Hydrol. Earth Syst. Sci. Discuss., 10, 11337–11383, 2013
www.hydrol-earth-syst-sci-discuss.net/10/11337/2013/
doi:10.5194/hessd-10-11337-2013

Hydrology and Earth System Sciences Discussions
Open Access

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

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

[1]Irstea Antony, 1 rue Pierre-Gilles de Gennes, CS10030, 92761 Antony, France
[2]EDF R & D LNHE, 6 quai Watier, 78401 Chatou, France

Received: 22 August 2013 – Accepted: 29 August 2013 – Published: 5 September 2013

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

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

This paper investigates the robustness of rainfall-runoff models when their parameters are transferred in time. More specifically, we studied their ability to simulate water balance on periods with different hydroclimatic characteristics. The testing procedure consisted in a series of parameter transfers between 10-yr periods and the systematic analysis of mean-volume errors. This procedure was applied to three conceptual models of different structural complexity over 20 mountainous catchments in southern France. The results showed that robustness problems are common. Errors on 10-yr-mean flows were significant for all three models and calibration periods, even when the entire record was used for calibration. Various graphical and numerical tools were used to show strong similarities between the shapes of mean flow biases calculated on a 10-yr-long sliding window when various parameter sets are used. Unexpected behavioural similarities were observed between the three models tested, considering their large differences in structural complexity. While the actual causes for robustness problems in these models remain unclear, this work stresses the limited transferability in time of the water balance adjustments made through parameter optimization. Although absolute differences between simulations obtained with different calibrated parameter sets were sometimes substantial, relative differences in simulated mean flows between time periods remained similar regardless of the calibrated parameter sets.

1 Introduction

1.1 Confidence and evaluation of rainfall–runoff modelling in a context of 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

HESSD
10, 11337–11383, 2013

conditions observed during historical periods cannot be easily considered as representative of other periods (historical 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. Several exhaustive studies from different countries 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 conceptual models lack robustness when used in contrasted climate conditions.

Long historical records that include contrasted sub-periods are needed for evaluation schemes of model 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, Nash and Sutcliffe (1970) efficiency (NSE) 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 also 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 these problems that models have simulating water balances 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, the blame for failure situations of parameter transfer seems to often 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; Smith and Gupta, 2012), MOPEX (Schaake et al., 2006; Chahinian et al., 2006) and HEPEX (Schaake et al., 2007; Thielen et al., 2008) are good examples of use for these testing schemes and they sometimes concluded on the equally good suitability of simple models. Such evaluation approaches must be generalized and innovative strategies should be imagined to make the best use of the extended times-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 (see previous section for references). Here, we propose a simple diagnostic approach to further investigate this question. Using extended 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 model complexity or the period used for parameter calibration on this capacity to simulate water balances.

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

HESSD
10, 11337–11383, 2013

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 (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).

Climate forcing and flow records are 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 with neighbouring stations) and erroneous data were considered as gaps. Total precipitation and air temperature series were computed using the SPAZM reanalysis (based on ground network data and weather patterns) made by Gottardi et al. (2012) and available at a daily time step from 1948 to 2010 for the main mountainous areas in France (Alps, Massif Central and Pyrenees). They can be considered high-quality data. Finally, potential evapotranspiration (PE) time series were computed with empirical formula using the air temperature from the SPAZM reanalysis (Thornthwaite, 1948; Oudin et al., 2005).

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 issues. 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 other two models, to confirm the findings presented here (the results can be found in the Appendix).

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

Three conceptual hydrological models are 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.

HESSD
10, 11337–11383, 2013

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). Its inputs are cumulated annual rainfall and PE data. The model can be described using a non-linear equation:

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

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

2.2.2 GR4J-CemaNeige

GR4J is a parsimonious daily lumped 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). This snow module has two free parameters, which are optimized 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 "production part" of the model is computed on a topography-based mesh. It includes a snow module and a parameterized function to adjust PE amounts (based on the Thornthwaite formula). A total of 19 parameters must be optimized.

HESSD
10, 11337–11383, 2013

2.2.4 Calibration procedure

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

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

with ρ, σ and μ being the Pearson correlation coefficient, the standard deviation and the average functions, 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 descent 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 optimized 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).

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 Generalized Split-Sample Test (GSST) procedure (Coron et al., 2012). The testing procedure proposed here is different. It consists in a series of model calibrations over various sub-periods and a single simulation period corresponding to the entire available time series. The calibration sub-periods were built using a sliding window that is moved by one hydrological year between two neighbouring sub-periods (i.e. overlap is allowed). The length of this sliding window is chosen as a compromise simultaneously allowing for correct parameter determination and a sufficient number of potentially contrasted sub-periods. This testing procedure is summarized in Fig. 3, where θ_i is the optimal set identified on the sub-period i. 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 52).

3.2 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 volume errors over the available data record for different calibration configurations. These errors were expressed as a dimensionless bias given by $\overline{\widehat{Q}_{10y.}}/\overline{Q_{10y.}}$, in which $\overline{\widehat{Q}_{10y.}}$ and $\overline{Q_{10y.}}$ are the 10-yr-mean simulated and observed flows, respectively. The results obtained with different parameter sets can be superimposed on the same graph. Thus, we built visual tools for analysing model behaviours. We illustrate their construction on the example case of the Ubaye River at Barcelonnette (540 km^2, case study 1 in Fig. 1) using the GR4J-Cemaneige model. Figure 4 shows the successive steps followed to plot the time series of relative bias.

