1.4 Scope

The main objective of this study is to investigate the impact of model structure on flow simulation for two hydrological modelling approaches: the fixed modelling approach, in which modellers would use a single predefined model structure, versus the flexible approach, in which the modeller could choose among a number of alternative model structures. This study relies on the previous work of Fenicia et al. (2011), Kavetski and Fenicia (2011) and Fenicia et al. (2013), where detailed investigations on small sets of catchments were performed. This study extends and generalizes previous work, considering a large set of 237 French catchments with diverse hydro-climatic conditions to better characterize the benefits and limitations of the two modelling approaches. Given the diversity of catchment characteristics, we also attempt to better understand differences in model performance and identify possible links between model structure, catchment characteristics and model performance.

The research questions we investigate are:

1. What is the influence of different model structures on average model performance?
2. What is the influence of different catchment characteristics on the relationship between model structure and performance?
3. What are the differences in performance between a fixed and flexible modelling approach?

Table 1. Characteristics of the 237 French catchments used in this study.

	Area [km^2]	Mean annual rainfall [mm yr^{-1}]	Mean annual PE [mm yr^{-1}]	Mean annual discharge [mm yr^{-1}]
Minimum	16	61	605	84
Average	567	988	732	383
Maximum	6836	1961	1182	1329

To address these questions, we will use SUPERFLEX as an example of a flexible modelling approach and the GR4H model as an example of a fixed modelling approach. SUPERFLEX provides a framework to construct and compare different model structures. The hourly GR4H model, and its daily version GR4J, are widely used models that have shown good average performance on many catchments in France and other countries (Perrin et al., 2003; Le Moine et al., 2007; Valéry et al., 2010; Coron et al., 2012). The GR4J and GR4H models were developed with the intention of providing the best average performance on a large array of conditions and catchments. However, structural inadequacy and/or the lack of flexibility in model structure may be one of the main reasons for its failures identified in previous studies (Perrin et al., 2003; Wagener, 2003; Le Moine, 2008; Andréassian et al., 2010; Kavetski and Fenicia, 2011). This study makes further inroads into understanding these limitations.

The paper is structured as follows. Section 2 presents the data and models used, as well as the evaluation methodology. Section 3 discusses the general results, followed by a discussion and conclusions of this work in Sects. 4 and 5 respectively.

2 Material and methods

2.1 Data

This study is based on a large data set of 237 catchments spread throughout France (Fig. 1). They represent a large variety of conditions in terms of physical characteristics (size, geology, etc.) and climate (Table 1). The catchments were selected to have limited snow influence (no catchment selected in the French Alps or Pyrenees) and limited lake influence. Precipitation, potential evapotranspiration (PE) and discharge data were available at hourly time steps from 1997 to 2006 for all catchments. Precipitation originates from the reanalysis produced by Météo-France (Tabary et al., 2012). PE data were estimated using the temperature-based formula proposed by Oudin et al. (2005), using temperature data produced by the SAFRAN reanalysis (Vidal et al., 2010). Discharge data were obtained from the French data base Banque Hydro (MEDD, 2007).

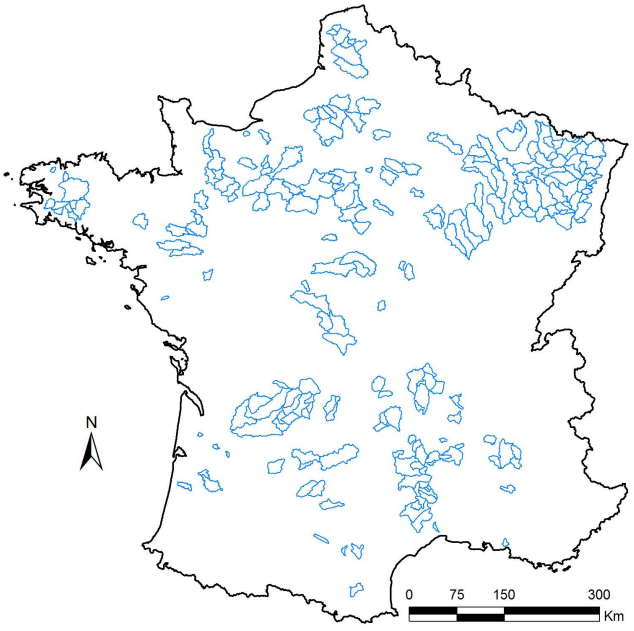

Fig. 1. Location and boundaries of the 237 French catchments used in this study (after Le Moine, 2008).

2.2 Catchment classification

To investigate the influence of different catchment characteristics on the relation between model structure and performance, the 237 catchments were classified based on three catchment characteristics: area, aridity index and the ratio of seasonal runoff coefficients in summer and winter ($RC_{S/W}$). The latter is considered a more integral property of climate and catchment characteristics and distinguishes groundwater-dominated catchments from catchments dominated by surface runoff. $RC_{S/W}$ was calculated using

$$RC_{S/W} = \frac{RC_S}{RC_W} = \frac{\overline{Q_S}/\overline{P_S}}{\overline{Q_W}/\overline{P_W}}, \quad (1)$$

where $\overline{P_S}$ and $\overline{Q_S}$ (resp. $\overline{P_W}$ and $\overline{Q_W}$) are the mean precipitation and discharge during three summer (resp. winter) months (Jul–Sep, resp. Jan–Mar) over ten-year time series (1997–2006). This ratio can be meaningfully computed and interpreted because the catchments were selected to have a limited snow influence. The meaning of $RC_{S/W}$ is best explained by two extreme cases: (1) a low $RC_{S/W}$ value (close to 0) means that the summer runoff coefficient, RC_S, is small compared to winter runoff coefficient, RC_W, i.e. a much lower percentage of rainfall will reach the catchment outlet in summer, which is an indication of catchments having limited baseflow and large summer losses by evapotranspiration. This case is classified as "Direct runoff"; (2) a high $RC_{S/W}$ value (closer to 1) means that RC_S and RC_W values are close, i.e. the propensity of the catchment to yield runoff is similar in summer and winter, which is an indication of

Table 2. Catchment characteristics and number of catchments in each classification range. Classifications are valid for the 237 French catchments used in this study.

Characteristic	Classification	Range	# Catchments
Area [km^2]	Small	16–6836	(237)
	Medium	< 200	85
	Large	200–600	79
		>600	73
Aridity index [−]	Dry	0.41–2.03	(237)
	Moist	< 1.2	78
	Wet	1.2–1.5	83
		> 1.5	76
$RC_{S/W}$ [−]	Direct runoff	0.03–1.00	(237)
	Mixed	< 0.15	87
	Groundwater	0.15–0.24	74
	dominated	> 0.24	76

groundwater-dominated catchments with limited impact of summer losses by evapotranspiration. This case is classified as "Groundwater-dominated runoff".

