Manuscript prepared for Hydrol. Earth Syst. Sci. Discuss.
with version 4.1 of the LaTeX class copernicus_discussions.cls.
Date: 19 November 2013

On the lack of robustness of hydrologic models regarding water balance simulation – a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments

L. Coron[1,2], **V. Andréassian**[1], **C. Perrin**[1], **M. Bourqui**[2], **and F. Hendrickx**[2]

[1]Irstea (formerly Cemagref), UR HBAN, 1 rue Pierre-Gilles de Gennes, 92761 Antony, France
[2]EDF R&D LNHE, 6 quai Watier, 78401 Chatou, France

Correspondence to: L. Coron (laurent.coron@irstea.fr)

Abstract

This paper investigates the robustness of rainfall-runoff models when their parameters are transferred in time. More specifically, we propose an approach to diagnose their ability to simulate water balance on periods with different hydroclimatic characteristics. The testing procedure consists in a series of parameter calibrations over 10-yr periods and the systematic analysis of mean flow volume errors on long records. This procedure was applied to three conceptual models of increasing structural complexity over 20 mountainous catchments in southern France. The results showed that robustness problems are common. Errors on 10-yr-mean flow volume were significant for all calibration periods and model structures. Various graphical and numerical tools were used to investigate these errors and unexpectedly strong similarities were found on the temporal evolutions of these volume errors. We indeed showed that relative changes in simulated mean flow between 10-yr periods can remain similar, regardless of the calibration period or the conceptual model used. Surprisingly, using longer records for parameters optimisation or using a semi-distributed 19-parameter daily model instead of a simple 1-parameter annual formula, did not provide significant improvements regarding these simulation errors on flow volumes. While the actual causes for these robustness problems can be manifold and are difficult to identify in each case, this work highlights that the transferability of water balance adjustments made during calibration can be poor, with potentially huge impacts in the case of studies in non-stationary conditions.

1 Introduction

1.1 Confidence and evaluation of rainfall–runoff modelling in a changing climate

Whether or not climate stationarity is an appropriate concept, it is becoming increasingly difficult to consider that catchments are static environmental systems (Milly et al., 2008; Koutsoyiannis, 2011; Matalas, 2012; Muñoz et al., 2013). The hydro-climatic conditions observed during historical periods cannot be easily considered as representative of other periods (histori-

cal or future). At the same time, hydrological models are increasingly used for water resources management or risk assessment, often for future, and different, climatic conditions. To date, many unknowns remain concerning the robustness of conceptual models in a changing climate.

The question of hydrological models' abilities in changing conditions has recently gained much interest, as demonstrated by the new IAHS Scientific Decade: "Panta Rhei" (Montanari et al., 2013). The temporal and climatic transferability of model parameters has been increasingly studied over the past few years, using the test procedures suggested by Klemeš (1986). It is now clear that a rainfall-runoff (RR) model calibrated on a given period will generally not be able to simulate flows with a similar efficiency on another period, especially when it differs climatically. Various research team throughout the world have documented this (see Rosero et al., 2010; Vaze et al., 2010; Merz et al., 2011; Coron et al., 2012; Seifert et al., 2012; Seiller et al., 2012; Brigode et al., 2013; Gharari et al., 2013). They agree that conceptual models lack robustness when used in contrasted climate conditions.

Long historical records that include contrasted sub-periods are needed for evaluating models robustness. Indeed, projections of future discharges under a changed climate cannot be compared to observations, by definition. The lack of model robustness is often measured through changes in root-mean-square error, NS efficiency (Nash and Sutcliffe, 1970) or similar quadratic error criteria, between different periods. These criteria have the advantage of reflecting the model efficiency on all simulated time steps and can even be used to build "model robustness criteria", as discussed by Coron et al. (2012). In several publications examining this issue, the authors showed the existence of almost systematic biases on simulated volumes, depending on the transfer conditions for model parameters (see Vaze et al., 2010; Merz et al., 2011; Coron et al., 2012; Seiller et al., 2012). Solving the problems of incorrect water balance simulation requires further investigations and has motivated the study reported herein. They are particularly relevant in the context of climate change impact studies, where conditions are known to evolve but biases on simulated volumes are commonly considered constant, for lack of true robustness assessment.

Moreover, in conceptual modelling, failure situations of parameter transfer often seem to be blamed on the overly simplistic model used or the inadequate calibration period chosen, without

proper checking. Yet, schemes for systematic model testing and comparison are valuable tools. They allow progress to be made on the evaluation of the models' suitability but also on the understanding of real-world hydrological system functioning (Seibert, 2001; Andréassian et al., 2009; Clark et al., 2011). International initiatives such as DMIP (Smith et al., 2004, 2012), MOPEX (Schaake et al., 2006; Chahinian et al., 2006) and HEPEX (Schaake et al., 2007; Thielen et al., 2008) are good examples of the use of these testing scheme. We think that this type of evaluation approaches must be generalised and innovative strategies should be devised to make the best use of the long time series now available.

1.2 Scope of the paper

This paper deals with the evaluation of model robustness and was motivated by the recent findings on the difficulties for RR model parameters to reproduce water balances. We propose a simple diagnostic approach to further investigate this question. Using long hydrological records, we tested the capacity of three different models to simulate mean flows over series of successive 10-yr periods different from the calibration one. Specifically, we aimed at evaluating the influence of the model complexity or the period used for parameter calibration on this capacity to simulate water balances.

This paper is organised as follows: the catchment set and models used are presented in the next section; the testing methodology and analysis techniques are discussed in Sect. 3 and the corresponding results provided in Sect. 4; a general discussion and the overall conclusions are given in Sects. 5 and 6, respectively.

2 Catchments and models

2.1 Set of 20 French catchments

2.1.1 Data description

A set of 20 catchments was used to evaluate the robustness of hydrological models, in their ability to simulate water balances. These 20 catchments are located in southern France, mostly in mountainous areas (Massif Central, Pyrennees and French Alps, see Fig. 1). They cover a relatively wide range of characteristics, in terms of size, mean elevation, snow influence and aridity index (see Table 1). The hydrological regimes are largely influenced by the processes of snow accumulation and melt for the most elevated catchments, and only governed by rainfall and evapotranspiration variations for the lowest ones. Three case studies were chosen to provide examples of detailed results: the Ubaye River at Barcelonnette (case study 1), the Lot River at Barnassac (case study 2) and the Drac River at Pont de la Guinguette (case study 3). Case studies 1 and 3 are medium-size high-elevation catchments located in the Alps. They have quite similar characteristics but marked differences in terms of precipitation. Case study 2 is a larger catchment in the Massif Central, with lower elevation and consequently a much more limited snow influence.

Climate forcings and flow records were at least 40 yr long, which cover a wide range of hydrometeorological conditions. Daily flow data were extracted from the HYDRO national archive (www.hydro.eaufrance.fr). They were checked for errors (by visual inspection and double mass curves analysis with neighbouring stations) and erroneous data were considered as gaps. Total precipitation and air temperature series were computed using the SPAZM reanalysis, which is based on ground network data and weather patterns. Developed by Gottardi et al. (2012), this reanalysis is available on 1x1 km cells at a daily time step from 1948 to 2010 for the main mountainous areas in France. These forcings can be considered high-quality data. Finally, potential evapotranspiration (PE) time series were computed using either Thornthwaite (1948) or Oudin et al. (2005) formula depending on the model considered. In both cases, PE series were computed using air temperature from the SPAZM reanalysis.

2.1.2 Comments on the catchment selection process

The impact of the case studies' particularities on the interpretations drawn is always subject to discussion.

When the catchment set used in this work was built, we attempted to neither exclude nor over-represent problematic situations. The availability of records of sufficient length and quality for our diagnostic approach mostly governed the selection procedure. Suspicious records were not kept and the catchments used here should be free of obvious quality problems. Moreover, all the selected catchments are unregulated and are not particularly known for changes in their hydrological functioning for other reasons than climate variability.

The size of the catchment set was largely impacted by the demanding computation times for the calibration of the most complex model used in this work. From the initial database of 365 eligible catchments, 20 catchments were kept to proceed with the full diagnostic approach. These catchments were also selected to be roughly representative of the variety of conditions in the initial database (although snow-dominated catchments are slightly over represented). The set of 365 catchments was used to apply our testing procedure with the two simpler models, to confirm the findings presented here (the results can be found in the Appendix).

[INSERT FIGURE 1]

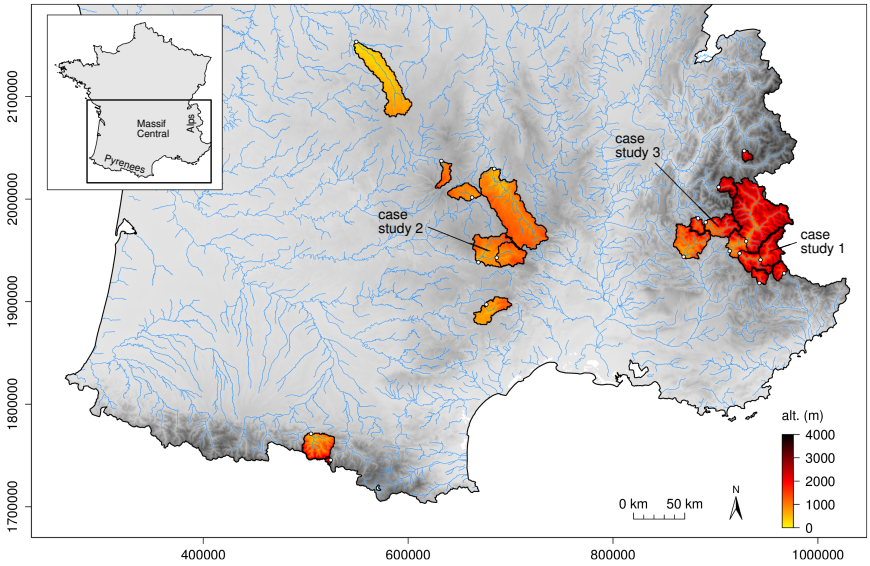

Fig. 1. Locations of the 20 catchments used in this study.

[INSERT TABLE 1]

Table 1. Characteristics of the 20-catchment set and the three case studies.

	Set of 20 catchments					Case studies		
	min	25th centile	median	75th centile	max	case study 1	case study 2	case study 3
Catchment surface [km^2]	24	170	490	1000	3600	540	1160	510
Mean elevation [m]	520	1100	1650	2180	2440	2270	1050	1700
Mean annual total precip. (P) [mm]	880	1180	1320	1460	2260	1210	990	1620
P_{solid}/P ratio (annual mean) [$-$]	4 %	11 %	38 %	46 %	59 %	47 %	11 %	42 %
Mean annual pot. evap. (PE_{Oudin}) [mm]	330	430	470	560	640	410	560	460
Mean annual discharge (Q) [mm]	370	550	710	980	1720	600	440	860
P/PE ratio (annual mean) [$-$]	1.55	1.98	2.97	3.23	5.23	2.94	1.78	3.51
Q/P ratio (annual mean) [$-$]	0.36	0.48	0.54	0.63	0.85	0.49	0.44	0.53
Available time series length [yr]	40	47	51	57	62	52	62	42

2.2 Three rainfall–runoff models of increasing complexity – a "modelling transect"

Three conceptual hydrological models were considered for this study and were chosen to cover a relatively wide range of structural complexity. Schematic diagrams of their structures are given in Fig. 2.