Here, time series of rainfall, 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

HESSD
10, 11337–11383, 2013

used to calibrate models and to compute volume errors. The building procedure is explained below:

3.2.1 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 the selected calibration criteria (KGE) accounts for bias, the volume error obtained for SP[08] is very small (i.e. $\overline{\widehat{Q}_{10y.}}/\overline{Q_{10y.}} \approx 1$). Then, from the flow series simulated over the whole period with the calibrated parameter set, one can compute the relative bias for each of the 40 remaining sub-periods and plot the relative bias 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 distance between starting years is less than 9 yr), but that the calibration and evaluation periods are independent in the other cases.

All 41 points can be joined and form a curve, which is specific to the parameter set. This curve, noted $\omega_{\theta_{SP[08]}}$, corresponds in fact to the 10-yr moving average error on mean flows when the model calibrated on SP[08] is used. One can note significant simulation errors. This indicates that it is difficult for the model to reproduce observed mean flows on this catchment over the whole period, with phases of mean-flow overestimation and underestimation. Since sub-periods overlap, there is a smoothing effect on these variations however.

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

The previous step is repeated with a second calibration sub-period SP[25]. Again, errors on mean flow are small on the calibration sub-period, but increase when the parameter set is 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.

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

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

3.2.4 Key questions

Numerous questions arise from the results obtained in the illustrative example of Fig. 4. First, each of the parallel curves illustrates a lack of robustness. A perfectly robust model would result in flat curves: the bias 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:

- Which model behaviour would we obtain with a parameter set optimized on the full record? The various parameter sets used to build Fig. 4c were optimized over 10 yr. 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 (i.e. lead to a flat $\omega_{\theta_{TP}}$ curve)?

- Which behaviours would result considering different model structures? Behavioural similarities were observed for GR4J-CemaNeige. Are these similarities observed for simpler or more complex conceptual models?

- Which model behaviours would be obtained on other catchments? We observed behavioural similarities between different parameter sets on the Ubaye River at Barcelonnette. Are these similarities observed on other catchments from the set?

HESSD
10, 11337–11383, 2013

3.3 Numerical criteria for analysis

We used numerical criteria to measure the degree of similarity between bias time series and to more easily generalize the evaluation over multiple catchments and models.

We aim at comparing the curves representing the temporal variations of model errors on mean flow volumes ($\overline{\widehat{Q}_{10y.}}/\overline{Q_{10y.}}$). These ω_θ curves can be defined as:

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

where SP[i] and SP[k] are the i-th and k-th 10-yr-long sub-periods used for parameter calibration and error computations, respectively.

For each hydrological model, we can compare various curves ($\omega_{\theta_{SP[i]}}$) corresponding to the different calibration sub-periods (SP[i]) and one additional curve ($\omega_{\theta_{TP}}$) corresponding to a calibration over the total period (TP), the latter being used as a reference for comparisons.

The standard deviation operator (σ) is used to measure both the scale of the volume error variations (criterion $\sigma[\omega_{\theta_{TP}}]$, see Eq. 4) and the significance of the "parallelism effect" between various ω_θ curves (criterion $\sigma[\omega_{\theta_{SP[i]}} - \omega_{\theta_{TP}}]$, see Eq. 5):

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

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀I ▶I
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

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

The first criterion ($\sigma[\omega_{\theta_{\mathrm{TP}}}]$) reveals the overall ability for a model to reproduce 10-yr-mean flows when this model is calibrated on the full available record. It varies between 0 (optimal situation with no errors) to $+\infty$. The second criterion ($\sigma[\omega_{\theta_{\mathrm{SP}[i]}} - \omega_{\theta_{\mathrm{TP}}}]$) expresses the similarity between relative variations of volume errors for different parameter sets. It takes values between 0 (situation where the $\omega_{\theta_{\mathrm{TP}}}$ curves are rigorously identical) and $+\infty$. We note that the mean volume error over the entire record $\left(\left[\overline{\widehat{Q}_{\mathrm{TP}}}\right]_{\theta_{\mathrm{SP}[i]}} / \overline{Q_{\mathrm{TP}}}\right)$ has no impact on this criterion. Indeed, only the shape similarities of the ω_θ curves are analysed and their vertical spacing is left out of consideration.

These standard deviations can be compared with each other using a ratio we have noted as ρ_i:

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

This ratio expresses the degree of "parallelism" relative to the magnitude of bias variations. In a way, ρ_i is a "noise-to-signal" ratio which highlights how strong the similarities are between different ω_θ curves.

A similar criterion can be built for inter-model comparisons where the "parallelism effect" is measured between volume bias variations for two models (M_1 and M_2), both calibrated over the entire time-series. In other words, we compare different $\omega_{\theta_{\mathrm{TP}}}$ curves.

This ratio, noted $\rho'_{M_1 M_2}$, is described in Eq. (7). The choice for the model serving as reference, whose corresponding $\sigma[\omega_{\theta_{\mathrm{TP}}}]$ constitutes the denominator in Eq. (7), is made arbitrarily.