The aridity index (Middleton and Thomas, 1992) was calculated as the ratio between mean precipitation and PE over the ten-year time series.

For each catchment characteristic, the catchment set is divided into three classes with approximately equal numbers of catchments. Table 2 shows the ranges and number of catchments in each class. In this study, the ranges are used solely to distinguish between catchments with below-median, median or above-median characteristics.

2.3 Models

The SUPERFLEX modelling approach was used to hypothesize 12 alternative model structures. The GR4H model was used as an example of a fixed model structure. All thirteen models are lumped and use the same rainfall and PE over the whole catchment as inputs and generate discharge as output.

2.3.1 SUPERFLEX

Twelve structures (SF01-SF12) as proposed by Fenicia et al. (2013) are generated using the SUPERFLEX framework. They cover a relatively broad range of conceptual model complexities (Fig. 2). Starting from a very simple structure (SF01), the complexity is gradually increased by adding reservoirs and lag-functions. In this way, the influence of individual components can be assessed.

In the SUPERFLEX structures, rainfall (Pt) and potential evapotranspiration (PE) are used as inputs. Potential evapotranspiration is systematically corrected with a calibrated factor Ce to fulfil the water balance. The choice to use a correction factor for PE was made because: (1) it is hypothesized that the main bias lies in the estimation of the potential evapotranspiration, (2) multiplication factors to the potential evapotranspiration are commonly used to account for different land use types (e.g. the "crop factor"), and (3) it was conceptually simple, though possibly less efficient for flow simulation than other correction functions (like underground exchanges, see e.g. Le Moine et al., 2007). Actual evapotranspiration (noted Ei, Eu or Ef for the interception, unsaturated zone and fast reservoirs respectively) applies to one or two reservoirs in each model structure.

SF01 consists of a single fast reservoir (FR) with a nonlinear storage-discharge relationship characterized by a time constant K_f and a power parameter α. SF02 also consists of a single reservoir, but it represents the unsaturated zone reservoir (UR) and uses a linear function to describe outflow and a power function (with parameter β) to describe the surface runoff.

In structures SF03 to SF05, an unsaturated zone reservoir (UR) is connected in series to a fast reservoir. These three structures differ in the constitutive functions used to describe the flows between the reservoirs and in the number of calibrated parameters. SF05 uses power functions to describe outflows from both reservoirs and a lag-function to represent channel routing. Compared to SF05, SF06 has an interception reservoir (IR) and SF07 has a riparian reservoir (RR).

SF08 is a simple structure with two parallel reservoirs, namely the fast reservoir (FR) and a slow reservoir (SR) with a time constant K_s. The structures SF09-SF12 build on SF08 with increasing complexity. Parameters M and D are ratios that divide the flows between different reservoirs. Table 3 summarizes the model structures. A more detailed description can be found in Fenicia et al. (2013).

2.3.2 GR4H

The GR4H model (Le Moine, 2008), which is an hourly version of GR4J (Perrin et al., 2003), is used as an example of a fixed model structure. As shown in Fig. 2, rainfall and PE are subtracted to determine the net precipitation P_n or net evaporation E_n. P_n is partitioned between storage in a soil moisture reservoir $S(P_s)$ and effective rainfall ($P_t = P_n - P_s$). The soil moisture reservoir is depleted by percolation *Perc*. Effective rainfall is then routed to the outlet via two branches: 10 % is routed via a single unit hydrograph, while 90 % is routed via a unit hydrograph and a nonlinear routing store R. A water gain/loss function F is applied to both flow components to represent groundwater exchanges with underlying aquifers and/or neighbouring catchments.

2.4 Model evaluation

2.4.1 Evaluation procedure

All model structures were calibrated and validated using the split sample test (Klemes, 1986), in which ten years of the available data were split into two independent periods

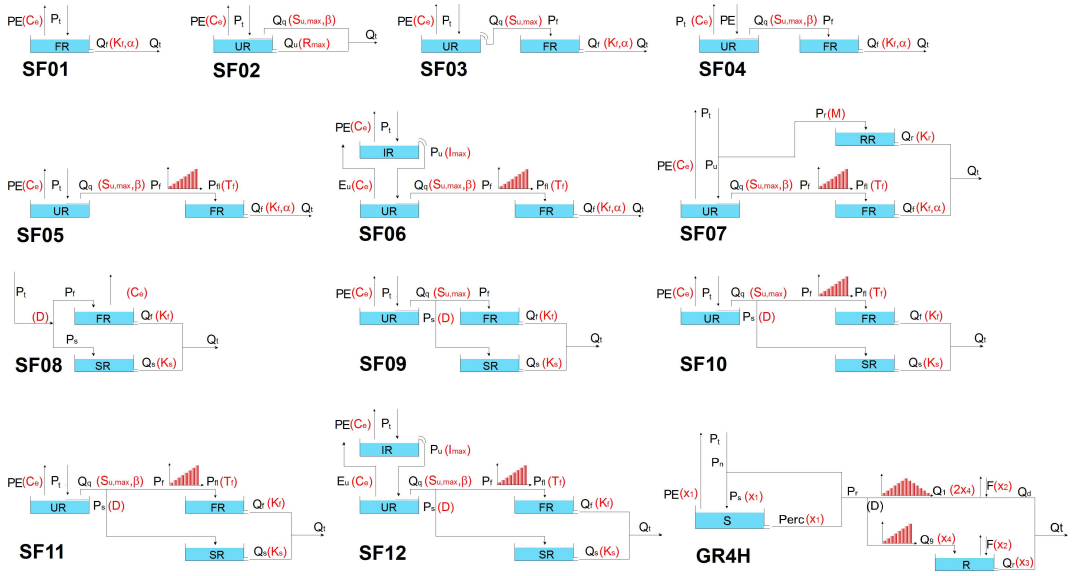

Fig. 2. Schematics of the 12 model structures generated using the SUPERFLEX framework (SF01–SF12 by Fenicia et al., 2013) and the GR4H model (Perrin et al., 2003). A flux of water is denoted in black next to an arrow line. Parameters are given in brackets in red font.

Table 3. Main characteristics of the twelve SUPERFLEX structures and GR4H. N_{res} and N_θ are the number of reservoirs and calibrated parameters respectively. "Rank" denotes the complexity rank of each structure based on the number of parameters, and number and type of function(s), where 1 is the simplest and 13 is the most complex.