2.2.1 Mouelhi formula

The formula proposed by Mouelhi et al. (2006) is a simple annual model with a single calibrated parameter. It originates from the well-known Turc–Mezentsev formula (Turc, 1954; Mezentsev, 1955; Lebecherel et al., 2013). Its inputs are cumulated annual precipitation and PE data (computed using Oudin's formula). The model can be described using a non-linear equation:

$$Q_{a(j)} = P_{a(j)} \cdot \left(1 - 1/\left[1 + \left(\frac{0.7 \cdot P_{a(j)} + 0.3 \cdot P_{a(j-1)}}{\alpha \cdot \text{PE}_{a(j)}} \right) \right]^{0.5} \right) \quad (1)$$

where $Q_{a(j)}$, $P_{a(j)}$ and $\mathrm{PE}_{a(j)}$ are the annual discharge, precipitation and PE, respectively, for a given year j, while $P_{a(j-1)}$ is the annual precipitation for the previous year $(j-1)$.

2.2.2 GR4J-CemaNeige

GR4J is a parsimonious daily model with four calibrated parameters, described by Perrin et al. (2003). For this study, it is used with the CemaNeige degree-day-type snow module, developed by Valéry (2010). The required inputs for GR4J-CemaNeige are daily series for min/mean/max air temperature, precipitation and PE (computed using Oudin's formula). Both CemaNeige and GR4J are run at a daily time step. The snow module is computed over five elevation layers of equal surface and its outputs are then aggregated to feed GR4J, which is a lumped model. The snow module has two free parameters, which are optimised together with the four GR4J parameters.

2.2.3 Cequeau

Cequeau is a daily semi-distributed conceptual model, initially developed at INRS-Eau (Charbonneau et al., 1977). Here we used a modified version described in detail by Le Moine and Monteil (2012). The model inputs are daily series for min/mean/max air temperature, and precipitation. Cequeau includes a snow module and a parameterised function to adjust PE amounts (based on the Thornthwaite formula). These functions are included in the soil moisture accounting (SMA) part of the model, which complies with a topography-based mesh. The number of cells in this mesh is adjusted to the catchment size and topography (for the 20-catchment set used in this work, this number ranges from 10 to 30). Considering the entire model structure, a total of 19 parameters must be optimised.

2.2.4 Calibration procedure

Model parameters were calibrated by maximising the Kling-Gupta efficiency (KGE), proposed by Gupta et al. (2009). This criterion is given by:

$$\mathrm{KGE} = 1 - \sqrt{\left(\rho[Q,\widehat{Q}] - 1\right)^2 + \left(\frac{\sigma[\widehat{Q}]}{\sigma[Q]} - 1\right)^2 + \left(\frac{\mu[\widehat{Q}]}{\mu[Q]} - 1\right)^2} \quad (2)$$

where Q and $\widehat{Q}$ are the time series of observed and simulated flow, respectively, at an annual time step for the Mouelhi formula and a daily time step for the GR4J-CemaNeige and Cequeau models ; ρ, σ and μ are the Pearson correlation coefficient, the standard deviation and the mean, respectively.

Given the small number of free parameters for the Mouelhi formula and the GR4J-CemaNeige model, we used a simple two-step calibration procedure: first the parameter space was screened using a gross predefined grid and the best parameter set was then used as a starting point for a simple steepest ascent local search algorithm. This approach proved efficient for such parsimonious models compared to more complex search algorithms (Edijatno et al., 1999; Mathevet, 2005). The parameters from Cequeau were optimised using a more complex procedure developed by Le Moine (2009), which combines the multi-objective evolutionary annealing-simplex (MEAS) algorithm proposed by Efstratiadis and Koutsoyiannis (2005) and the multi-objective genetic algorithm, ε-NSGA-II, detailed by Reed and Devireddy (2004). This procedure has proved to be efficient in past applications of the Cequeau model for water resources assessment and dam management in France (Bourqui et al., 2011; François et al., 2013).

[INSERT FIGURE 2]

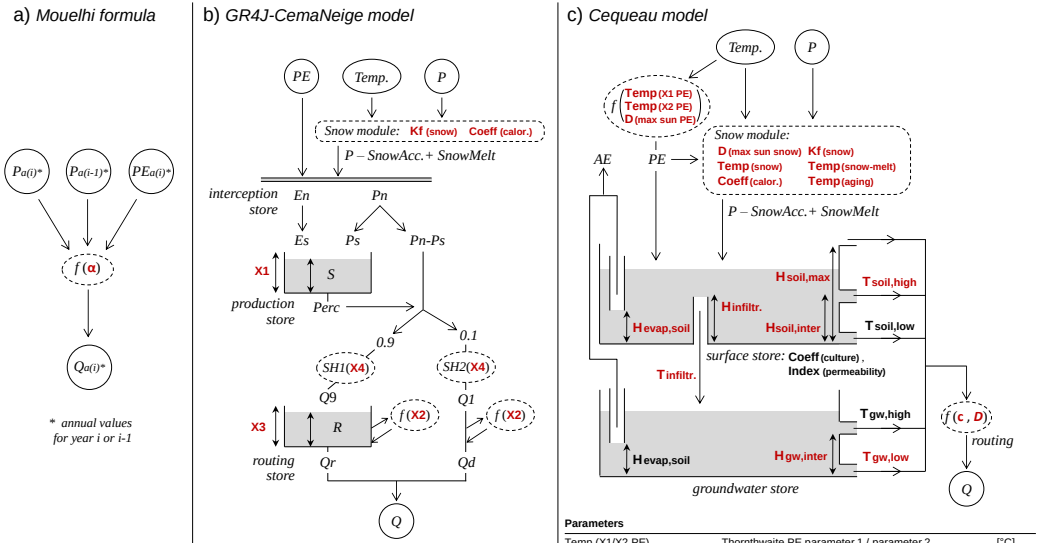

Fig. 2. Structural schemes of the three models tested: **(a)** the Mouelhi formula, **(b)** GR4J-CemaNeige and **(c)** Cequeau (optimised parameters are in red bold characters).

3 Robustness testing procedure

3.1 Sub-period calibration procedure

In a previous article, we proposed a testing methodology based on multiple transfer tests: the Generalised Split-Sample Test (GSST) procedure (Coron et al., 2012). It consists of a series of calibration-validation tests on independent sub-periods of equal length, considering all possible sub-period pairs. This testing procedure has been simplified for this study. The calibration sub-periods are built as in the GSST, i.e. using a sliding window that is moved by one hydrological year between two neighbouring sub-periods (overlap is allowed). However, we considered for this study a unique simulation period corresponding to the entire available time series, contrary to what was done in the GSST. As a result, the calibration and simulation periods were not independent and the transfer tests presented here should not be interpreted as strict split-sample tests. This testing procedure is illustrated in Fig. 3, where θ_i is the optimal parameter set identified on the sub-period i.

The testing procedure implemented in this work is highly dependent on the length of the sliding window used to build the calibration sub-periods. This length is chosen as a compromise simultaneously allowing for correct parameter determination and a sufficient number of contrasted sub-periods. Here, we considered 10-yr-long calibration sub-periods (SP), while the available total periods (TP) were at least 40-yr long and at most 62- yr long for the catchment set (i.e. the number of sub-periods built per catchment ranged from 31 to 53).

Hydrological years starting on October 1st from calendar year j and ending on September 30th from calendar year j+1 were used, for the time series split. Using hydrological instead of calendar years is important since some of the catchments considered in this work are snow-dominated (i.e. precipitations are stored as snow during the winter and only become runoff when spring arrives).

[INSERT FIGURE 3]

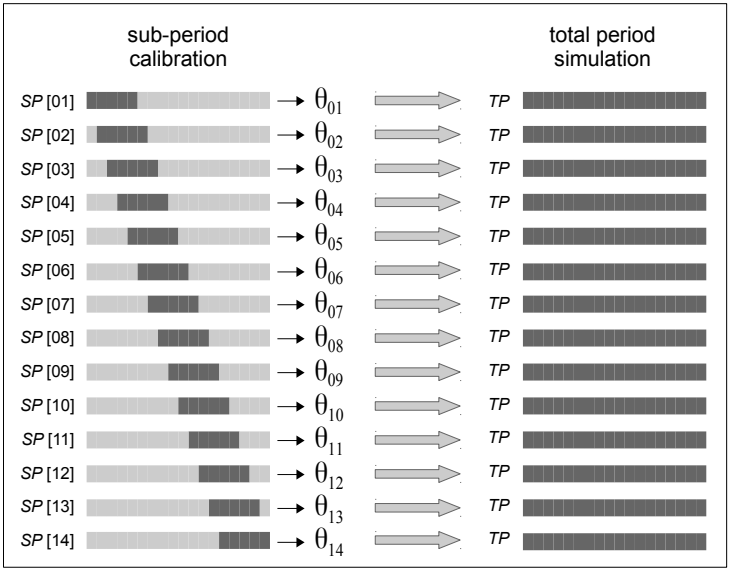

Fig. 3. Sub-period (SP) calibration procedure and simulation over the total period (TP) (example of 5-yr-long sub-periods within an 18-yr-long total period).

3.2 Model efficiencies

An overview of the model performances over the catchment set is provided in Table 2. For each catchment, KGE values were computed over the total available record, considering the various parameter sets stemming from our sub-period calibration procedure (see Fig. 3). For each model, the efficiencies were computed at the time step used to run the model, i.e. annual for the Mouelhi formula and daily for the GR4J-CemaNeige and Cequeau models. Additional

KGE were computed at the annual time step for GR4J-CemaNeige and Cequeau (after series aggregation).

For these tests, the calibration periods (SP) are included in the simulation period (TP). The KGE values in Table 2 are therefore not exactly "validation" efficiencies. Still, they give a good idea of the models' performances over the catchment set. On average, high efficiencies are reached for the daily models. Cequeau shows the highest criteria computed at both annual and daily time steps. The Mouelhi formula provides the lowest performances, but they remain acceptable on average over the set.

[INSERT TABLE 2]

Table 2. Model efficiencies computed over the total available records, considering sub-period calibrated parameter sets: $[\mathrm{KGE}_{\mathrm{TP}}]_{\theta_{\mathrm{SP}}}$.

		Set of 20 catchments					Case studies		
		min	25th centile	median	75th centile	max	case study 1	case study 2	case study 3
KGE at the annual time step	Mouelhi	0.048	0.572	0.760	0.883	0.942	0.899	0.713	0.919
	GR4J-CemaNeige	0.497	0.814	0.871	0.905	0.968	0.868	0.883	0.896
	Cequeau	0.277	0.810	0.881	0.921	0.971	0.884	0.898	0.901
KGE at the daily time step	GR4J-CemaNeige	0.670	0.828	0.866	0.899	0.943	0.864	0.848	0.838
	Cequeau	0.724	0.845	0.878	0.902	0.943	0.890	0.881	0.876

3.3 Visual tools for robustness analysis

Previous studies on the temporal robustness of conceptual hydrological models have shown that volume errors can be significant as a result of parameter transfer (Merz et al., 2011; Coron et al., 2012). To further investigate this issue, we studied the temporal variations of medium-term flow volume errors over the available records for different calibration configurations. These errors were expressed as a dimensionless bias given by $\overline{\widehat{Q}_{10\mathrm{yr}}}/\overline{Q_{10\mathrm{yr}}}$, in which $\overline{\widehat{Q}_{10\mathrm{yr}}}$ and $\overline{Q_{10\mathrm{yr}}}$ are the 10-yr-mean simulated and observed flows, respectively. The results obtained with dif-

ferent parameter sets can be superimposed on the same graph. Thus, we built visual tools for analysing model behaviours. We illustrate their construction with the example of the Ubaye River at Barcelonnette (case study 1 in Fig. 1) using the GR4J-Cemaneige model. Figure 4 shows the successive steps followed to plot the variations of mean flow volume errors.