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

As for $\sigma[\omega_{\theta_{\mathrm{SP}[i]}} - \omega_{\theta_{\mathrm{TP}}}]$, the criteria detailed in Eqs. (6) and (7) range between 0 and $+\infty$ (the smaller the value, the stronger the similarities between the ω_θ curves).

4 Results

4.1 Case studies – graphical analyses on three catchments

The graphical procedure illustrated in Fig. 4 was applied to the 20 catchments with the 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 red curve corresponds to the calibration over the entire record.

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

First of all, let us analyse each graph independently. It can be seen that the "parallelism effect" noted in Fig. 4 can also be observed here: the model calibration on different sub-periods leads to errors on 10-yr-mean flows, which vary similarly over time (see grey $\varpi_{\theta_{SP}}$) curves on graphs 5d to 5l). Moreover, the parameter set optimized on the full record does not yield a flatter ϖ_θ curve and hence does not provide a better simulation of mean flows simultaneously on every 10-yr-long sub-period (see red $\varpi_{\theta_{TP}}$) curves on graphs 5d to 5l). Logically, this curve is placed so that the mean volume bias of the entire period remains close to 1 (i.e. $\left(\left[\overline{\widehat{Q}_{TP}}\right]_{\theta_{TP}} / \overline{Q_{TP}} \approx 1\right)$. If we follow the terminology from Singh et al. (2013), the error analyses on the sub-period calibrations ($\varpi_{\theta_{SP}}$ curves) mostly concern "extrapolation cases", where the information content may differ between calibration and validation and greater errors could therefore be expected. However, when the parameter set optimized on the full record is used ($\varpi_{\theta_{TP}}$ curves), this is an "interpolation case" with a stable information content and where smaller errors are expected, which is obviously not the case for the catchments considered here.

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 bias (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). Our recent investigations on this topic, however, showed that these correlations are not systematic and that their significance greatly varies from one catchment to another (Coron, 2013). Additionally, on these three illustrative examples, we note that the available period for analysis is shorter for the Drac River than on the other two catchments, but the extent of the changes on observed data (rainfall, temperature, discharges) is similar for the three catchments over the common period. Therefore, the smaller range of bias variation 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 bias 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), which explicitly includes the bias. Indeed, the various $\omega_{\theta_{\mathrm{SP}[k]}}$ curves are "positioned" to ensure $\left(\left[\overline{\widehat{Q}_{\mathrm{SP}[k]}}\right]_{\theta_{\mathrm{SP}[i=k]}} / \overline{Q_{\mathrm{SP}[k]}} \approx 1\right)$, as shown in Fig. 4. The most spaced out curves are the ones whose corresponding calibration sub-periods constitute the upper and lower extremes in terms of relative variations. Likewise, for catchments where model errors on volumes are almost time-invariant, all $\omega_{\theta_{\mathrm{SP}}}$ curves are nearly flat and superimposed.

Thirdly, the graphs placed in columns (Fig. 5) show strong similarities, indicating similar behaviours of the three models tested on each catchment. 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 optimized parameters). The $\omega_{\theta_{\mathrm{TP}}}$ curve shapes (and indirectly the $\omega_{\theta_{\mathrm{SP}}}$ curve shapes) are not strictly identical between the three models, however.

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

The numerical criteria introduced in Sect. 3.3 can be used to measure these behavioural similarities systematically over a large 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 summarized 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

HESSD
10, 11337–11383, 2013

catchment). Relatively similar situations are observed for all three models, with median values around 4 %. Yet, small differences can be noted: Results for the Mouelhi formula and GR4J-CemaNeige model are almost identical, with the $\sigma[\omega_{\theta_{\mathrm{TP}}}]$ values slightly smaller for the latter. Larger differences are obtained with the Cequeau model, whose errors on simulated mean flows vary less with time. This model appears to be slightly more robust than the other two, at least with regard to its ability to simulate water balances simultaneously on various periods. Possible explanations for Cequeau's better robustness might be related to its greater structural complexity (in terms of conceptualization, parameterization and/or spatial distribution) or to the different ways snow storage or PE data are computed.

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). 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. 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 0.2. 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 with a median 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 $\omega_{\theta_{\mathrm{TP}}}$ 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. Yet, the smaller $\sigma[\omega_{\theta_{\mathrm{TP}}}]$ values obtained with Cequeau in some cases are not the only explanation for the greater ρ_i values observed. It seems likely that they result from the larger differences between ω_θ curves with this model (see Fig. 5 for examples of "parallelism

HESSD
10, 11337–11383, 2013

imperfections”). The reasons for these greater differences may stem from Cequeau's greater complexity compared to the Mouelhi formula and GR4J-CemaNeige. Because a larger number of parameters had to be optimized, some 10-yr-long sub-periods may not have been informative enough to allow their optimization. This could explain the fewer similarities between ω_θ trajectories.

4.3 Direct comparison of the three models' behaviours

The issue discussed in this paper has been broken down into three questions (see Sect. 3.2.4). 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 0.3 to 0.65. 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 volume bias (1.6 to 3 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 (cf. right versus left parts of Fig. 7), likely because Cequeau is slightly more robust on the catchment set (cf. lower $\sigma[\omega_{\theta_{\mathrm{TP}}}]$ on average).