Structure	N_{res}	Connection	Type of function(s)	N_θ	Rank
SF01	1	–	Power	3	1
SF02	1	–	Power + Linear	4	2
SF03	2	Series only	Threshold + Power	4	3
SF04	2	Series only	Power	5	6
SF05	2	Series only	Lag + Power	6	8
SF06	3	Series only	Threshold + Lag + Power	7	10
SF07	3	Series and parallel	Lag + Power + Linear	8	12
SF08	2	Parallel only	Linear	4	4
SF09	3	Series and parallel	Linear	5	7
SF10	3	Series and parallel	Lag + Linear	6	9
SF11	3	Series and parallel	Lag + Power + Linear	7	11
SF12	4	Series and parallel	Threshold + Lag + Power + Linear	8	13
GR4H	2	Series and parallel	Lag + Power + Linear	4	5

(1997–2001 and 2002–2006). Calibration was performed on each period followed by validation on the other period. To reduce model initialization problems, the reservoirs were initialized using three years of warm-up data (1994–1996 and 1999–2001 respectively) preceding each period. The data used for warm-up prior to 1997 originate from the SAFRAN reanalysis (Vidal et al., 2010).

Parameter optimization was carried out using the weighted least squares (WLS) scheme as described in Kavetski and Fenicia (2011), implemented within the Bayesian total error analysis (BATEA) framework and software (Kavetski et al., 2006; Kavetski and Evin, 2011). The WLS scheme accounts for the heteroscedasticity of the errors in the model predictions (i.e. larger errors in the larger flows) and provides a better balance between fitting high and low flows. A quasi-Newton optimization was applied using 20 different initial values across the parameter space. This local optimization method was shown to be effective and efficient when applied on smooth parameter spaces (Kavetski and Kuczera, 2007).

Model performance was evaluated based on the validation results using the four criteria described in the next section. First, the thirteen model structures were compared as individual models using an averaged evaluation criterion for all catchments. Then, for comparison of the flexible SUPERFLEX approach with the fixed GR4H model, consistency rules were applied to evaluate the consistency of the model

structure and parameter identifications on the two independent test periods for each catchment.

2.4.2 Evaluation criteria

The four evaluation criteria CR1 to CR4, given in Eq. (2) to 2.4.3, focus on different aspects of model performance: high flows, low flows, volume error and variability of predictions:

$$\mathrm{CR1} = \frac{\mathrm{CR1}^*}{2 - \mathrm{CR1}^*} \text{ with } \mathrm{CR}\ 1^* = 1 - \frac{\sum_{i=1}^{N} \left(Q_{\mathrm{obs},i} - Q_{\mathrm{sim},i}\right)^2}{\sum_{i=1}^{N} \left(Q_{\mathrm{obs},i} - \bar{Q_{\mathrm{obs}}}\right)^2} \quad (2)$$

$$\mathrm{CR2} = \frac{\mathrm{CR2}^*}{2 - \mathrm{CR2}^*} \text{ with } \mathrm{CR2}^* = 1 - \frac{\sum_{i=1}^{N} \left(\frac{1}{Q_{\mathrm{obs},i}+\epsilon} - \frac{1}{Q_{\mathrm{sim},i}+\epsilon}\right)^2}{\sum_{i=1}^{N} \left(\frac{1}{Q_{\mathrm{obs},i}+\epsilon} - \frac{\bar{1}}{Q_{\mathrm{obs}}+\epsilon}\right)^2} \quad (3)$$

$$\mathrm{CR3} = \frac{\mathrm{CR3}^*}{2 - \mathrm{CR3}^*} \text{ with } \mathrm{CR3}^* = 1 - \left| \sqrt{\frac{\sum_{i=1}^{N} Q_{\mathrm{sim},i}}{\sum_{i=1}^{N} Q_{\mathrm{obs},i}}} - \sqrt{\frac{\sum_{i=1}^{n} Q_{\mathrm{obs},i}}{\sum_{i=1}^{n} Q_{\mathrm{sim},i}}} \right| \quad (4)$$

$$\mathrm{CR4} = \begin{cases} -1 + 2\frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}} \sigma_{\mathrm{obs}} > \sigma_{\mathrm{sim}} \\ -1 + 2\frac{\sigma_{\mathrm{obs}}}{\sigma_{\mathrm{sim}}} \sigma_{\mathrm{obs}} < \sigma_{\mathrm{sim}} \end{cases}, \quad (5)$$

where Q_{obs} and Q_{sim} represent the observed and simulated discharge respectively; at time step i, N is the number of time steps, the over bar represents an average over the selected period, ϵ is a small constant (1 % of the mean streamflow, see Pushpalatha et al., 2012 for more information) and σ is the standard deviation of the streamflow over the selected period.

CR1* is the well-known Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) which is most sensitive to peaks in discharge (Perrin et al., 2003). CR2* is the Nash–Sutcliffe efficiency based on the inversed discharge emphasizing low flow errors (Pushpalatha et al., 2012). CR3* is based on the relative volume error and thus emphasizes any error in the water balance between observed and simulated discharge (Perrin et al., 2003). CR1* to CR3* have values between 1 (perfect fit) and $-\infty$ and are transformed to a value between 1 and -1 to avoid the influence of very low negative values on the calculation of mean performance (Mathevet et al., 2006; Pushpalatha et al., 2012).

The fourth criterion (CR4) is the ratio of standard deviations of observed and simulated discharges, with a maximum value of 1 indicating that the simulated discharge reproduces exactly the variability in the observed discharge versus a minimum value of -1 indicating large differences in the variability of the simulated and observed time series (Gupta et al., 2009).

The average of CR1–CR4 over the validation periods is used as the single metric of overall performance to compare model structures. Although this comprehensive criterion hides differences between the four individual statistics (CR1 to CR4), the distinction between poor and good models was quite similar for all of them. Therefore, we used this average criterion to summarize and compare model performance.

2.4.3 Consistency

The consistency of the model between the two calibration periods for a given catchment was evaluated by checking for parameter and structural consistency. Here, a model structure was considered parametrically inconsistent when at least one of the parameters differed by more than 50 % from the average between the two calibration periods. In that case, this structure was left out of the final comparison.