Here, time series of precipitation, temperature and discharges were available over the 1959–2009 period. We built a total of 41 continuous sub-periods using a 10-yr-long sliding window following the procedure presented in Fig. 3. These sub-periods were used to calibrate models and to compute volume errors. The building procedure is explained is the next three subsections.

[INSERT FIGURE 4]

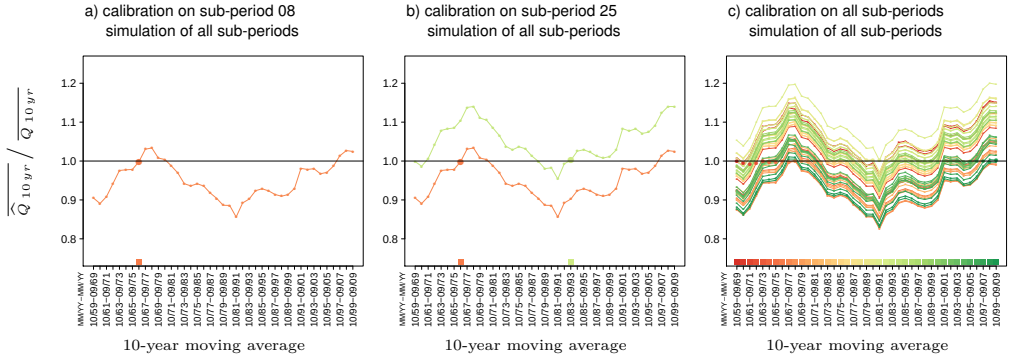

Fig. 4. Construction of the graphical representation of the series of 10-yr mean flow volume errors.

3.3.1 First step: using a single calibration period (Fig. 4a)

Let us consider the example of sub-period SP[08] and plot the point corresponding to the errors in calibration (large circle). Since volume errors are an important component of the calibration

criteria (KGE), the mean flow volume error obtained for SP[08] were small (i.e. $\overline{\widehat{Q}_{10yr}}/\overline{Q_{10yr}} \approx 1$). Then, from the simulated flows over the whole record using the calibrated parameter set, we could compute the mean flow volume error for each of the 40 remaining sub-periods and plot these errors for each of them (small dots). Note that there is an overlap between the calibration period and the neighbouring evaluation periods (for which the time lapse between starting years is less than nine years), but that the calibration and evaluation periods are independent in the other cases.

All 41 points were joined to form a curve, which is specific to the parameter set. This curve, noted $\omega_{\theta_{SP[08]}}$, corresponds to the 10-yr moving average of mean flow volume errors when the model calibrated on SP[08] is used. One can note significant simulation errors for this example, the range of volume error variations being 17.7%, with a standard deviation of 4.7%. This indicates that it is difficult for the model to reproduce observed 10-yr-mean flows on this catchment over the whole period. Phases of mean flow overestimation and underestimation are observed, but because of the sub-periods overlap, there is a smoothing effect on these variations.

3.3.2 Second step: adding another calibration period (Fig. 4b)

The previous step was repeated with a second calibration sub-period SP[25]. Again, mean flow volume errors were small on the calibration sub-period, but increased when the parameter set was transferred to simulate other parts of the time series. Interestingly, the shapes of the $\omega_{\theta_{SP[08]}}$ and $\omega_{\theta_{SP[25]}}$ curves are similar, although their vertical positioning on the graph differs.

3.3.3 Last step: combining all calibration periods (Fig. 4c)

This plotting procedure was used with all available parameter sets, i.e. considering all sub-periods as parameter "donors". In each case, the entire time series was simulated and errors were computed on the 10-yr sub-periods. It can be noted that mean flow volume errors remain small during calibration in all cases and that the shapes of all the curves are similar, showing a "parallelism effect".

3.3.4 Key questions

Numerous questions arose from the results obtained in the example of Fig. 4. First, each of the parallel curves illustrates a lack of robustness. A perfectly robust model would result in flat curves, i.e. the mean flow volume error would not depend on the period considered. Beyond noting alternating phases of 10-yr-mean flow over- and underestimation, we then focused on the following questions:

- The various parameter sets used to build Fig. 4c were optimised over 10 years. Are these calibration periods too short for the model to capture long-term dynamic processes? Would a calibration over the full record lead to correct volume simulations over the different parts of the time series?
- We observed behavioural similarities between different parameter sets on the Ubaye River at Barcelonnette. Are these similarities observed on other catchments from the set?
- Behavioural similarities were observed for GR4J-CemaNeige. Are these similarities observed for simpler or more complex conceptual models?

3.4 Numerical criteria for analysis

Numerical criteria were built to measure the parameter transferability issues in terms of volume errors and to assess the degree of similarity between series of mean flow volume errors obtained with different parameter sets. These criteria enabled us to generalise our analyses over multiple catchments and models.

3.4.1 Measures of transferability

Since the focus was put here on mean flow volume errors ($\overline{\widehat{Q}_{10yr}}/\overline{Q_{10yr}}$) and their temporal variations, we defined series of ω_θ curves as:

$$\omega_{\theta_{\mathrm{SP}[i]}} = (u_k)_{k\in[1:p]}\,; \quad u_k = \frac{\left[\overline{\widehat{Q}_{\mathrm{SP}[k]}}\right]_{\theta_{\mathrm{SP}[i]}}}{\overline{Q_{\mathrm{SP}[k]}}} \tag{3}$$

where $\mathrm{SP}[i]$ and $\mathrm{SP}[k]$ are the i-th and k-th 10-yr-long sub-periods chosen among the p possible ones; $\overline{Q_{\mathrm{SP[k]}}}$ is the mean observed flow on $\mathrm{SP}[k]$ and $[\overline{\widehat{Q}_{\mathrm{SP}[k]}}]_{\theta_{\mathrm{SP}[i]}}$ is the mean simulated flow on $\mathrm{SP}[k]$ using the parameter set optimised on $\mathrm{SP}[i]$.

Computable for each hydrological model, these $\omega_{\theta_{\mathrm{SP}[i]}}$ curves reflect the extent of mean flow volume errors. They can be compared to assess the impact of changing the calibration sub-period on these errors (as shown in the example from Fig. 4). An $\omega_{\theta_{\mathrm{TP}}}$ curve can be additionally considered in the comparison. It indicates the mean flow volume errors under calibration conditions, when both the calibration and simulation period correspond to the total period (TP). Because volume errors are an important component of the KGE calibration criterion, we expect $\omega_{\theta_{\mathrm{TP}}}$ to be the flattest of all the ω_θ curves. For this reason, we chose to consider the $\omega_{\theta_{\mathrm{TP}}}$ as a reference in the comparison criteria proposed hereafter.

In order to measure the magnitude of the volume error temporal variations, we used the standard deviation operator (σ) on the ω_θ curves. An example for the $\omega_{\theta_{\mathrm{TP}}}$ curve is given in Eq. (4).

$$\sigma\left[\omega_{\theta_{\mathrm{TP}}}\right] = \sqrt{\left(\frac{1}{p}\sum_{k=1}^{p}\left(u_k\right)^2\right) - \left(\frac{1}{p}\sum_{k=1}^{p}\left(u_k\right)\right)^2} \quad ; \quad u_k = \frac{\left[\overline{\widehat{Q}_{\mathrm{SP}[k]}}\right]_{\theta_{\mathrm{TP}}}}{\overline{Q_{\mathrm{SP}[k]}}} \tag{4}$$

with the same notations as in Eq. (3).

This criterion reveals the overall ability for a model to reproduce 10-yr-mean flow on various sub-periods when it is calibrated on the full available record. It varies between 0 (optimal situ-

ation with no errors) and $+\infty$. The largest the values, the smallest the model transferability in time (at least with respect to mean flow volume errors).

3.4.2 Measures of behavioural similarity

Other criteria were designed to specifically address the question of behavioural similarity highlighted in Fig. 4c.

In line with the criterion of Eq. (4), the standard deviation operator was used again, but with a different objective this time: measuring the similarity between ω_θ obtained from different parameter sets. The corresponding criterion is given in Eq. (5).

$$\sigma\left[\omega_{\theta_{\mathrm{SP}[i]}} - \omega_{\theta_{\mathrm{TP}}}\right] = \sqrt{\left(\frac{1}{p}\sum_{k=1}^{p}\left(v_k\right)^2\right) - \left(\frac{1}{p}\sum_{k=1}^{p}\left(v_k\right)\right)^2} \quad ; \quad v_k = \frac{\left[\overline{\widehat{Q}_{\mathrm{SP}[k]}}\right]_{\theta_{\mathrm{SP}[i]}} - \left[\overline{\widehat{Q}_{\mathrm{SP}[k]}}\right]_{\theta_{\mathrm{TP}}}}{\overline{Q_{\mathrm{SP}[k]}}} \tag{5}$$

with the same notations as in Eq. (3).

As opposed to the previous one, this criterion is not informative on the transferability level of a model, but measures the degree of "parallelism" between two ω_θ curves. It takes values between 0 (situation where the shapes of the $\omega_{\theta_{\mathrm{SP}[i]}}$ and $\omega_{\theta_{\mathrm{TP}}}$ curves are rigorously identical) and $+\infty$. We note that, by construction, the mean flow volume error over the entire record ($[\overline{\widehat{Q}_{\mathrm{TP}}}]_{\theta_{\mathrm{SP}[i]}}/\overline{Q_{\mathrm{TP}}}$) has no impact on this second criterion. In other words, only the shape similarities between the ω_θ curves are analysed, while their vertical spacing is left out of consideration.

This measure of similarity was then normalised by the magnitude of volume error variations ($\sigma[\omega_{\theta_{\mathrm{TP}}}]$) to build a non-dimensional criterion (ρ_i), given in Eq. (6). In a way, ρ_i is a "noise-to-signal ratio" which highlights how similar ω_θ curves are.

$$\rho_i = \frac{\sigma\left[\omega_{\theta_{\mathrm{SP}[i]}} - \omega_{\theta_{\mathrm{TP}}}\right]}{\sigma\left[\omega_{\theta_{\mathrm{TP}}}\right]} \tag{6}$$

Similarly, a criterion was built for inter-model comparisons where the "degree of parallelism" on volume error variations is measured between two models (M_1 and M_2), both calibrated over

the entire time series. Noted $\rho'_{M_1 M_2}$, this ratio, is described in Eq. (7) and corresponds to the comparison between different $\omega_{\theta_{\text{TP}}}$ curves. The choice for the model serving as reference, whose corresponding $\sigma[\omega_{\theta_{\text{TP}}}]$ constitutes the denominator, is made arbitrarily.

$$\rho'_{M_1 M_2} = \frac{\sigma\left[\omega_{\theta_{\text{TP}}}^{M_2} - \omega_{\theta_{\text{TP}}}^{M_1}\right]}{\sigma\left[\omega_{\theta_{\text{TP}}}^{M_1}\right]} \tag{7}$$

As for $\sigma[\omega_{\theta_{\text{SP}[i]}} - \omega_{\theta_{\text{TP}}}]$, the criteria detailed in Eqs. (6) and (7) range between 0 and $+\infty$. The smaller the ρ_i value, the stronger the similarities between the $\omega_{\theta_{\text{SP}[i]}}$ and $\omega_{\theta_{\text{TP}}}$ curves for the model considered. Similarly, the smaller the $\rho'_{M_1 M_2}$ value, the stronger the similarities between the $\omega_{\theta_{\text{TP}}}$ curves from the models compared (M_1 and M_2).