HESSD
10, 11337–11383, 2013

The differences in volume bias variations caused by a change of hydrological model were expected, especially considering the large complexity gaps between the model structures used here. It is nevertheless surprising to see that the length of the calibration period has a limited impact on the relative variations of these biases for all three models. Indeed, volume bias variations are consistent within each structure when the sub-period or total-period calibrations are used. This consistency remains when all the information is used for calibration but a model change is considered, although it is not as strong in the latter case (see Fig. 6 vs. Fig. 7). One important point must not be forgotten however: only relative variations are considered here and the overall bias (i.e. the ϖ_{θ} curves' vertical positioning) is not measured. As can be seen from Fig. 5, calibrations on various sub-periods result in different overall biases, since there is a "parallelism effect" but no superposition with the $\varpi_{\theta_{\mathrm{SP}[i]}}$ curves. Conversely, overall biases close to 1 are reached for all $\varpi_{\theta_{\mathrm{TP}}}^{M_j}$ curves, since the objective function used (KGE) constrains the water balance adjustment.

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 balances 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 flows. Observed changes are plotted as well as changes simulated by each model. When expressed in a relative way (e.g. $\triangle\overline{Q}_{[A/B]} = \overline{Q_{\mathrm{SP}[A]}}/\overline{Q_{\mathrm{SP}[B]}}$), changes from different pairs of sub-periods 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 $\triangle\widehat{\overline{Q}}_{[A/B]}$ simulated using the parameter set optimized over the full record (θ_{TP}) and the numerous parameter sets ($\theta_{\mathrm{SP}[i]}$) obtained from

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

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 $\left[\triangle\widehat{\overline{Q}}_{[A/B]}\right]_{\theta_{\mathrm{TP}}}$ to $\triangle\overline{Q}_{[A/B]}$. 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 $\left[\triangle\widehat{\overline{Q}}_{[A/B]}\right]_{\theta_{\mathrm{TP}}}$ values are summarized by boxplots (showing the 5th, 25th, 50th, 75th and 95th percentiles). We see how the models used face difficulties to reproduce the climate elasticity of 10-yr-mean flows, i.e. larger changes are underestimated, whether they are positive or negative. Cequeau shows the best ability and the Mouelhi formula the worst, which is in accordance with the $\sigma[\omega_{\theta_{\mathrm{TP}}}]$ 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 $\left[\triangle\widehat{\overline{Q}}_{[A/B]}\right]_{\theta_{\mathrm{SP}[i]}}$ and $\left[\triangle\widehat{\overline{Q}}_{[A/B]}\right]_{\theta_{\mathrm{TP}}}$ for all the couples of sub-periods A and B. These values were used as coordinates to build clouds of points, which show whether all calibration periods lead to similar simulated mean-flow changes. Aggregated over the 20 catchments, the results for the three models 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 flows 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.

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 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 10-yr-mean simulated volumes are expressed relative to the mean volume 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 hypothesized that the bias 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 (cf. 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 determine a deeper diagnosis, including 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 conceptualization

The role of inappropriate model structure must of course be questioned regarding robustness problems. For instance, Hartmann et al. (2013) give an example of a need for adaptation of a model structure to ground realities in karstic zones. Simple or complex approaches can be used to investigate the question of structural deficit. 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, with a possible link to the changes in climatic conditions (Coron et al., 2012). This comparison could be extended to other model structures, although a relatively large complexity range has been considered here, from an annual 1-parameter formula to a semi-distributed daily model with 19 optimized parameters.

Problems of miscalibration or overcalibration of model parameters may also cause robustness problems. For the work reported here, different calibration criteria were tested, including the well-known NSE and a modified KGE where the weight of volume bias within the formula was reduced. We also attempted to calibrate the GR4J-

HESSD
10, 11337–11383, 2013

CemaNeige model on the total records with the exclusive aim of minimizing the standard deviation on the 10-yr-mean volume errors ($\sigma[\varpi_{\theta_{TP}}]$). None of these criteria could significantly reduce the robustness problems observed in this study. A brief review of the authors discussing parameters' miscalibration or overcalibration in hydrology include Wagener et al. (2003), Hartmann and Bárdossy (2005), Son and Sivapalan (2007), Gupta et al. (2009), Ebtehaj et al. (2010), Efstratiadis and Koutsoyiannis (2010), Andréassian et al. (2012), Gharari et al. (2013) and Zhan et al. (2013). They propose new calibration criteria or optimization strategies to reduce these problems, some of which seem promising. Yet, the risks for non-optimal parameterization occuring depend a great deal on the choices made on the model structure. Further investigations are required to confirm the deficiencies on water balance simulation highlighted here and should include both aspects of model structure and calibration strategy. While they may conclude on the sole responsibility of the conceptualization process for these deficiencies, other causes can contribute and should not be neglected.

5.2 Robustness and data

In spite of the quality verifications of the records to be used, the potential role of input errors on modelling performance must not be forgotten (Oudin et al., 2006; McMillan et al., 2010, 2011). Such errors can occur during the measurement or treatment phase. They may induce poor temporal transferability of model parameters if they vary temporally, for example in relation to human activities or climatic conditions. The incorrect estimation of precipitation and evapotranspiration fluxes may explain temporal robustness problems.