In the case of SUPERFLEX, structural inconsistency was considered to occur when the best model structure identified on the two calibration periods was not the same. Therefore, in any given catchment, a SUPERFLEX structure is considered consistent and eligible for comparison only if it has an average of CR1–CR4 within 10 % of the best performing structure in both calibration periods. The "best" SUPERFLEX structure for a given catchment was identified as the simplest consistent structure with an average of CR1–CR4 within 10 % of the results for the best performing structure. Here, model complexity is quantified by the number of parameters. When two equivalently performing structures have the same number of parameters, then the one with the least number of flow paths is considered simpler (see "Rank" in Table 3).

Using the split-sample test procedure and the consistency rules, models that give accurate and consistent results on a catchment can be identified. Despite some inevitable subjectivity, the consistency rules make the model evaluation more stringent.

3 Results

3.1 Average performance of the thirteen individual structures

Figure 3 shows the distribution of performance for all model structures on the 237 catchments for the two validation periods. The figure shows that the seven best performing model structures (GR4H, SF04–SF07, SF11 and SF12) show very similar average performance despite their structural differences. Six of the model structures (SF01, SF02, SF03, SF08, SF09 and SF10) perform poorly compared to the best seven structures, and are hence poor candidates for a fixed-structure approach. It can also be seen that structures with an unsaturated zone reservoir or a power function perform on average considerably better than structures without these components.

The calibrated values of the power function parameter α (describing the outflow of the fast reservoir FR) vary over a wide range for the different catchments and show a high

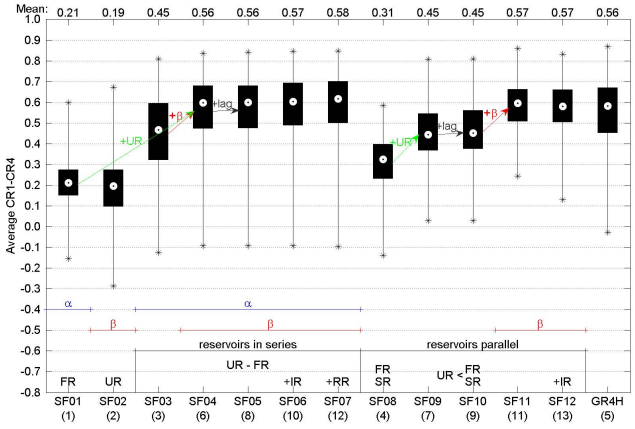

Fig. 3. Box plots (maximum, 75th percentile, median, 25th percentile and minimum) of averaged CR1–CR4 values of all model structures on the 237 catchments in the validation periods, including notes on differences in SUPERFLEX structures. Where "+β" means adding a power function β, "+lag" means adding a lag-function, UR = unsaturated zone reservoir, FR = fast reservoir, SR = slow reservoir, IR = interception reservoir and RR = riparian zone reservoir. The x axis shows the twelve SUPERFLEX structures plus GR4H. The number of model parameters (used as a measure of complexity) is given in brackets after the model name. The numbers across the top of the figure give the mean values of model performance.

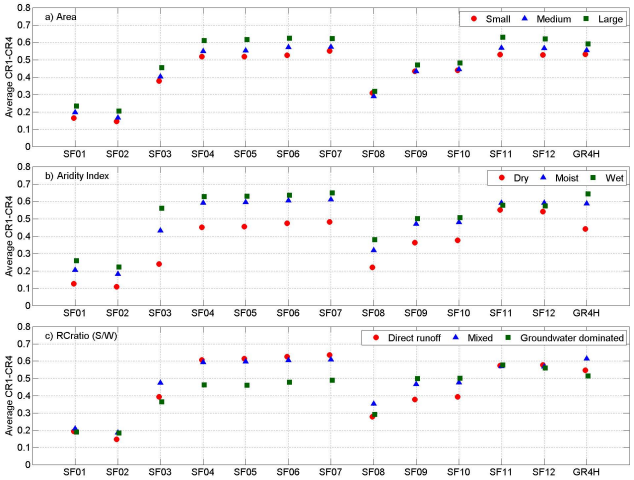

Fig. 4. Average of CR1-CR4 in validation in three classes of: **(a)** catchment area, **(b)** aridity index, and **(c)** $RC_{S/W}$.

correlation between the two periods (not shown here; see Van Esse, 2012, for more details). This indicates that this is an effective parameter to calibrate. The same was found for the power function parameter β (describing the outflow of the unsaturated zone reservoir UR).

However, for structures with both power functions, the values of power β are clustered around 1. This indicates that adding a second power function to the structure is far less effective than the first one, and suggests that using a nonlinear reservoir near the outlet of the flow network is more effective than placing such a reservoir upstream in the flow network.

Model comparisons show that the addition of a lag-function or an interception reservoir (IR) does not increase performance for most individual catchments. For the riparian zone reservoir (RR), the analysis shows larger differences: for some catchments this reservoir does increase performance while for others, the effects are negative. Overall, at least within the SUPERFLEX configurations applied in this work, the lag-function and the interception and riparian zone reservoirs do not increase average model performance, and therefore their inclusion into a "default" general model can be questioned.

The average performance and performance range of GR4H are close to those of SF04–SF07. The fixed power functions describing reservoir outflow in GR4H are expected to be important components, just as the power functions in the SUPERFLEX structures. The more complex models SF11 and SF12 are more robust in the sense of having a lower number of strong model failures, which is very valuable. However, on average, they are not able to outperform the simpler models (including GR4H).

3.2 Performance across catchment classes

To investigate the effect of catchment characteristics on model performance, average performance is analysed for three catchment characteristics: catchment area, aridity index, and the ratio of summer/winter runoff coefficients, $RC_{S/W}$.

Figure 4a shows that model performance is generally better on larger catchments than on smaller catchments. It is found to be easier to capture the rainfall-runoff relationship in catchments where hydrological processes are mixed and have a smoother behaviour. This corroborates previous findings by Merz et al. (2009) on a large set of Austrian catchments.

Figure 4b shows that model performance on wetter catchments is generally better than on drier catchments. The difference in performance on dry versus wet catchments is large for almost all model structures. This is in agreement with literature showing that drier catchments are generally more difficult to model due to the higher nonlinearity in the hydrological processes (e.g. Atkinson et al., 2002) and more complex error structure (Smith et al., 2010). The recent review study by Parajka et al. (2013) also shows that runoff hydrograph predictions in ungauged catchments tend to be more accurate in humid and large catchments.