4 Results

4.1 Case studies – graphical analyses on three catchments

The graphical procedure illustrated in Fig. 4 was applied to the 20 catchments and three hydrological models described in Sect. 2.2 (the 1-parameter Mouelhi formula, the 6-parameter GR4J-CemaNeige model and the 19-parameter Cequeau model). Examples of results are given in Fig. 5 for three catchments: the Ubaye River at Barcelonnette (540 km^2, case study 1), the Lot River at Barnassac (1160 km^2, case study 2) and the Drac River at Pont de la Guinguette (510 km^2, case study 3). This figure is composed of 12 graphs, where the results obtained on the same catchment are in columns, while data and simulations with the same model are in rows. In all cases, we plotted the 10-yr moving average of the variables considered. For each graph showing simulation results, the grey curves correspond to the sub-period calibration procedure previously introduced (see Figs. 3 and 4), while the single black curve corresponds to the calibration over the entire record.

The graphs from Fig. 5 provide useful elements that help determine the impact of the calibration period on model robustness.

First of all, let us analyse each graph independently. The "parallelism effect" observed in Fig. 4 is again visible here. Indeed, the model calibration on different sub-periods lead to errors on 10-yr-mean flows, which vary similarly over time (cf. similarly shaped grey $\omega_{\theta_{SP}}$ curves on graphs 5d to 5l). Concerning the cases where parameter sets were optimised on the full record, the corresponding $\omega_{\theta_{TP}}$ curves are (as expected) not randomly vertically placed. Logically the mean flow volume ratio of the entire period remains close to 1. However, we surprisingly did not obtained flatter ω_θ curves (cf. black curves on graphs 5d to 5l). This shows that even when they are calibrated over the full records, the models tested are unable to provide a better simulation of 10-yr-mean flows than when only a small part of the information is used for parameters optimisation.

Secondly, we observe different behaviours depending on the catchment considered. On some catchments, temporal variations are clearly visible on model volume errors, with amplitudes often around 20 %. This is the case for the Ubaye River at Barcelonnette (already discussed) but also for the Lot River at Barnassac (Fig. 5, case study 2), where an increasing trend is observed on the mean flow volume error (from underestimation to overestimation). Conversely, these errors are almost invariant on other catchments, for example the Drac River at Pont de la Guinguette (Fig. 5, case study 3). Explaining why these errors occur is complex. Some causal links may be inferred from these examples, related to changes in climate forcings (e.g. changes in mean air temperature for the Lot River). However, our recent investigations on this topic showed that if such correlations can be establish in numerous cases, there are not systematic and their significance greatly varies from one catchment to another (Coron, 2013). To date, we remain unable to draw general conclusions regarding the spatial similarities in model volume error variations and can only acknowledge for the need to further investigate this complex question.

Additionally, on these three illustrative examples, we note that the available period for analysis is shorter for the Drac River than for the other two catchments, but the magnitude of the changes on observed data (precipitation, temperature, discharges) is similar for the three catchments over the common period. Therefore, the smaller range of volume error variations obtained for the Drac River catchment truly reflects better model performance in this case.

From these comparisons, we note that the greater the amplitude of volume error variations, the more vertically spaced the $\omega_{\theta_{\mathrm{SP}}}$ curves are on these graphs. This is a consequence of the calibration criterion used (KGE), where volume errors are explicitly targeted. The various $\omega_{\theta_{\mathrm{SP}[i]}}$ curves are indeed "positioned" to ensure $\overline{[\widehat{Q}_{\mathrm{SP}[k]}]}_{\theta_{\mathrm{SP}[i=k]}} / \overline{Q_{\mathrm{SP}[k]}} \approx 1$. When the sub-period used for calibration corresponds to a lower or upper extreme of the $\omega_{\theta_{\mathrm{SP}}}$ curves, it is "vertically positioned" above or below the other ω_θ curves, respectively. This can be seen in Fig. 4, with the curves whose corresponding calibration sub-periods are 10/1968-09/1978 and 10/1981-09/1991. Likewise, for catchments where model errors on mean flow volumes are almost time-invariant, all $\omega_{\theta_{\mathrm{SP}}}$ curves are nearly flat and thus superimposed.

Thirdly, the graphs placed in columns (Fig. 5) show strong similarities, indicating similar behaviours of the three models tested on each catchment. The $\omega_{\theta_{\mathrm{TP}}}$ curve shapes (and indirectly the $\omega_{\theta_{\mathrm{SP}}}$ curve shapes) are not strictly identical between the three models. Still, the overall shapes of the 10-yr moving average curves look alike, in spite of the large differences in complexity between the models used (structure, time step, number of optimised parameters).

[INSERT FIGURE 5]

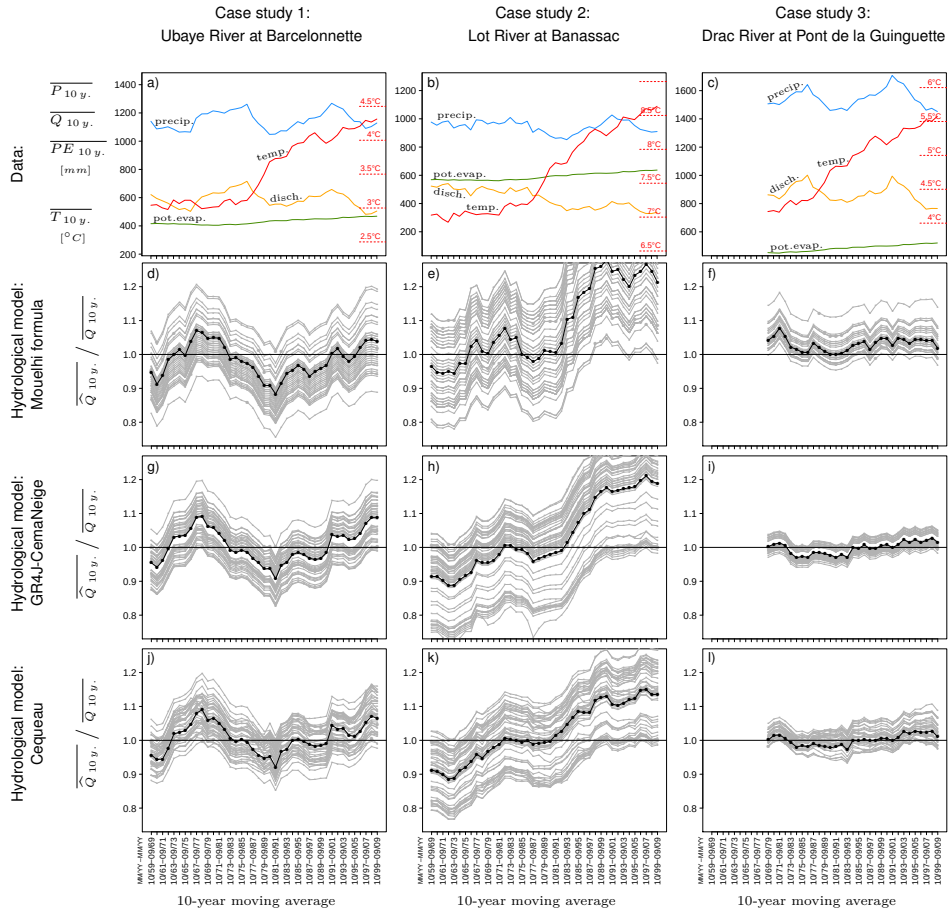

Fig. 5. Examples of behavioural similarities observed on three catchments with the three models tested (for **d** to **l**, the various $\omega_{\theta_{\mathrm{SP}[i]}}$ curves are in grey and the single $\omega_{\theta_{\mathrm{TP}}}$ curve is in black).

4.2 Generalisation of the results (three models over 20 catchments)

The criteria introduced in Sect. 3.4 were used to measure these behavioural similarities systematically over a larger number of tests: we tested the three models over 20 catchments (see characteristics in Sect. 2.1).

First, we computed the standard deviation on the $\omega_{\theta_{\mathrm{TP}}}$ curves, which measures the scale of the volume error variations with time (see Eq. 4). These results are summarised in Fig. 6. For each model, the boxplot provides the 5th, 25th, 50th, 75th and 95th percentile values of the $\sigma[\omega_{\theta_{\mathrm{TP}}}]$ distribution over the catchment set (one value per catchment). Relatively similar medians are obtained for all three models, with values around 4 %. Yet, small differences can be noted between the distributions. The distributions obtained for the Mouelhi formula and GR4J-CemaNeige model are almost identical, and differ from the results obtained with the Cequeau model, whose errors on simulated mean flows are less variable in time (as shown by smaller $\sigma[\omega_{\theta_{\mathrm{TP}}}]$ values). Therefore, it seems that Cequeau is slightly more robust than the other two models, at least with regard to its ability to simulate water balances simultaneously on various periods. The small number of available points (20) limits the possibilities to perform relevant statistical tests to confirm these qualitative assessments. However, we can note that these results are in accordance with the model efficiencies presented in Table 2, the Mouelhi formula and Cequeau being, on average, the worst and best performing models during the transferability tests on the catchment set, respectively. Possible explanations for this might be the differences in structural complexity (in terms of conceptualisation, parametrisation and spatial distribution). Other reasons for Cequeau's better robustness might be related to the different ways snow storage and PE data are computed, but further tests focused on these aspects are necessary to provide a better understanding of these differences.

The ρ_i ratio was then used to measure the significance of behavioural similarities on these volume errors over the catchment set (see Eq. 6). We remind that only "relative" variations are considered in this criterion and the overall volume error (i.e. the ω_θ curves' vertical positioning) is not measured. The "parallelism imperfections" between various ω_θ curves are compared to the scale of the temporal variations of volume errors shown in Fig. 6. Since numerous sub-

period calibrations were made for each catchment, a large number of ρ_i can be computed over the 20 catchments considered. The distributions of the values obtained for each model are given in Fig. 7, using a boxplot representation (5th, 25th, 50th, 75th and 95th percentiles).

Values of ρ_i obtained for the Mouelhi formula and GR4J-CemaNeige model are small, with more than 95 % of them smaller than 0.25. The median value of 0.1 means that, on average and for both models, the "parallelism imperfections" between ω_θ curves (i.e. the "noise") are 10 times smaller than the temporal variations observed (i.e. the "signal"). The results are different for the Cequeau model but the values obtained remain small: the median is around 0.3 and 75 % of them are smaller than 0.5 (value for which the noise's significance is half the signal's). Because the reference $\omega_{\theta_{\text{TP}}}$ curves differ between models, we must add that any inter-model comparison based on Fig. 7 should be analysed together with the distributions shown in Fig. 6. However, the smaller $\sigma[\omega_{\theta_{\text{TP}}}]$ values obtained with Cequeau in some cases are likely not the only explanation for the greater ρ_i values observed. They may also result from the larger differences between ω_θ curves with this model (see Fig. 5 for examples of "parallelism imperfections"). The reasons for these greater differences could stem from Cequeau's greater complexity compared to the Mouelhi formula and GR4J-CemaNeige. Because a larger number of parameters had to be optimised, some 10-yr-long sub-periods may not have been informative enough to allow their optimisation. This could explain the fewer similarities between ω_θ trajectories.