The inaccurate estimation of evapotranspiration is particularly suspected, since uncertainties are associated with the computation of potential evapotranspiration (PE) first and of actual evapotranspiration (AE) thereafter. Evapotranspiration is indeed 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; Herrnegger et al., 2012). Concerning the work presented here, quality checks were performed on rainfall, temperature and discharge series to detect obvious problems. Regarding PE series, complementary tests were made 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 volume bias were neither better nor exactly similar to those shown here and we could not conclude with certainty on this potential role of PE data on models' robustness deficiencies.

5.3 Robustness and changes in catchment functioning

Finally, although poor modelling strategies or data quality are major sources for model failure, other explanations are worth considering. Working on an (until then) unexplained over-estimation of the Meuse River runoff between 1930 and 1965, Fenicia et al. (2009) showed the major role of changes in land use management and forest age on the catchment's functioning. Such temporary or permanent changes of a catchment functioning will 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). 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 question the robustness of rainfall-runoff models, regarding their ability to reproduce water balances simultaneously on different temporal

HESSD
10, 11337–11383, 2013

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 if tests are performed on long records. When temporal transferability poses problems, choosing another calibration sub-period induces no significant difference on the relative change in 10-yr-mean simulated flows. For example, if we consider two temporal periods A and B, the $\overline{\widehat{Q}_A}/\overline{\widehat{Q}_B}$ ratio remains very stable regardless of the calibration period, even when the full record is used to optimize model parameters. The choice of the calibration period affects how the moving average curve of volume bias is positioned, but the relative changes between periods remain comparable. 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.

The three models tested in this study show significant 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 the structural complexity of the 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 estimation, for example. It is 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 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

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

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 summarized in Fig. A1 and Table A1.

These additional results are in accordance with those exposed in the article. First, the Mouelhi formula and GR4J-CemaNeige model show difficulties to reproduce water balances simultaneously on different temporal periods. Then, the "parallelism effect" observed during the study of volume errors variations is confirmed for these models (see Figs. A2 and A3). Again with this new catchment set the $\omega_{\theta_{\mathrm{TP}}}$ curve shapes (and indirectly the $\omega_{\theta_{\mathrm{SP}}}$ curve shapes) remain very similar for both models. This is shown in Fig. A2b by the low ρ_i values, whose distribution is 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 here.

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.

HESSD
10, 11337–11383, 2013

References

Andréassian, V.: Impact de l'évolution du couvert forestier sur le comportement hydrologique des bassins versants, Ph.D. thesis, UPMC, Paris, France, 262 pp., 2002. 11360

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. 11340

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, Hydrol. Process., 26, 2206–2210, doi:10.1002/hyp.9264, 2012. 11359

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, International Association of Hydrological Sciences, Melbourne, Australia, 32–38, 2011. 11344

Brigode, P., Oudin, L., and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, J. Hydrol., 476, 410–425, doi:10.1016/j.jhydrol.2012.11.012, 2013. 11339

Bulygina, N. and Gupta, H.: Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation, Water Resour. Res., 45, W00B13, doi:10.1029/2007WR006749, 2009. 11358

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, J. Hydrol., 298, 242–266, doi:10.1016/j.jhydrol.2004.03.042, 2004. 11358

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, edited by: Andréassian, V., Hall, A., Chahinian, N., Schaake, J., IAHS, Wallingford, 313–338, 2006. 11340

Charbonneau, R., Fortin, J., and Morin, G.: The CEQUEAU model: description and examples of its use in problems related to water resource management, Hydrolog. Sci. Bull., 22, 93–202, 1977. 11343

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, Stoch. Environ. Res. Risk Assess., doi:10.1007/s00477-013-0755-5, in press, 2013. 11360

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, 2011. 11340

Coron, L.: Les modèles hydrologiques conceptuels sont-ils robustes face á un climat en évolution? Diagnostic sur un échantillon de bassins versants français ausraliens, PhD thesis, AgroParisTech, Paris, France, 235 pp., 2013. 11351

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. 11339, 11344, 11345, 11358

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, J. Hydrol., 386, 186–197, doi:10.1016/j.jhydrol.2010.03.020, 2010. 11359

Ebtehaj, M., Moradkhani, H., and Gupta, H. V.: Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling, Water Resour. Res., 46, W07515, doi:10.1029/2009WR007981, 2010. 11359

Edijatno, Nascimento, N. D. O., Yang, X., Makhlouf, Z., and Michel, C.: GR3J: a daily watershed model with three free parameters, Hydrolog. Sci. J., 44, 263–277, doi:10.1080/02626669909492221, 1999. 11344

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, Vienna, Austria, p. 04593, 2005. 11344

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multiobjective calibration approaches in hydrological modelling: a review, Hydrolog. Sci. J., 55, 58–78, 2010. 11359

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?, Hydrol. Earth Syst. Sci., 13, 1727–1737, doi:10.5194/hess-13-1727-2009, 2009. 11360