From all models considered in this study, only SF11 and SF12 show little difference in performance over the three classes (however their performance is still worse for drier catchments). The parallel fast and slow reservoirs plus the

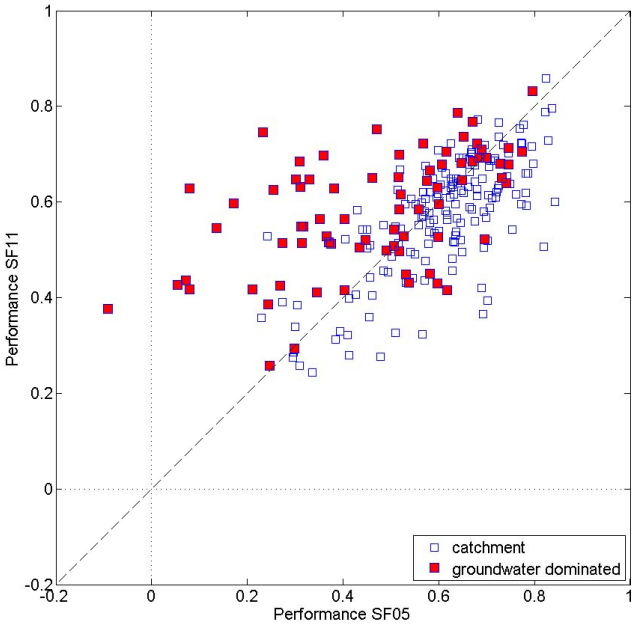

Fig. 5. Average performance of model structure SF05 (without parallel slow reservoir) against SF11 (with parallel slow reservoir). Groundwater-dominated catchments are marked in red.

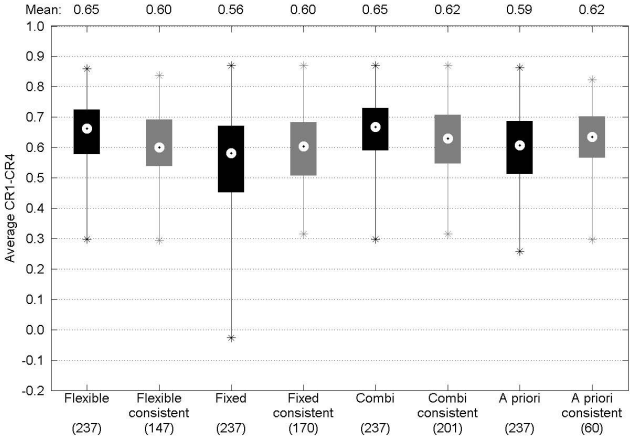

Fig. 6. Distribution of average performance in validation for the flexible, fixed, combined and a priori approaches. The whiskers denote the minimum and maximum values, the shaded boxes indicate the 25th–75th percentiles, and the circles indicate the median value. The dark boxes represent the distributions including inconsistent results, and the grey boxes represent only consistent results. The number of catchments in each group is given in brackets.

unsaturated zone reservoir with a power function seem to be better able to simulate both wetter and drier catchments.

The $RC_{S/W}$ classification in Fig. 4c separates the structures in an interesting way. Structures SF04 to SF07 perform much worse on the groundwater-dominated catchments than on the other catchments, while SF09 and SF10 simulate these catchments best of the three classes. This reversed order of performance can be explained by the parallel slow reservoir component in SF09 and SF10. This component allows independent fast and slow flow, hence enabling a slow groundwater component while maintaining the ability to produce high flow in case of a storm event. The parallel riparian zone reservoir in SF07 does not have this effect because it is not connected to the unsaturated zone reservoir and only a maximum of 20 % of rainfall is routed through this reservoir.

Differences in performance between serial versus parallel reservoir structures have already been investigated in the literature (e.g. Jakeman and Hornberger, 1993, and more recently by Kavetski and Fenicia, 2011, and Fenicia et al., 2013). Jakeman and Hornberger (1993) found that the most commonly identified configuration for a rainfall-runoff model is two parallel reservoirs. This difference is clearly observed when comparing the performance per catchment of SF05 and SF11 using the $RC_{S/W}$ classification. Figure 5 shows the average performance of SF05 against that of SF11, two identical models apart from the use of a parallel slow reservoir in SF11. The figure confirms that the more complex SF11 does not perform better overall than SF05 (mean average of CR1–CR4 of 0.57 and 0.56 respectively), but that it performs significantly better for most groundwater-dominated catchments. This shows the value of the flexible distribution of flow (ratio D) and the independent residence times (K_f and K_s) of the parallel reservoirs for these type of catchments.

3.3 Comparison of fixed and flexible modelling approaches

Figure 6 shows the performance distribution of the flexible and fixed approaches with and without accounting for the cases where inconsistent results between calibration periods occurred. Without the consistency rules, the flexible approach performs better on the catchment set. However, when the inconsistent catchments are removed from the set, both approaches perform very similarly.

The flexible approach is inconsistent more often than the fixed approach: 38 % versus 28 % of the catchments, respectively. If the two approaches are combined (i.e. using GR4H as one of the SUPERFLEX structures), then a significant gain in consistency is achieved: the combined approach is inconsistent for only 36 catchments (i.e. 15 % of the catchment set). This suggests that GR4H should be included as part of the "default" model selection offered in SUPERFLEX.

Note that applying the consistency rules creates two catchment sub-sets of different sizes, which may lead to some bias in comparing the performances. However, the mean results on the overlapping (116) catchments are very similar: an average of CR1–CR4 of 0.62 for GR4H and 0.61 for SUPERFLEX. The choices for the 50 % threshold for parameter inconsistency and the 10 % range for structural inconsistency also considerably influence these results: e.g. a 5 % structural inconsistency range would classify 134 catchments as

Table 4. Number of times each structure is selected as "best-performing" for the flexible approach alone (column labelled "Flexible") and when GR4H is considered as one of thirteen structures (column labelled "Combined").

Model (Rank)	Flexible	Combined
SF01 (1)	0	0
SF02 (2)	0	0
SF03 (3)	3	3
SF04 (6)	50	22
SF05 (8)	1	1
SF06 (10)	0	0
SF07 (12)	13	7
SF08 (4)	2	2
SF09 (7)	38	13
SF10 (9)	7	0
SF11 (11)	32	15
SF12 (13)	1	0
Total SUPERFLEX	147	63
GR4H (5)	–	138
Inconsistent	90	36
Total	237	237

having inconsistent results for the flexible approach, while a 25 % range would result in only 26 inconsistent catchments. Finally, the SUPERFLEX structures used in this study might be more prone to parameter inconsistency as most of them have more fitted parameters than GR4H.