[INSERT FIGURE 6]

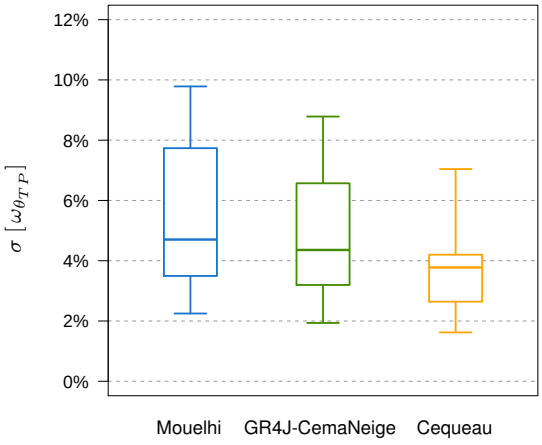

Fig. 6. Standard deviations of the 10-yr mean flow volume errors obtained during calibration over the full record (distribution for each model over 20 catchments).

[INSERT FIGURE 7]

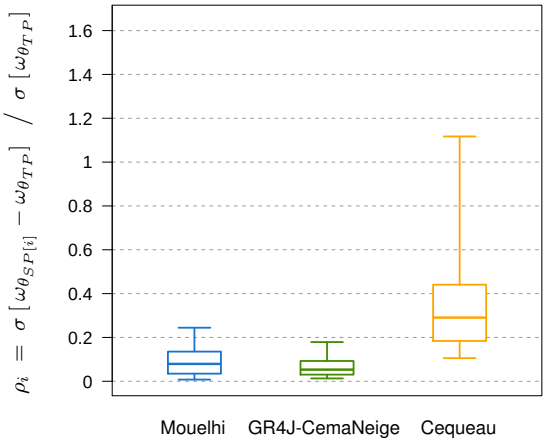

Fig. 7. Behavioural similarities observed between sub-period and full record calibrations in terms of 10-yr mean flow volume errors (distribution for each model over 20 catchments).

4.3 Direct comparison of the three models' behaviours

The issue discussed in this paper has been broken down into three questions (see Sect. 3.3.4). The distributions obtained on the catchment set for the ρ_i criterion are quite informative with respect to the first two questions on the volume error similarities between sub-period and total-period calibration for each model over different catchments. Analysing the distributions of $\rho'_{M_1 M_2}$ should provide insights into the question of inter-model similarities.

For each catchment, we consider the simulations obtained with the models for a full-record calibration. The three corresponding $\omega_{\theta_{\mathrm{TP}}}$ curves (one per model) are compared through a ratio of standard deviation similar to ρ_i (see Eqs. 6 and 7). $\rho'_{M_1 M_2}$ values can be interpreted like the ρ_i values. These distributions are presented in Fig. 8, where two pairs of comparisons are made

depending on the model used as a reference for $\rho'_{M_1 M_2}$ computations (here, either the simplest or the most complex of the three models is used as *M1*).

In the vast majority of situations, the values taken by $\rho'_{M_1 M_2}$ are below 1, with median values ranging from 0.4 to 0.8. It shows that behavioural similarities exist between different models and that the scale of the differences remains smaller than the scale of temporal variations of the 10-yr-mean flow volume errors (1.25 to 2.5 times smaller on average). $\rho'_{M_1 M_2}$ values are higher when the Cequeau model is used as a reference than when the Mouelhi formula plays this role (see right versus left parts of Fig. 7), likely because Cequeau is more robust on the catchment set (see higher KGE in Table 2 and lower $\sigma[\omega_{\theta_{\mathrm{TP}}}]$ in Fig. 6).

Differences on mean flow volume errors could be expected from a change of hydrological model, especially considering the large complexity gaps between the model structures used here. Nevertheless, it is surprising that they remain limited, although the shape similarities between $\omega^M_{\theta_{\mathrm{TP}}}$ curves are not as strong as the ones between $\omega_{\theta_{\mathrm{SP}}}$ curves (see Fig. 7 vs. Fig. 6).

[INSERT FIGURE 8]

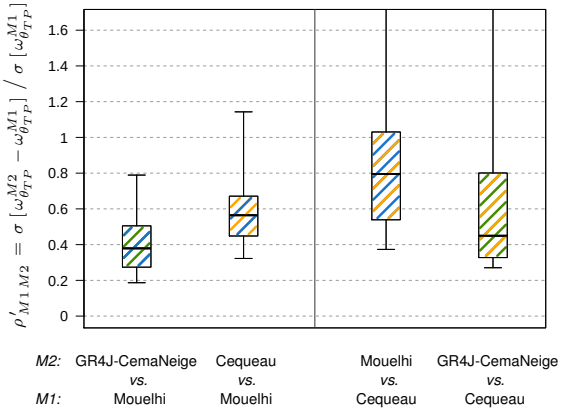

Fig. 8. Behavioural similarities observed between different models in terms of 10-yr mean flow volume errors. Calibrations over the full record (distributions over 20 catchments).

4.4 Alternative graphical representation

We have shown the existence of a "parallelism effect" in the previous evaluation of the models' ability to reproduce water balance over time. The behavioural similarities observed in our tests can be viewed in another (maybe simpler) way.

Let us start again with the sub-periods built for each catchment using a 10-yr-long sliding window. For each catchment, we considered all possible pairs of sub-periods A and B and we compared the relative changes in mean flow, either observed or simulated. Because they are expressed in a relative way (e.g. $\triangle\overline{Q}_{[A/B]} = \overline{Q_{\mathrm{SP}[A]}}/\overline{Q_{\mathrm{SP}[B]}}$), values from different sub-period pairs and different catchments can be analysed together. For each pair (A and B), we computed the $\triangle\overline{Q}_{[A/B]}$ observed and the various $\triangle\overline{\widehat{Q}}_{[A/B]}$ simulated using the parameter set optimised

over the full record (θ_{TP}) and the numerous parameter sets ($\theta_{\text{SP}[i]}$) obtained from the sub-period calibrations (see Fig. 3). These changes were then used as coordinates to build large scatterplots.

Comparing observed and simulated changes provides information on the models' ability to reproduce the variations in water balance equilibrium over different periods. We only considered here the parameter set obtained from the calibration on the entire record and therefore compared $\left[\triangle\overline{\widehat{Q}}_{[A/B]}\right]_{\theta_{\text{TP}}}$ with $\triangle\overline{Q}_{[A/B]}$. Aggregated over the 20 catchments, the results of these comparisons are given in Fig. 9a–c for the three models considered in this study. To extract the information contained in the graphs, the point clouds are divided into vertical slices and the distributions of $\left[\triangle\overline{\widehat{Q}}_{[A/B]}\right]_{\theta_{\text{TP}}}$ values are summarised by boxplots (showing the 5th, 25th, 50th, 75th and 95th percentiles). We see how these models face difficulties to reproduce the climate elasticity of 10-yr-mean flows, i.e. larger changes are underestimated, be they positive or negative. Cequeau shows the best ability and the Mouelhi formula the worst, which is in accordance with the $\sigma[\omega_{\theta_{\text{TP}}}]$ previously obtained (see Fig. 6).

Comparing mean flow changes simulated by the same model but with different parameter sets reveals how the choice of the calibration period affects the model outputs. Every $\theta_{SP[i]}$ parameter set was considered together with the θ_{TP}. The corresponding simulations were analysed to extract $\left[\triangle\overline{\widehat{Q}}_{[A/B]}\right]_{\theta_{\text{SP}[i]}}$ and $\left[\triangle\overline{\widehat{Q}}_{[A/B]}\right]_{\theta_{\text{TP}}}$ for all the sub-period pairs (A and B). These values were used as coordinates to build clouds of points and aggregated over the 20 catchments. The corresponding results are given in Fig. 9d–f. These graphical representations provide another way to measure behavioural similarities on medium-term volume errors between sub-period and total-period calibration. The conclusions inferred from Fig. 7 are confirmed. The choice of the calibration period has very little impact on the simulated changes of 10-yr-mean flow between periods. Similarities are the strongest for the Mouelhi formula and the GR4J-CemaNeige model, with an R^2 coefficient of 0.997 (Pearson coefficient). For the Cequeau model, a larger number of cases where simulated changes are different between sub-period and total-period calibrations can be seen. Nevertheless, behavioural similarities remain strong on average over the 20 catchments, with an R^2 coefficient around 0.95.

[INSERT FIGURE 9]

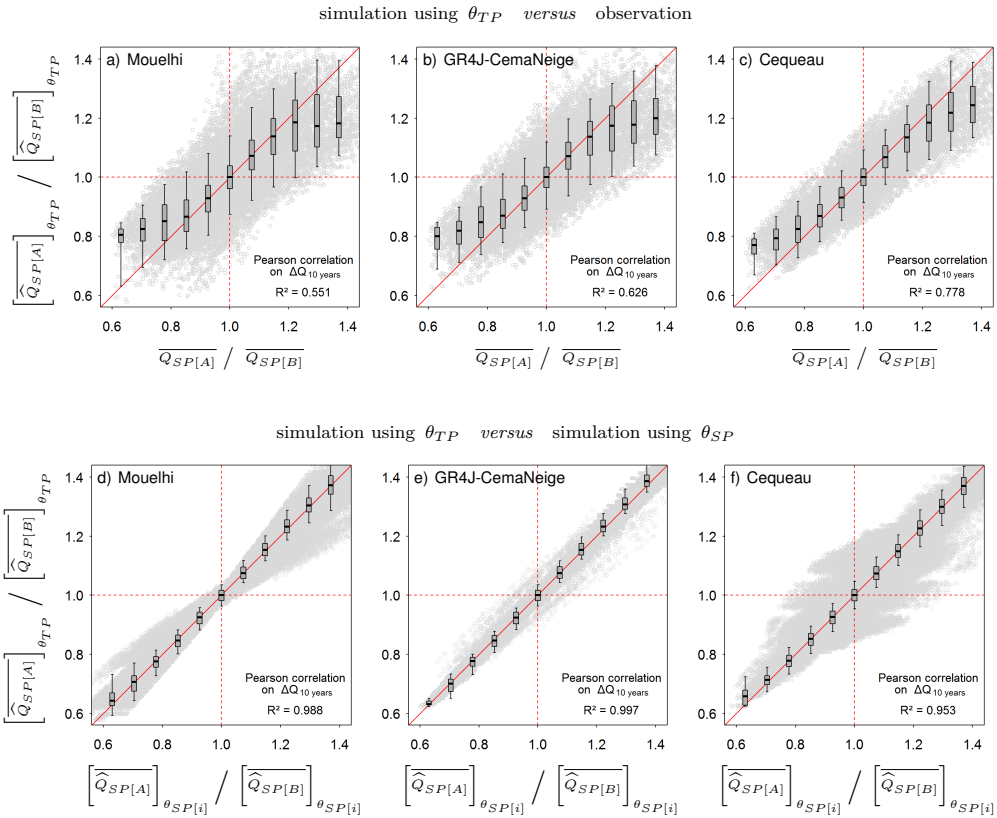

Fig. 9. Comparisons of relative changes in 10-yr-mean flow, observed and simulated (aggregation of results from 20 catchments, considering any possible pair of 10-yr sub-periods A and B).

4.5 Possible implications for climate change impact studies

The models' behaviours highlighted throughout this work are quite remarkable. If a study was to be conducted on the impact of the calibration period over the 10-yr-mean flow volume errors, we would probably rate the uncertainties as "high" for some catchments. Indeed, for a catchment where the ω_θ curves are not flat, choosing one calibration period or another determines the vertical positioning of the corresponding curve, which impacts the absolute errors on every sub-period taken independently (see Fig. 4, for example). However, when the simulated 10-yr-mean flows are expressed relatively to the 10-yr-mean flow simulated during calibration, the same analysis would conclude that these uncertainties are "low", especially for the Mouelhi formula and GR4J-CemaNeige model (as shown in Figs. 7 and 9). People who are both optimistic and familiar with climate change impact studies might see this as good news, because it advocates for the validity of the delta-change approach used to present changes in hydrological simulations, in which it is hypothesised that the mean flow volume error remains constant. Yet, this is not entirely satisfactory and we would strongly prefer to understand and thus avoid these parameter transferability problems from the start.