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

François, B., Hingray, B., Hendrickx, F., and Creutin, J. D.: Storage water value as a signature of the climatological balance between resource and uses, Hydrol. Earth Syst. Sci. Discuss., 10, 8993–9025, doi:10.5194/hessd-10-8993-2013, 2013. 11344

Gharari, S., Hrachowitz, M., Fenicia, F., and Savenije, H. H. G.: An approach to identify time consistent model parameters: sub-period calibration, Hydrol. Earth Syst. Sci., 17, 149–161, doi:10.5194/hess-17-149-2013, 2013. 11339, 11359

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, J. Hydrol., 432–433, 154–167, doi:10.1016/j.jhydrol.2012.02.014, 2012. 11341

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. 11344, 11359

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, Adv. Water Resour., 54, 149–160, doi:10.1016/j.advwatres.2013.01.010, 2013. 11358

Hartmann, G. and Bárdossy, A.: Investigation of the transferability of hydrological models and a method to improve model calibration, Adv. Geosci., 5, 83–87, doi:10.5194/adgeo-5-83-2005, 2005. 11359

Herrnegger, M., Nachtnebel, H.-P., and Haiden, T.: Evapotranspiration in high alpine catchments – an important part of the water balance, Hydrol. Res., 43, 460–475, doi:10.2166/nh.2012.132, 2012. 11359

Klemeš, V.: Operational testing of hydrological simulation models, Hydrolog. Sci. J., 31, 13–24, doi:10.1080/02626668609491024, 1986. 11339

Koutsoyiannis, D.: Hurst-Kolmogorov Dynamics and Uncertainty, J. Am. Water Resour. Assoc., 47, 481–495, doi:10.1111/j.1752-1688.2011.00543.x, 2011. 11338

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. 11344

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. 11343

Lin, Z. and Beck, M. B.: Accounting for structural error and uncertainty in a model: An approach based on model parameters as stochastic processes, Environ. Model. Softw., 27–28, 97–111, doi:10.1016/j.envsoft.2011.08.015, 2012. 11358

Matalas, N.: Comment on the Announced Death of Stationarity, J. Water Resour. Plan. Manage., 138, 311–312, doi:10.1061/(ASCE)WR.1943-5452.0000215, 2012. 11338

Mathevet, T.: Quels modèles pluie-débit globaux au pas de temps horaire? Développements empiriques et comparaison de modle sur un large échantillon de bassins versants, Ph.D. thesis, ENGREF, Paris, France, 354 pp., 2005. 11344

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, Hydrol. Process., 24, 1270–1284, doi:10.1002/hyp.7587, 2010. 11359

McMillan, H., Jackson, B., Clark, M., Kavetski, D., and Woods, R.: Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, J. Hydrol., 400, 83–94, doi:10.1016/j.jhydrol.2011.01.026, 2011. 11359

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. 11339, 11345

Mezentsev, V.: Du nouveau sur le calcul de l'évaporation totale (Yechio raz o rastchetie srednevo summarnovo isspareniia), Meteorologiya i Gidrologiya (Russian Meteorology and Hydrology), 5, 24–26, 1955. 11343

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. 11338

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, Hydrolog. Sci. J., 58, 1256–1275, doi:10.1080/02626667.2013.809088, 2013. 11339

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, Cambridge University Press, Swansea, Royaume-Uni, 205–234, 1965. 11360

Mouelhi, S., Michel, C., Perrin, C., and Andréassian, V.: Linking stream flow to rainfall at the annual time step: The Manabe bucket model revisited, J. Hydrol., 328, 283–296, doi:10.1016/j.jhydrol.2005.12.022, 2006. 11343

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. 11338

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I. A discussion of principles, J. Hydrol., 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970. 11339

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, J. Hydrol., 303, 290–306, doi:10.1016/j.jhydrol.2004.08.026, 2005. 11341, 11360

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, J. Hydrol., 320, 62–83, doi:10.1016/j.jhydrol.2005.07.016, 2006. 11359

Perrin, C. and Andréassian, V. E.: The Court of Miracles of Hydrology, Hydrolog. Sci. J., 55, 849–1084, 2010. 11361

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. 11343

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, edited by: Coello, C. A. and Lamont, G. B., World Scientific, New York, USA, 79–100, 2004. 11344

Reusser, D. E. and Zehe, E.: Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity, Water Resour. Res., 47, W07550, doi:10.1029/2010WR009946, 2011. 11358

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, J. Geophys. Res., 115, D03106, doi:10.1029/2009JD012035, 2010. 11339

Schaake, J., Duan, Q., Andréassian, V., Franks, S., Hall, A., and Leavesley, G.: The model parameter estimation experiment (MOPEX) – Preface, J. Hydrol., 320, 1–2, doi:10.1016/j.jhydrol.2005.07.054, 2006. 11340

Schaake, J., Hamill, T., Buizza, R., and Clark, M.: HEPEX, the Hydrological Ensemble Prediction Experiment, B. Am. Meteorol. Soc., 88, 1541–1547, doi:10.1175/BAMS-88-10-1541, 2007. 11340

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrol. Process., 15, 1063–1064, doi:10.1002/hyp.446, 2001. 11340

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 Resour. Res., 48, W06503, doi:10.1029/2011WR011149, 2012. 11339