To find out more about which of the flexible model structures perform well, we look at the number of times each structure is selected as best. Table 4 shows that SF04, SF09 and SF11 are selected most often. These structures are successful because they only use the most effective model components. Very simple structures and structures with only small differences from these best three are selected much less often. This shows a disadvantage of hypothesising a model structure a priori: many structures give similar results, which makes choosing the best one difficult. On the other hand this may also imply that selecting a sub-optimal model structure would not substantially deteriorate the model predictions compared to selecting the "best" structure. Combining all 13 structures, GR4H is selected the largest number of times (138), showing the good generality of the model. For 63 other catchments, however, one of the twelve SUPERFLEX structures is favoured. For these catchments, choosing a customized structure gives a better performance.

3.4 Possible link between structures and catchment type

The diversity of model structures and catchment conditions considered in this study lends itself to investigating possible relationships between catchment characteristics and well-performing model structures. The ability to a priori select model structures that perform well on particular types of catchments would be very helpful, in particular for predictions in ungauged basins.

Hence an attempt was made to find a best a priori model for different catchment types. To this end, 27 (3^3) sub-sets were created by crossing the three qualitative ranges (high, low and medium) of the catchment characteristics (area, aridity index and the ratio of seasonal runoff coefficients). Each catchment pertains to one of the sub-sets depending on its characteristics. On average, there are about eight catchments per sub-set. For each catchment sub-set, the model structure that performs best on average was selected as the best a priori model. The performance of this a priori model was then determined for all the catchment of the sub-set.

The results of this analysis are shown in the two box plots at the right of Fig. 6. The average performance of the a priori model is not better than that of the fixed or flexible approaches and is found to be much less consistent. Neither do the results show that this particular a priori method is useful for some specific types of catchments, as none of the sub-sets have a distinctly higher average performance. This is not to say that the a priori selection of a model structure is not possible, but further research is needed (e.g. along the lines of Fenicia et al., 2013).

4 Discussion

Fenicia et al. (2011), Kavetski and Fenicia (2011) and Fenicia et al. (2013) performed similar studies, but only on a few experimental catchments. The results of this study on a large set of 237 French catchments extend the previous works in several ways.

4.1 Catchment-to-model relationship

The results of this study show similar relationships between catchment characteristics and model structure as those found by the above-mentioned authors, especially in the case of groundwater-dominated catchments. Results obtained here are also consistent with results shown by Perrin et al. (2001), who showed that choosing a specific structure for a given catchment significantly improved the performance compared to using a fixed structure. However, efforts into defining a best a priori model for each type of catchment did not improve average performance compared to the fixed and flexible approaches. There are (at least) two hypotheses for this: either the used characteristics are not sufficient to fully characterize the hydrological behaviour of the catchment, or establishing a link between model structure and catchment properties should be investigated using more sophisticated approaches than the ones tested here. It should be noted that the differences in performance between quite a few models used here are probably well within the data uncertainty.

4.2 Challenges to the application of the flexible modelling approach

This research has illustrated important challenges in applying the flexible modelling approach on a large scale: similarities in performance of different model hypotheses complicate the selection of the best model and, in the absence of additional insights (such as in Fenicia et al., 2013), make it difficult to unambiguously relate model structure and hydrological processes. Relating model structure and hydrological processes becomes even more difficult given the large differences in results between the two calibration periods and the demand that the best model structure should work well for that specific catchment under all possible conditions (see also best-compromise model, Seiller et al., 2012). Furthermore, working with multiple model alternatives (as is the case in SUPERFLEX) requires more effort, and becoming experienced with the methodology can take time. On the other hand, in many catchments, multi-hypothesis frameworks such as SUPERFLEX can be used to improve the performance of fixed-structure models, and, in general, be used to construct model ensembles to describe structural uncertainties (see Clark et al., 2011b). The general advantages and disadvantages of model structures and components and their relevance for different catchment types can only be studied through a combination of large-scale studies (e.g., Andréassian et al., 2006; Gupta et al., 2013) and detailed insights from experimental catchments (e.g., Fenicia et al., 2013).

4.3 Influence of model structure

Considering the components of the SUPERFLEX structures, new systematic tests could be additionally carried out to further investigate the sensitivity of model results to the various modelling options identified as relevant here (power functions and parallel flow paths). Interestingly, in the case of the GR4H model and its daily version GR4J, recent investigations independent of the present study corroborate to some extent the results of this research: the addition of a free parameter in the formulation of the water exchange function significantly improves the simulation of low flows (Le Moine, 2008), and the addition of a second routing store (with an additional parameter) in parallel to the existing ones also significantly improves results (Pushpalatha et al., 2011).

4.4 Reasons for model failure

Identification of the cause for model failure and inconsistency can significantly help the process of model improvement and is consequently of major practical interest. Based on this study, it is difficult to clearly explain the reasons for poor model performance and/or inconsistency. Nonetheless, catchments for which all models performed poorly (below 0.5 on average of CR1–CR4) or were inconsistent could be separated into three groups, based on the likely cause of poor model performance:

1. Catchments where wet and dry periods have led to severe differences in observed flow between calibration and validation periods. These catchments proved to be difficult to model, especially when the effect of one wet or dry year lasts over multiple years (see also Coron et al., 2012). Many structures give inconsistent results on these catchments, but those with independent parallel flow paths have a higher success rate.

2. Catchments showing flashy flow, generally relatively small catchments located near the Mediterranean Sea. Many structures were inconsistent or gave poor performance due to poor simulation of sharp flow peaks. On these catchments, simple structures tend to perform relatively better and were often selected as best.

3. Catchments with unexplained variations in low flow measurements. These problems in observed flow may be attributed to downstream obstacles or measurement errors and are the most likely reason for poor model performance.

Further analysis of these model failures can prove useful for model development. For example, in-depth diagnosis of model parameter transposability in time was recently investigated by several authors (Vaze et al., 2010; Merz et al., 2011; Coron et al., 2012) and could be used to investigate the model failures reported in this study in more detail.

4.5 Low-flow simulations

In this study, model parameter calibration was carried out using a weighted least square (WLS) scheme, which puts more emphasis on low flows compared to the standard least square (SLS) scheme, because data and structural errors are generally higher for high flows. However, despite of this improved objective function, all model structures scored poorly on the low flow criterion (CR2*). The poor performance of the models in the low-flow simulations, even with a variety of modelling options, corroborates earlier findings by Smith et al. (2010), Pushpalatha et al. (2011), and others.

5 Conclusions

This study found that relatively simple model structures with carefully selected components can produce accurate simulations of catchment discharge, which corroborates previous findings (Jakeman and Hornberger, 1993; Perrin et al., 2001; Clark et al., 2008; and others).