5 Discussion

Series of simulations from three models calibrated on different periods have been compared in this work. Differences were expected between their accuracy regarding the simulation of water balances. However, it was surprising to see how limited these differences were in practice on the catchment set used here (see results of similarity measurements in Sect. 4). Yet, we must acknowledge that after these tests we still do not know whether the three models share the same deficiency or suffer from the same external factors.

As a result, this work may appear incomplete to some readers who expected more explanations or even solutions to the modelling deficiencies presented here. We agree that the diagnosis should ideally be followed by solutions, but our attempts to diagnose these problems, includ-

ing analyses of model parameters, remained unsuccessful. The possible causes for the lack of temporal robustness are numerous and hard to distinguish from one another.

5.1 Robustness and modelling choices

The role of inappropriate model structure must of course be questioned regarding robustness problems. For instance, Hartmann et al. (2013) showed the need for adaptation of a model structure to ground realities in karstic zones. Simple or complex approaches can be used to investigate this question. For several examples, see Butts et al. (2004), Bulygina and Gupta (2009), Reusser and Zehe (2011), Lin and Beck (2012) and Seiller et al. (2012). Here, we investigated this issue through a comparison between three models of increasing complexity. The results suggest that the structures of all three models may not be suitable to allow for water balance adjustments simultaneously on various periods. This comparison could be extended to other model structures, although a relatively large complexity range was considered: from an annual 1-parameter formula to a semi-distributed daily model with 19 optimised parameters.

Problems of miscalibration or overcalibration of model parameters may also cause robustness problems. A review of the authors discussing this issue in hydrology includes Wagener et al. (2003), Hartmann and Bárdossy (2005), Son and Sivapalan (2007), Bai et al. (2009), Gupta et al. (2009), de Vos et al. (2010), Ebtehaj et al. (2010), Efstratiadis and Koutsoyiannis (2010), Pechlivanidis et al. (2010), Zhang et al. (2011), Andréassian et al. (2012), Gharari et al. (2013) and Zhan et al. (2013). Some of these studies present new calibration criteria better balancing the weight of different error types (e.g. wrong volume, wrong variability, etc.). Other studies propose optimisation strategies involving multi-period calibration, these sub-periods being selected according to their relevance with respect to the calibration objectives (e.g. informative content, hydro-climatic characteristics, etc.). For the work reported here, different calibration criteria were tested, including the well-known NSE and a modified KGE where the weight of volume error within the formula was reduced. We also attempted to calibrate the GR4J-CemaNeige model on the total records with the exclusive aim of minimising the standard deviation on the 10-yr-mean flow volume errors ($\sigma[\omega_{\theta_{\mathrm{TP}}}]$). None of these criteria could significantly reduce the robustness problems observed in this study.

Other tests could be made to determine the potential impact of the sub-period length in the calibration procedure. However, as we have shown in this paper, a significant part of the efficiency loss during parameters transfer is caused by the models' difficulties to reproduce mean flow volumes on the calibration period and other periods simultaneously. Increasing the sub-period length in our procedure mechanically decreases the contrast between the conditions under which the model is tested. Indeed, whichever the causes of the robustness issues are (e.g. changes in measurement biases, changes in climatic conditions). Although smaller contrast may lead to smaller efficiency loss during the transfer tests, the corresponding flattening of the ω_θ curves nonetheless remains a mechanical effect, similar to changing the lens of a magnifying glass. The absence of true improvement from using a longer calibration period was proved in our work when parameter sets optimised on the full records were used. Indeed, we showed how these θ_{TP} could not allow a reduction in the mean flow volume error variations (see Fig. 7). Concerning now the impact of reducing the sub-period length, it is logically different. Indeed, below a certain length, the parameters would be optimised on insufficiently informative periods, therefore causing a drop in the model efficiencies during validation.

In spite of these various calibration criteria tested and the relatively large range of model complexity considered in this study, further investigations are still necessary to confirm the deficiencies reported in this paper regarding mean flow volume simulation. Such investigations should extend both testing on model structure and calibration strategy. While they may conclude on the sole responsibility of the conceptualisation process, it remains impossible at the moment to determine with certainty the causes for transferability issues. All potential causes must therefore be considered.

5.2 Robustness and data quality

The level of achievable modelling performances surely depends on the model used but also on the quality of the data it is fed with. Errors may occur during the measurements recording or their post-processing (e.g. aggregation, interpolation, etc.). Depending on the error type they may have a negative impact on the modelling performances, which must be considered (Oudin et al., 2006; McMillan et al., 2010, 2011). If these errors vary temporally, they will induce poor

temporal transferability of model parameters. This can for instance be the case when the measurement techniques are changed or when the sensor network evolves. This may also indirectly result from vegetation growth or changing climatic conditions if they impact the biases on model input estimates. In the case of hydrological modelling, the incorrect estimation of discharges, precipitation and evapotranspiration fluxes may explain temporal robustness problems.

For the work presented here, we remind that precipitation, temperature and discharge series could be considered to be of high quality. Yet, we performed additional quality checks using visual inspection and double mass curves comparisons with neighbouring stations. In spite of these verifications, the contribution of the data to model robustness issues is hard to exclude with certainty. Among the potential input errors, particular attention should be given to the estimation of evapotranspiration. Uncertainties are indeed associated with the computation of potential evapotranspiration (PE) as well as actual evapotranspiration (AE), which depends on the later. Evapotranspiration is an important part of the water balance and it may not be adequately estimated in the context of a changing climate, depending on the approach used (Donohue et al., 2010; Milly and Dunne, 2011; Herrnegger et al., 2012). In an attempt to investigate the potential contribution of PE estimates on our modelling results, we performed complementary tests using the Penman–Monteith formula (instead of Oudin's) to feed the Mouelhi formula and the GR4J-CemaNeige model (Monteith, 1965; Oudin et al., 2005). The corresponding variations on 10-yr mean volume errors were neither better nor exactly similar to those shown here. Therefore, we could not exclude a potential role of the PE and AE computational choices on the models' robustness deficiencies and we can only acknowledge for the strong need for further work on this question. Among the potential directions for further research, we could mention the need to test multiple formula to compute PE, experiment various modelling strategies to estimate AE from PE and soil moisture conditions, or compare modelled AE with other AE estimates (e.g. from lysimeter or flux stations).

Finally, investigations on the spatial similarities of model volume errors can help assess the role of data quality issues on models robustness issues. Indeed, strong dissimilarities between the volume error curves of different catchments, in spite of their common characteristics, may be caused by time variant errors in discharge measurements. Conversely, similarly shaped volume

error curves of neighbouring catchments may be obtained as a result of inaccurate regional estimates for the model's input forcings (e.g. if the bias on precipitation estimates evolves in time or if the method used for PE computation is inappropriate).

5.3 Robustness and changes in catchment functioning

Although poor modelling strategies or data quality are likely to be the major sources for model failure, other explanations are worth considering. Working on an (until then) unexplained overestimation of the Meuse River runoff between 1930 and 1965, Fenicia et al. (2009) showed the major impact of changes in land use management and forest age on the catchment's functioning. Such temporary or permanent changes of a catchment functioning result in significant model robustness problems if not included in the modelling framework. While limited human impacts on the water balances are expected for the 20 catchments used in this study, we agree that these impacts may be hard to quantify in practice (Andréassian, 2002). Besides, human activities are not the only source for changes in the rainfall–runoff relationship, which may also result from natural events. For example, Chiew et al. (2013) discussed how the "Millennium drought" reduced the surface-groundwater connection in south-eastern Australia, thus dramatically modifying the dominant hydrological processes. Although this example relates to an extreme event, we believe that, in the context of global climate change, such explanations must not be underrated when analysing models' temporal robustness.

6 Summary and conclusions

The purpose of this paper was to propose tools to help diagnose the robustness of rainfall-runoff models, regarding their ability to reproduce water balances simultaneously on different temporal periods. A comparison framework was implemented over 20 mountainous catchments in France using three models of increasing complexity: the annual Mouelhi formula, the daily-lumped GR4J-CemaNeige model and the daily semi-distributed Cequeau.

The results show that failure situations are common when models are evaluated on long records. When temporal transferability posed problems, choosing another calibration sub-period induced no significant difference on the 10-yr-mean flow volume errors. Indeed, when we considered two temporal periods A and B, the $\overline{\widehat{Q}_A}/\overline{\widehat{Q}_B}$ ratio remained stable regardless of the calibration period, even when the full record was used to optimise model parameters. This reveals that the lack of robustness identified for some catchments on 10-yr-mean flows is not caused by a poor choice of calibration period but rather stems from the models' overall inability to reproduce water balances simultaneously on different sub-periods (considering their usage conditions: input data sets, modelling choices, etc.).

The three models tested in this study showed strong similarities in their (in)ability to simulate water balances. Some differences exist but they are smaller than expected with regards to the large differences in complexity level between the tested models. At this stage, however, we cannot conclude whether these three models share the same deficiency or suffer from the same external causes, related to input data estimation for example. It is indeed difficult to apportion blame between the potential explanations for robustness problems, which remain numerous: ineffective model structure, inappropriate calibration strategy as well as temporal changes in input errors, the catchments' natural functioning or anthropogenic impact.

The present study differs from previous works in that we highlighted strong behavioural similarities between different model structures and calibration periods. We used simple but relevant graphical and numerical tools to show how limited the impact of a model's complexity or calibration period can be regarding its capacity to reproduce the temporal variations in water budget equilibrium. In agreement with the participants at the "Court of Miracles of Hydrology" workshop (Perrin and Andréassian, 2010), we believe that modelling failures should be seen positively as challenges and can be substantial sources of information on model imperfections and catchment functioning. This study showed that blaming the excessively short calibration period or the overly simplistic structure without a more detailed examination is not necessarily the best option when discussing temporal robustness in hydrological modelling. In order to progress on this issue, advances are needed on both the quantification of medium-term water

exchanges at the catchment scale and the way these exchanges can be modelled to account for temporal variations.

Appendix A

The procedure presented in this paper has been applied over a larger catchment set for the Mouelhi formula and GR4J-CemaNeige model. This set is composed of 365 French catchments, whose locations and properties are summarised in Fig. A1 and Table A1.

These additional results are in accordance with those exposed in the article. The difficulties for the Mouelhi formula and GR4J-CemaNeige model to reproduce water balances simultaneously on different temporal periods were confirmed. The "parallelism effect" observed during the study of volume errors variations for these models was again visible on this much larger catchment set (see Figs. A2 and A3). Our findings that $\omega_{\theta_{\mathrm{SP}}}$ and $\omega_{\theta_{\mathrm{TP}}}$ curve have similar shapes were reproduced on this new set for both models. This is shown in Fig. A2b by the low ρ_i values, whose distributions are similar to the one obtained for the 20 catchment set. This can also be seen in Fig. A3, where the ratio $\overline{\widehat{Q}_A}/\overline{\widehat{Q}_B}$ remains very stable regardless the calibration period (where A and B are 10-yr-long temporal periods, see Sect. 4.4). Indeed, the Pearson correlation coefficient (R^2) between simulated changes are equivalent when results are aggregated over the 20 catchments used in the article or the 365 catchments considered in this appendix.