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrol. Earth Syst. Sci., 16, 1171–1189, doi:10.5194/hess-16-1171-2012, 2012. 11339, 11358

Singh, S. K., McMillan, H., and Bardossy, A.: Use of the data depth function to differentiate between case of interpolation and extrapolation in hydrological model prediction, J. Hydrol., 477, 213–228, doi:10.1016/j.jhydrol.2012.11.034, 2013. 11351

Smith, M. B. and Gupta, H. V.: The Distributed Model Intercomparison Project (DMIP) – Phase 2 experiments in the Oklahoma Region, USA, J. Hydrol., 418–419, 1–2, doi:10.1016/j.jhydrol.2011.09.036, 2012. 11340

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, J. Hydrol., 298, 4–26, doi:10.1016/j.jhydrol.2004.03.040, 2004. 11340

Son, K. and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, Water Resour. Res., 43, W01415, doi:10.1029/2006WR005032, 2007. 11359

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, Atmos. Sci. Lett., 9, 29–35, doi:10.1002/asl.168, 2008. 11340

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

Thornthwaite, C. W.: An approach toward a rational classification of climate, Geograph. Rev., 38, 55–94, 1948. 11341

Turc, L.: Le bilan d'eau des sols: relation entre les précipitations, l'évapotranspiration et l'écoulement, Annales agronomiques, Série A, 491–595, 1954. 11343

Valéry, A.: Modélisation précipitations-débit sous influence nivale, élaboration d'un module neige et évaluation sur 380 bassins versants, Ph.D. thesis, AgroParisTech, Paris, France, 303 pp., 2010. 11343

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, J. Hydrol., 394, 447–457, doi:10.1016/j.jhydrol.2010.09.018, 2010. 11339

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, Hydrol. Process., 17, 455–476, doi:10.1002/hyp.1135, 2003. 11359

Zhan, C.-S., Song, X.-M., Xia, J., and Tong, C.: An efficient integrated approach for global sensitivity analysis of hydrological model parameters, Environ. Model. Softw., 41, 39–52, doi:10.1016/j.envsoft.2012.10.009, 2013. 11359

HESSD

10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

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

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

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 (annual mean) [–]	2 %	3 %	7 %	13 %	30 %
Mean annual pot. evap. PE($_{Oudin}$) [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

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

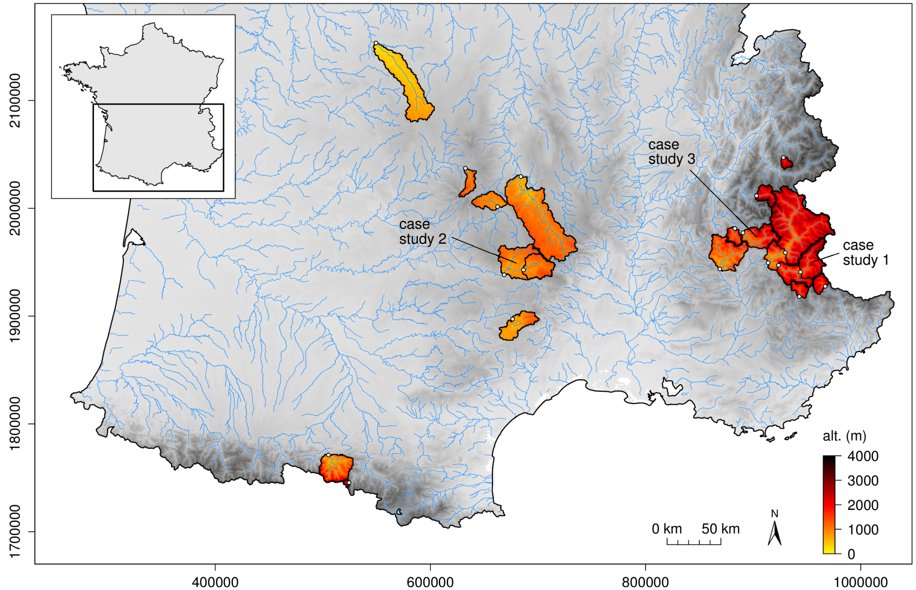
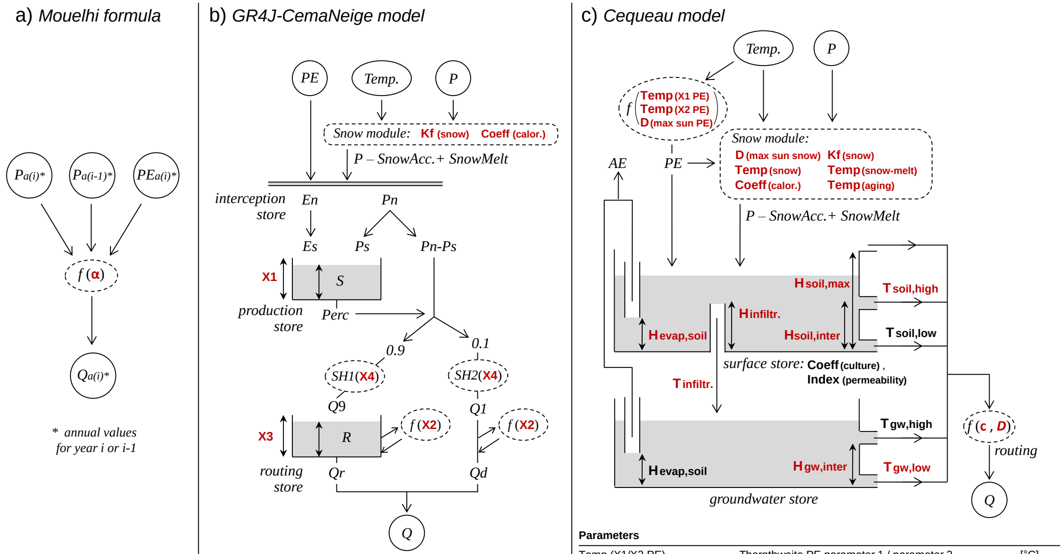