The analysis of thirteen individual lumped conceptual model structures on 237 French catchments with a diverse range of characteristics showed that:

1. Increasing model complexity does not always lead to higher performance for a given catchment. However, complex structures perform poorly for fewer catchments;

2. Conceptual hydrological models generally perform better on larger and/or wetter catchments than on smaller and/or drier catchments;

3. The use of a power function to describe reservoir outflow significantly increases mean model performance compared to models that use linear outflow functions;

4. Parallel reservoirs with independent time constants increase model performance in groundwater-dominated catchments; and

5. The addition of a lag-function between reservoirs, or of an interception reservoir, does not lead to a significant increase in average model performance for a given catchment.

On the full catchment set, the flexible modelling approach provides better average results than the fixed modelling approach. Generally, selecting the best model structure for each catchment gives the best results. However, the results of the two approaches are comparable when applying consistency rules on parameters and structures for two calibration periods. The difficulties in identifying optimal model structures and parameter sets from rainfall-runoff data alone, as well as the frequent inconsistencies in the model performance across different time periods, illustrate the need for careful controlled approaches for model selection and development. Therefore, stringent model evaluation schemes should be designed to enhance the robustness of the selected model structures and reduce the likelihood of unexpected model failures (Andréassian et al., 2009, 2010; Clark et al., 2011b).

Acknowledgements. The authors thank Météo-France and SCHAPI for providing meteorological and hydrological data. Special thanks go to Vazken Andréassian from Irstea for his useful suggestions during this study.

Edited by: R. Merz

References

Andréassian, V., Hall, A., Chahinian, N., and Schaake, J.: Introduction and Synthesis: Why should hydrologists work on a large number of basin data sets?, IAHS Publ., 307, 1–5, 2006.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., and Valéry, A.: HESS Opinions "Crash tests for a standardized evaluation of hydrological models", Hydrol. Earth Syst. Sci., 13, 1757–1764, doi:10.5194/hess-13-1757-2009, 2009.

Andréassian, V., Perrin, C., Parent, E., and Bárdossy, A.: Editorial - The Court of Miracles of Hydrology: can failure stories contribute to hydrological science?, Hydrolog. Sci. J., 55, 849–856, doi:10.1080/02626667.2010.506050, 2010.

Atkinson, S. E., Woods, R. A., and Sivapalan, M.: Climate and landscape controls on water balance model complexity over changing timescales, Water Resour. Res., 38, 1314, doi:10.1029/2002WR001487, 2002.

Bergström, S.: The HBV model, in: Computer Models of Watershed Hydrology, Chapter 13, edited by: Singh, V. P., Water Resources Publications, 443–476, 1995.

Beven, K. J., Lamb, R., Quinn, P., Romanowicz, R., and Freer, J.: TOPMODEL, in: Computer models of watershed hydrology, edited by: Singh, V. P., Water Resources Publications, Highlands Ranch, Colorado, 627–668, 1995.

Booij, M. J.: Determination and integration of appropriate spatial scales for river basin modelling, Hydrol. Process., 17, 2581–2598, doi:10.1002/hyp.1268, 2003.

Breuer, L., Huisman, J. A.,Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM), I: Model intercomparison with current land use, Adv. Water Res., 32, 129–146, 2009.

Chiew, F., Stewardson, M., and McMahon, T.: Comparison of six rainfall-runoff modelling approaches, J. Hydrol., 147, 1–36, 1993.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007wr006735, 2008.

Clark, M. P., McMillan, H., Collins, D., Kavetski, D., and Woods, R. A.: Hydrological field data from a modeller's perspective: Part 2. Process-based evaluation of model hypotheses, Hydrol. Process., 400, 523–543, doi:10.1002/hyp.7902, 2011a.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resour. Res., 47, W09301, doi:10.1029/2010wr009827, 2011b.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments., Water Resour. Res., 48, W05552, doi:10.1029/2011wr011721, 2012.

DHI: MIKE11 – Reference and User's Manual, DHI Water and Environment, Hørsholm, Denmark, 2008.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrol. Earth Syst. Sci., 17, 1893–1912, doi:10.5194/hess-17-1893-2013, 2013.

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G.: Learning from model improvement: On the contribution of complementary data to process understanding, Water Resour. Res., 44, W06419, doi:10.1029/2007wr006386, 2008.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resour. Res., 47, W11510, doi:10.1029/2010wr010174, 2011.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., and Freer, J.: Catchment properties, function, and conceptual model representation: is there a correspondence?, Hydrol. Process., doi:10.1002/hyp.9726, online first, 2013.

Goswami, M. and O'Connor, K. M.: A "monster" that made the SMAR conceptual model "right for the wrong reasons", Hydrolog. Sci. J., 55, 913–927, 2010.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Gupta, H. V., Perrin, C., Kumar, R., Blöschl, G., Clark, M., Montanari, A., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci. Discuss., 10, 9147–9189, doi:10.5194/hessd-10-9147-2013, 2013.

Jakeman, A. J. and Hornberger, G. M.: How much complexity is warranted in a rainfall-runoff model ?, Water Resour. Res., 29, 2637–2649, 1993.

Kavetski, D. and Evin, G.: BATEA The User Guide (v 7.020.005), University of Newcastle, Australia, 2011.

Kavetski, D. and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resour. Res., 47, W11511, doi:10.1029/2011wr010748, 2011.

Kavetski, D. and Kuczera, G.: Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration, Water Resour. Res., 43, W03411, doi:10.1029/2006wr005195, 2007.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resour. Res., 42, W03407, doi:10.1029/2005wr004368, 2006.

Klemes, V.: Operational Testing of Hydrological Simulation-Models, Hydrolog. Sci. J., 31, 13–24, 1986.

Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J. A., Bilotta, G. S., Brazier, R. E., Butler, P., and Haygarth, P. M.: Ensemble evaluation of hydrological model hypotheses, Water Resour. Res., 46, W07516, doi:10.1029/2009WR007845, 2010.

Kuczera, G., Renard, B., Thyer, M., and Kavetski, D.: There are no hydrological monsters, just models and observations with large uncertainties!, Hydrol. Sci. J.-J. Sci. Hydrol., 55, 980–991, doi:10.1080/02626667.2010.504677, 2010.

Leavesley, G. H., Markstrom, S. L., Restrepo, P. J., and Viger, R. J.: A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modelling, Hydrol. Process., 16, 173–187, doi:10.1002/hyp.344, 2002.

Le Moine, N.: Le bassin versant de surface vu par le souterrain: une voie d'amélioration des performances et du réalisme des modèles pluie-débit?, Ph.D. thesis, Université Pierre et Marie Curie, Paris, 324 pp., 2008.