[INSERT FIGURE A01]

[INSERT TABLE A01]

[INSERT TABLE A02]

[INSERT FIGURE A02]

[INSERT FIGURE A03]

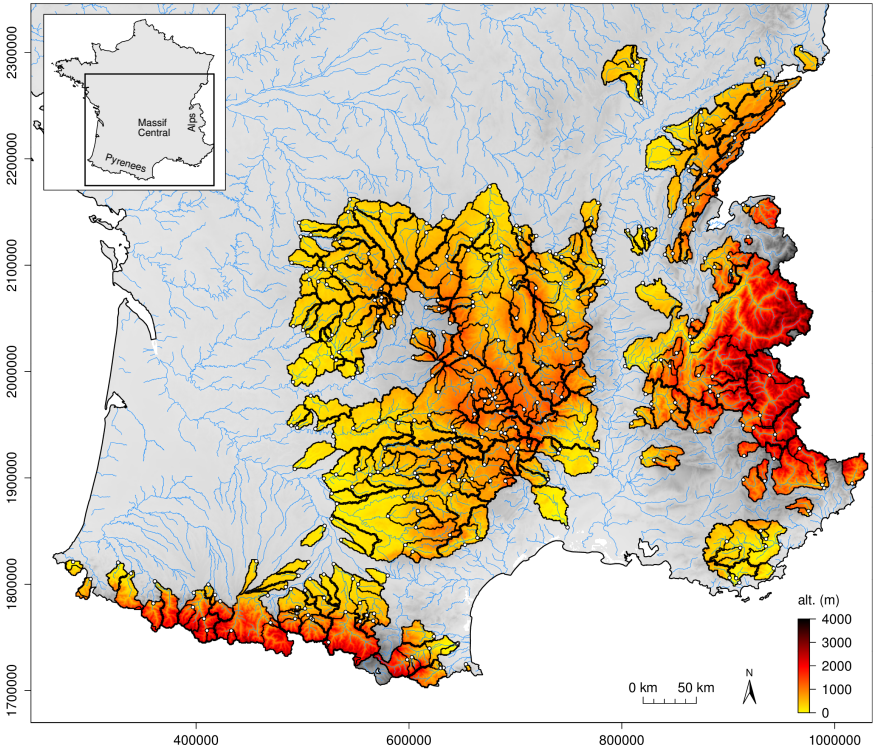

Fig. A1. Locations of the 365 catchments used in the additional testing with the Mouelhi formula and GR4J-CemaNeige model.

Table A1. Characteristics of the enlarged catchment set used in the additional testing (365 catchments).

	5th centile	25th centile	median	75th centile	95th centile
Catchment surface [km^2]	34	100	220	590	2510
Mean elevation [m]	260	490	750	1070	1660
Mean annual total precip. (P) [mm]	850	990	1160	1440	1860
P_{solid}/P ratio ratio (annual mean) [$-$]	2 %	3 %	7 %	13 %	30 %
Mean annual pot. evap. PE$_{(\text{Oudin})}$ [mm]	500	560	630	680	770
Mean annual discharge (Q) [mm]	220	370	540	880	1410
P/PE ratio (annual mean) [$-$]	1.15	1.49	1.85	2.46	3.52
Q/P ratio (annual mean) [$-$]	0.23	0.36	0.47	0.60	0.84
Available time series length [yr]	33	40	43	52	62

Table A2. Model efficiencies computed over the total available records, considering sub-period calibrated parameter sets: $[\text{KGE}_{\text{TP}}]_{\theta_{\text{SP}}}$
(results for the enlarged catchment set).

		5th centile	25th centile	median	75th centile	95th centile
KGE at the annual time step	Mouelhi	0.301	0.541	0.687	0.782	0.897
	GR4J-CemaNeige	0.649	0.774	0.842	0.893	0.937
KGE at the daily time step	GR4J-CemaNeige	0.704	0.810	0.860	0.897	0.931

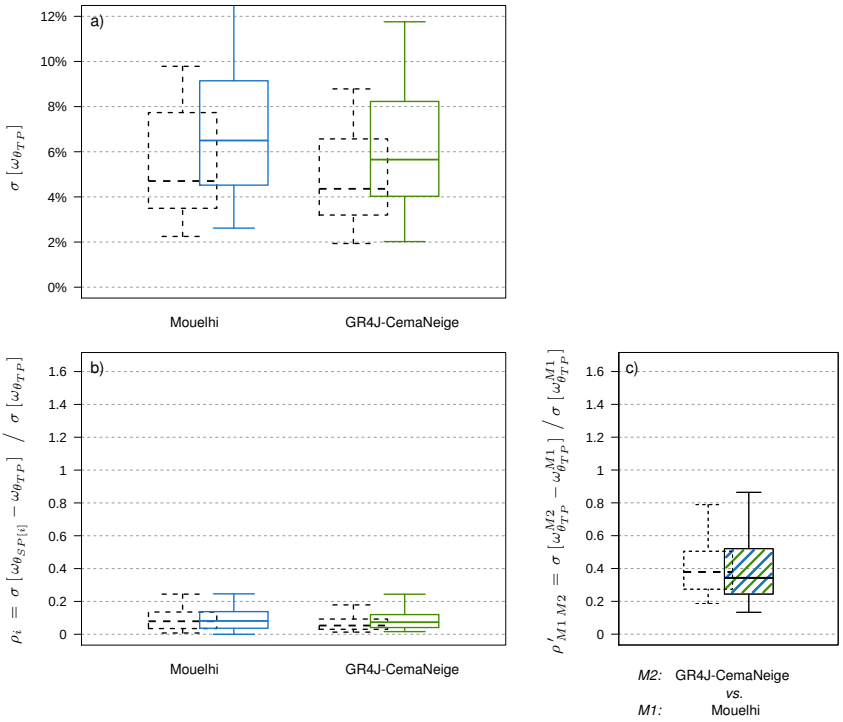

Fig. A2. Distributions of $\sigma[\omega_{\theta_{\mathrm{TP}}}]$, ρ_i and $\rho'_{M_1 M_2}$ values obtained for the set of 365 catchments (solid coloured lines) and comparison with the previous results obtained on 20 catchments (dashed black lines).

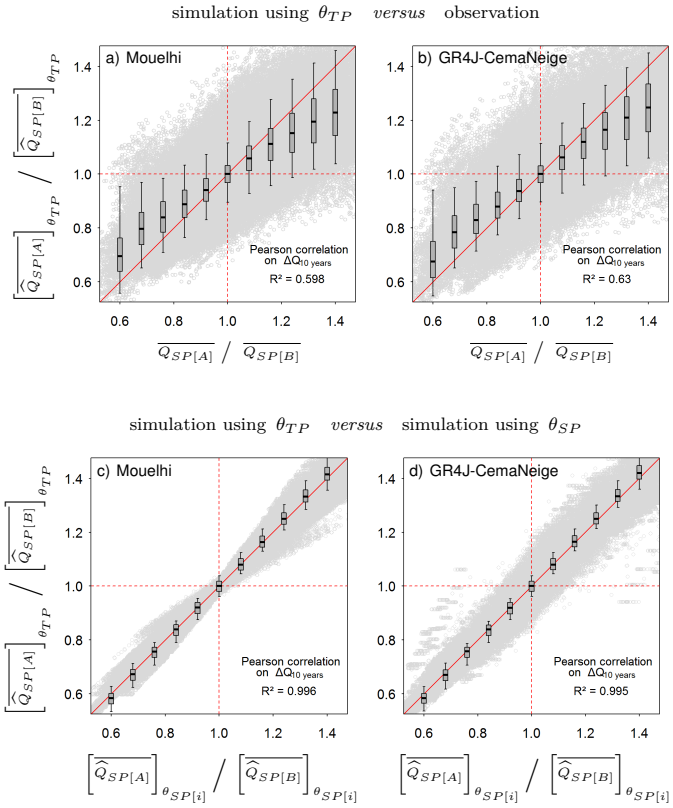

Fig. A3. Comparisons of relative changes in 10-yr-mean flow, observed and simulated (aggregation of results from 365 catchments, considering any possible pair of 10-yr sub-periods A and B).

Acknowledgements. The authors would like to thank EDF R & D LNHE and Irstea HBAN (France) for supporting this study and providing the datasets used in this work. We also like to thank Ilias Pechlivanidis and two anonymous reviewers for their relevant and constructive criticism, which helped to improve the quality of the manuscript.

References

Andréassian, V.: Impact de l'évolution du couvert forestier sur le comportement hydrologique des bassins versants, PhD thesis, UPMC, Paris, France, pp. 262, 2002.

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

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.-H., Oudin, L., Mathevet, T., Lerat, J., and Berthet, L.: All that glitters is not gold: the case of calibrating hydrological models, Hydrological Processes, 26, 2206–2210, doi:10.1002/hyp.9264, 2012.

Bai, Y., Wagener, T., and Reed, P.: A top-down framework for watershed model evaluation and selection under uncertainty, Environmental Modelling & Software, 24, 901–916, doi:10.1016/j.envsoft.2008.12.012, 2009.

Bourqui, M., Mathevet, T., Gailhard, J., and Hendrickx, F.: Hydrological validation of statistical downscaling methods applied to climate model projections, in: Hydro-climatology: Variability and change (IUGG2011), vol. 344, pp. 32–38, International Association of Hydrological Sciences, Melbourne, Australia, 2011.

Brigode, P., Oudin, L., and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, Journal of Hydrology, 476, 410–425, doi:10.1016/j.jhydrol.2012.11.012, 2013.

Bulygina, N. and Gupta, H.: Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation, Water Resources Research, 45, W00B13, doi:10.1029/2007WR006749, WOS:000263332500004, 2009.

Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, Journal of Hydrology, 298, 242–266, doi:10.1016/j.jhydrol.2004.03.042, 2004.

Chahinian, N., Andréassian, V., Duan, Q., Fortin, V., Gupta, H., Hogue, T., Mathevet, T., Montanari, A., Moretti, G., Moussa, R., Perrin, C., Schaake, J., Wagener, T., and Xie, Z.: Compilation of the MOPEX 2004 results, in: Large sample basin experiments for hydrological model parameterization, no. 307 in IAHS Red Book Series, pp. 313–338, Andréassian, V., A. Hall, N. Chahinian, J. Schaake, Wallingford, IAHS (red book series n°307) edn., 2006.

Charbonneau, R., Fortin, J., and Morin, G.: The CEQUEAU model: description and examples of its use in problems related to water resource management, Hydrological Sciences Bulletin, 22, 93–202, 1977.

Chiew, F. H. S., Potter, N. J., Vaze, J., Petheram, C., Zhang, L., Teng, J., and Post, D. A.: Observed hydrologic non-stationarity in far south-eastern Australia: implications for modelling and prediction, Stochastic Environmental Research and Risk Assessment, pp. 1–13, doi:10.1007/s00477-013-0755-5, 2013.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, W09301, doi:10.1029/2010WR009827, WOS:000295610200001, 2011.

Coron, L.: Les modèles hydrologiques conceptuels sont-ils robustes face à un climat en évolution ? Diagnostic sur un échantillon de bassins versants français et australiens, PhD thesis, AgroParisTech, Paris, France, pp. 235, http://pastel.archives-ouvertes.fr/pastel-00879090/, 2013.

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 Resources Research, 48, W05552, doi:10.1029/2011WR011721, 2012.

de Vos, N. J., Rientjes, T. H. M., and Gupta, H. V.: Diagnostic evaluation of conceptual rainfall-runoff models using temporal clustering, Hydrological Processes, 24, 2840–2850, doi:10.1002/hyp.7698, 2010.