Le Moine, N., Andréassian, V., Perrin, C., and Michel, C.: How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments, Water Resour. Res., 43, W06428, doi:10.1029/2006WR005608, 2007.

Le Moine, N., Andréassian, V., and Mathevet, T.: Confronting surface- and groundwater balances on the La Rochefoucauld-Touvre karstic system (Charente, France), Water Resour. Res., 44, W03403, doi:10.1029/2007wr005984, 2008.

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

Linsley, R. K. and Crawford, N. H.: Computation of a synthetic streamflow record on a digital computer, IAHS Publ., 51, 526–538, 1960.

Mathevet, T., Michel, C., Andréassian, V., and Perrin, C.: A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins, Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment–MOPEX, IAHS Publ., 307, 211–219, 2006.

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M., and Woods, R. A.: Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure, Hydrol. Process., 25, 511–522, doi:10.1002/hyp.7841, 2011.

MEDD: Banque Hydro, Retrieved 2012, from Le Ministère de l'Ecologie et du Développement Durable (MEDD), Direction de l'Eau et le SCHAPI, available at: www.hydro.eaufrance.fr/ (last access : 27 August 2012), 2007.

Merz, R., Parajka, J., and Blöschl, G.: Scale effects in conceptual hydrological modeling, Water Resour. Res., 45, W09405, doi:10.1029/2009wr007872, 2009.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resour. Res., 47, W02531, doi:10.1029/2010WR009505, 2011.

Middleton, N. and Thomas, D. S. G.: World Atlas of Desertification, United Nations Environment Programme (UNEP), London, 1992.

Moussa, R. and Bocquillon, C.: Criteria for the choice of flood-routing methods in natural channels, J. Hydrol., 186, 1–30, 1996.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a rainfall-runoff model? Part 2 – Towards a simple and efficient PE model for rainfall-runoff modelling, J. Hydrol., 303, 290–306, doi:10.1016/j.jhydrol.2004.1008.1026, 2005.

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies, Hydrol. Earth Syst. Sci., 17, 1783–1795, doi:10.5194/hess-17-1783-2013, 2013.

Pechlivanidis, I. G., Jackson, B. M., McIntyre, N. R., and Wheater, H. S.: Catchment scale hydrological modelling: a review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications, Global Nest J., 13, 193–214, 2011.

Perrin, C., Michel, C., and Andreassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, J. Hydrol., 242, 275–301, 2001.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, doi:10.1016/s0022-1694(03)00225-7, 2003.

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., and Andréassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, J. Hydrol., 411, 66–76, 2011.

Pushpalatha, R., Perrin, C., Le Moine, N., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, J. Hydrol., 420–421, 171–182, doi:10.1016/j.jhydrol.2011.11.055, 2012.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D., and Dmip participants: Overall distributed model intercomparison project results, J. Hydrol., 298, 27–60, 2004.

Refsgaard, J. C. and Hansen, J. R.: A good-looking catchment can turn into a modeller's nightmare, Hydrolog. Sci. J., 55, 899–912, doi:10.1080/02626667.2010.505571, 2010.

Refsgaard, J. C. and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models, Water Resour. Res., 32, 2189, doi:10.1029/96WR00896, 1996.

Refsgaard, J. C., Henriksen, H. J., Harrar, W. G., Scholten, H., and Kassahun, A.: Quality assurance in model based water management – review of existing practice and outline of new approaches, Environ. Modell. Softw., 20, 1201–1215, 2005.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., and Franks, S.W.: Toward a reliable decomposition of predictive uncertainty in hydrological modeling – characterizing rainfall errors using conditional simulation", Water Resour. Res., 47, W11516, doi:10.1029/2011WR010643, 2011.

Rogers, P.: On the choice of "appropriate model" for water resources planning and management, Water Resour. Res., 14, 1003–1010, 1978.

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrol. Earth Syst. Sc., 16, 1171–1189, doi:10.5194/hess-16-1171-2012, 2012.

Singh, V. P. and Frevert, D. K. (Eds.): Mathematical Models of Small Watershed Hydrology and Applications, Water Resources Publications, Highlands Ranch, Colorado, 2002a.

Singh, V. P. and Frevert, D. K. (Eds.): Mathematical Models of Large Watershed Hydrology. Water Resources Publications, Highlands Ranch, Colorado, 2002b.

Smith, T., Sharma, A., Marshall, L., Mehrotra, R., and Sisson, S.: Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, Water Resour. Res., 46, W12551, doi:10.1029/2010wr009514, 2010.

Tabary, P., Dupuy, P., L'Henaff, G., Gueguen, C., Moulin, L., Laurantin, O., Merlier, C., and Soubeyroux, J.-M.: A 10-year (1997–2006) reanalysis of quantitative precipitation estimation over France: methodology and first results. IAHS Publ., 351, 255–260, 2012.

Tromp-van Meerveld, H. J. and McDonnell, J. J.: Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis, Water Resour. Res., 42, W02411, doi:10.1029/2004WR003800, 2006.

Valéry, A., Andréassian, V., and Perrin, C.: Regionalization of preciptation and air temperature over high altitude catchments – learning from outliers, Hydrolog. Sci. J., 55, 928–940, doi:10.1080/02626667.2010.504676, 2010.

Van Esse, W. R.: Demystifying hydrological monsters – Can flexibility in model structure help explain monster catchments?, M.Sc. thesis, University of Twente, Enschede, the Netherlands, 111 pp., 2012.

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., and Teng, J.: Climate non-stationarity - Validity of calibrated rainfall-runoff models for use in climate change studies, J. Hydrol., 394, 447–457, 2010.

Velázquez, J. A., Anctil, F., and Perrin, C.: Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, Hydrol. Earth Syst. Sci., 14, 2303–2317, doi:10.5194/hess-14-2303-2010, 2010.

Vidal, J.-P., Martin, E., Franchisteguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system. Int. J. Climatol., 30, 1627–1644, 2010.

Wagener, T.: Evaluation of catchment models, Hydrol. Process., 17, 3375–3378, doi:10.1002/hyp.5158, 2003.

Wagener, T., Lees, M. J., and Wheater, H. S.: A toolkit for the development and application of parsimonious hydrological models, in: Mathematical models of small watershed hydrology, edited by: Singh, V. P., Frevert, R., and Meyers, D., Water Resources Publications, LLC, USA, 2001.

Zhao, R. J. and Liu, X. R.: The Xinanjiang model, in:Computer models of watershed hydrology, edited by: Singh, V. P., Water Resources Publications, Highlands Ranch, Colorado, 215–232, 1995.