Donohue, R. J., McVicar, T. R., and Roderick, M. L.: Assessing the ability of potential evaporation formulations to capture the dynamics in evaporative demand within a changing climate, Journal of Hydrology, 386, 186–197, doi:10.1016/j.jhydrol.2010.03.020, 2010.

Ebtehaj, M., Moradkhani, H., and Gupta, H. V.: Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling, Water Resources Research, 46, W07515, doi:10.1029/2009WR007981, 2010.

Edijatno, Nascimento, N. D. O., Yang, X., Makhlouf, Z., and Michel, C.: GR3J: a daily watershed model with three free parameters, Hydrological Sciences Journal, 44, 263–277, doi:10.1080/02626669909492221, 1999.

Efstratiadis, A. and Koutsoyiannis, D.: The multiobjective evolutionary annealing-simplex method and its application in calibrating hydrological models, in: European Geosciences Union General Assembly 2005, Geophysical Research Abstracts, vol. 7, p. 04593, Vienna, Austria, 2005.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multiobjective calibration approaches in hydrological modelling: a review, Hydrological Sciences Journal, 55, 58–78, 2010.

Fenicia, F., Savenije, H. H. G., and Avdeeva, Y.: Anomaly in the rainfall-runoff behaviour of the Meuse catchment. Climate, land-use, or land-use management?, Hydrology and Earth System Sciences, 13, 1727–1737, doi:10.5194/hess-13-1727-2009, 2009.

François, B., Hingray, B., Hendrickx, F., and Creutin, J. D.: Storage water value as a signature of the climatological balance between resource and uses, Hydrology and Earth System Sciences Discussions, 10, 8993–9025, doi:10.5194/hessd-10-8993-2013, 2013.

Gharari, S., Hrachowitz, M., Fenicia, F., and Savenije, H. H. G.: An approach to identify time consistent model parameters: sub-period calibration, Hydrology and Earth System Sciences, 17, 149–161, doi:10.5194/hess-17-149-2013, 2013.

Gottardi, F., Obled, C., Gailhard, J., and Paquet, E.: Statistical reanalysis of precipitation fields based on ground network data and weather patterns: Application over French mountains, Journal of Hydrology, 432–433, 154–167, doi:10.1016/j.jhydrol.2012.02.014, 2012.

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, Journal of Hydrology, 377, 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Hartmann, A., Antonio Barbera, J., Lange, J., Andreo, B., and Weiler, M.: Progress in the hydrologic simulation of time variant recharge areas of karst systems - Exemplified at a karst spring in Southern Spain, Advances in Water Resources, 54, 149–160, doi:10.1016/j.advwatres.2013.01.010, WOS:000317344300011, 2013.

Hartmann, G. and Bárdossy, A.: Investigation of the transferability of hydrological models and a method to improve model calibration, Advances in Geosciences, 5, 83–87, 2005.

Herrnegger, M., Nachtnebel, H.-P., and Haiden, T.: Evapotranspiration in high alpine catchments – an important part of the water balance!, Hydrology Research, 43, 460–475, doi:10.2166/nh.2012.132, 2012.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31, 13–24, doi:10.1080/02626668609491024, 1986.

Koutsoyiannis, D.: Hurst-Kolmogorov Dynamics and Uncertainty, Journal of the American Water Resources Association, 47, 481–495, doi:10.1111/j.1752-1688.2011.00543.x, 2011.

Le Moine, N.: Description de l'algorithme développé pour le calage automatique du modèle Cequeau (rapport intermédiaire de post-doctorat), Tech. rep., UPMC - EDF R&D, Chatou, France, 2009.

Le Moine, N. and Monteil, C.: CEQUEAU - EDF R&D version 5.1.1, Note de principe, Tech. rep., EDF R&D, Chatou, France, 2012.

Lebecherel, L., Andréassian, V., and Perrin, C.: On regionalizing the Turc-Mezentsev water balance formula, Water Resources Research, in press, doi:10.1002/2013WR013575, 2013.

Lin, Z. and Beck, M. B.: Accounting for structural error and uncertainty in a model: An approach based on model parameters as stochastic processes, Environmental Modelling & Software, 27–28, 97–111, doi:10.1016/j.envsoft.2011.08.015, 2012.

Matalas, N.: Comment on the Announced Death of Stationarity, Journal of Water Resources Planning and Management, 138, 311–312, doi:10.1061/(ASCE)WR.1943-5452.0000215, 2012.

Mathevet, T.: Quels modèles pluie-débit globaux au pas de temps horaire? Développements empiriques et comparaison de modèle sur un large échantillon de bassins versants, PhD thesis, ENGREF, Paris, France, pp. 354, 2005.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M.: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, Hydrological Processes, 24, 1270–1284, doi:10.1002/hyp.7587, 2010.

McMillan, H., Jackson, B., Clark, M., Kavetski, D., and Woods, R.: Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, Journal of Hydrology, 400, 83–94, doi:10.1016/j.jhydrol.2011.01.026, 2011.

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

Mezentsev, V.: Du nouveau sur le calcul de l'évaporation totale (Yechio raz o rastchetie srednevo summarnovo isparениia), Meteorologiya i Gidrologiya (Russian Meteorology and Hydrology), 5, 24–26, 1955.

Milly, P. C. D. and Dunne, K. A.: On the Hydrologic Adjustment of Climate-Model Projections: The Potential Pitfall of Potential Evapotranspiration, Earth Interactions, 15, 1–14, doi:10.1175/2010EI363.1, 2011.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J.: Stationarity Is Dead: Whither Water Management?, Science, 319, 573–574, doi:10.1126/science.1151915, 2008.

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., Blöschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S. J., Di Baldassarre, G., Yu, B., Hubert, P., Huang, Y., Schumann, A., Post, D. A., Srinivasan, V., Harman, C., Thompson, S., Rogger, M., Viglione, A., McMillan, H., Characklis, G., Pang, Z., and Belyaev, V.: "Panta Rhei—Everything Flows": Change in hydrology and society—The IAHS Scientific Decade 2013–2022, Hydrological Sciences Journal, 58, 1256–1275, doi:10.1080/02626667.2013.809088, 2013.

Monteith, J.: Evaporation and environment, in: Symposia of the Society for Experimental Biology (in The State and Movement of Water in Living Organisms), vol. 19, p. 205–234, Cambridge University Press, Swansea, Royaume-Uni, 1965.

Mouelhi, S., Michel, C., Perrin, C., and Andréassian, V.: Linking stream flow to rainfall at the annual time step: The Manabe bucket model revisited, Journal of Hydrology, 328, 283–296, doi:10.1016/j.jhydrol.2005.12.022, 2006.

Muñoz, E., Arumí, J. L., and Rivera, D.: Watersheds are not static: Implications of climate variability and hydrologic dynamics in modeling, Bosque (Valdivia), 34, 3–4, doi:10.4067/S0717-92002013000100002, 2013.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I. A discussion of principles, Journal of Hydrology, 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2-Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling, Journal of Hydrology, 303, 290–306, doi:10.1016/j.jhydrol.2004.08.026, 2005.

Oudin, L., Perrin, C., Mathevet, T., Andréassian, V., and Michel, C.: Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, Journal of Hydrology, 320, 62–83, doi:10.1016/j.jhydrol.2005.07.016, 2006.

Pechlivanidis, I., McIntyre, N., and Wheater, H.: Calibration of the semi-distributed PDM rainfall–runoff model in the Upper Lee catchment, UK, Journal of Hydrology, 386, 198–209, doi:10.1016/j.jhydrol.2010.03.022, 2010.

Perrin, C. and Andréassian, V. e.: The Court of Miracles of Hydrology, Hydrological Sciences Journal (Special Issue), 55, 849–1084, 2010.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003.

Reed, P. and Devireddy, D.: Groundwater monitoring design : a case study combining epsilon-dominance archiving and automatic parameterization for the NSGA-II, in: Applications of multi-objective evolutionary algorithms, Advances in natural computation series, vol. 1, pp. 79–100, Carlos A. Coello Coello & Gary B. Lamont (editors), New-York, USA, world scientific edn., 2004.

Reusser, D. E. and Zehe, E.: Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity, Water Resources Research, 47, W07550, doi:10.1029/2010WR009946, 2011.

Rosero, E., Yang, Z.-L., Wagener, T., Gulden, L. E., Yatheendradas, S., and Niu, G.-Y.: Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced versions of the Noah land surface model over transition zones during the warm season, Journal of Geophysical Research, 115, D03106, 21 pp., doi:10.1029/2009JD012035, 2010.

Schaake, J., Duan, Q., Andréassian, V., Franks, S., Hall, A., and Leavesley, G.: The model parameter estimation experiment (MOPEX) - Preface, Journal of Hydrology, 320, 1–2, doi:10.1016/j.jhydrol.2005.07.054, 2006.

Schaake, J., Hamill, T., Buizza, R., and Clark, M.: HEPEX, the Hydrological Ensemble Prediction Experiment, Bulletin of the American Meteorological Society, 88, 1541–1547, doi:10.1175/BAMS-88-10-1541, 2007.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrological Processes, 15, 1063–1064, doi:10.1002/hyp.446, 2001.

Seifert, D., Sonnenborg, T. O., Refsgaard, J. C., Højberg, A. L., and Troldborg, L.: Assessment of hydrological model predictive ability given multiple conceptual geological models, Water Resources Research, 48, W06503, doi:10.1029/2011WR011149, 2012.

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

Smith, M. B., Seo, D.-J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., Moreda, F., and Cong, S.: The distributed model intercomparison project (DMIP): motivation and experiment design, Journal of Hydrology, 298, 4–26, doi:10.1016/j.jhydrol.2004.03.040, 2004.

Smith, M. B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., Cui, Z., Mizukami, N., Anderson, E. A., and Cosgrove, B. A.: The distributed model intercomparison project – Phase 2: Motivation and design of the Oklahoma experiments, Journal of Hydrology, 418–419, 3–16, doi:10.1016/j.jhydrol.2011.08.055, 2012.

Son, K. and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, Water Resources Research, 43, W01415, doi:10.1029/2006WR005032, 2007.

Thielen, J., Schaake, J., Hartman, R., and Buizza, R.: Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007, Atmospheric Science Letters, 9, 29–35, doi:10.1002/asl.168, 2008.

Thornthwaite, C. W.: An approach toward a rational classification of climate, Geographical Review, 38, 55–94, 1948.

Turc, L.: Le bilan d'eau des sols : relation entre les précipitations, l'évapotranspiration et l'écoulement, Annales agronomiques, Série A, 491–595, 1954.

Valéry, A.: Modélisation précipitations débit sous influence nivale : Elaboration d'un module neige et évaluation sur 380 bassins versants, PhD thesis, AgroParisTech, Paris, France, 2010.

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J.-M., Viney, N. R., and Teng, J.: Climate nonstationarity - Validity of calibrated rainfall-runoff models for use in climatic changes studies, Journal of Hydrology, 394, 447–457, doi:10.1016/j.jhydrol.2010.09.018, 2010.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, Hydrological Processes, 17, 455–476, doi:10.1002/hyp.1135, 2003.

Zhan, C.-s., Song, X.-m., Xia, J., and Tong, C.: An efficient integrated approach for global sensitivity analysis of hydrological model parameters, Environmental Modelling & Software, 41, 39–52, doi:10.1016/j.envsoft.2012.10.009, WOS:000315974500004, 2013.

Zhang, H., Huang, G. H., Wang, D., and Zhang, X.: Multi-period calibration of a semi-distributed hydrological model based on hydroclimatic clustering, Advances in Water Resources, 34, 1292–1303, doi:10.1016/j.advwatres.2011.06.005, 2011.