Fig. 1. Locations of the 20 catchments used in this study.

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.



Parameters

α	multiplication factor	[-]

Parameters

X1	production store capacity	[mm]
X2	groundwater exchange coefficient	[-]
X3	routing store capacity	[mm]
X4	unit hydrograph time constant	[day]
Kf (snow)	degree-day melt coefficient	[mm.j^{-1}]
Coeff (calor.)	weighting coefficient for snow pack thermal state	[-]

Parameters

Temp (X1/X2 PE)	Thornthwaite PE parameter 1 / parameter 2	[°C]
D (max sun PE/snow)	julian day of maximum PE / maximum snow	[-]
Kf (snow)	degree-day melt coefficient	[mm.j^{-1}]
Temp (snow/snow-melt/aging)	temperature threshold for snow- making/melt/aging	[°C]
Coeff (calor.)	weighting coefficient for snow pack thermal state	[-]
T infiltr./soil/gw	time constant for filling or emptying of stores	[d]
H evap./infiltr./soil/gw	max. capacity of stores or height threshold for emptying	[mm]
c	celerity coefficient for routing (1D Hayami)	[d]
D	diffusion coefficient for routing (1D Hayami)	[d]

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

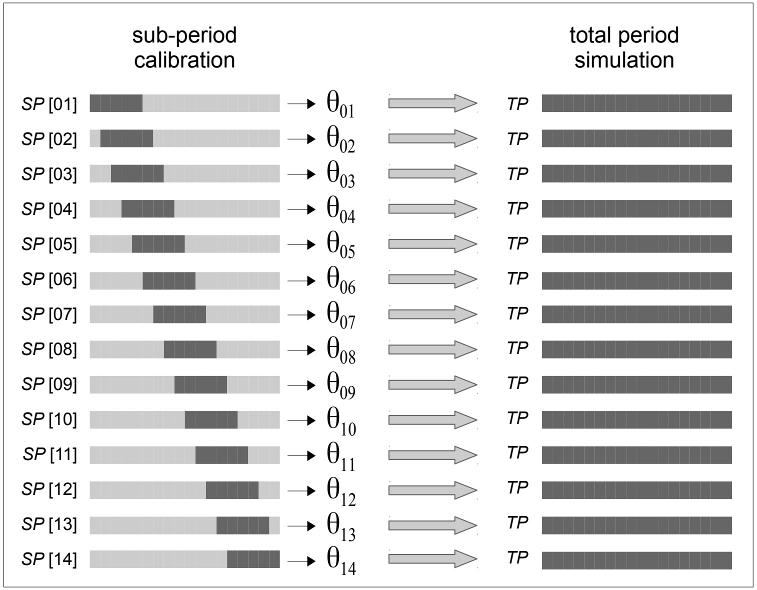


Fig. 3. Sub-period (SP) calibration procedure and simulation over the total period (TP) (example of 5-yr-long sub-periods within an 18-yr-long period).

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

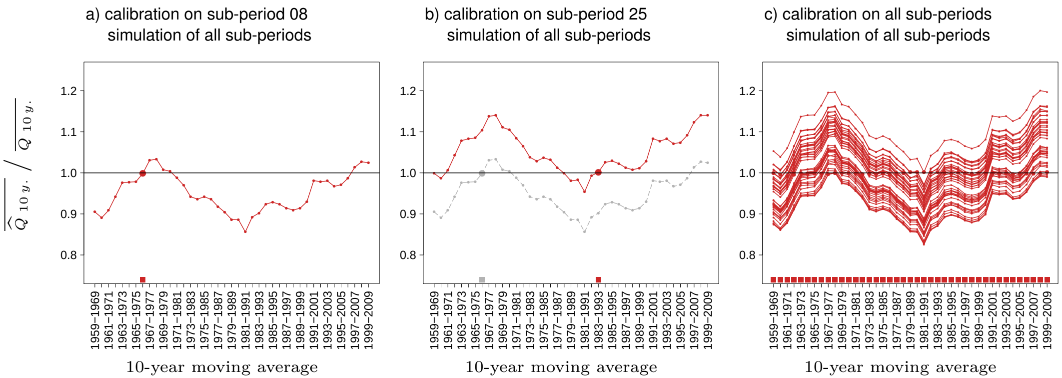


Fig. 4. Construction of the graphical representation of the series of 10-yr moving average biases with the GR4J-CemaNeige model.

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

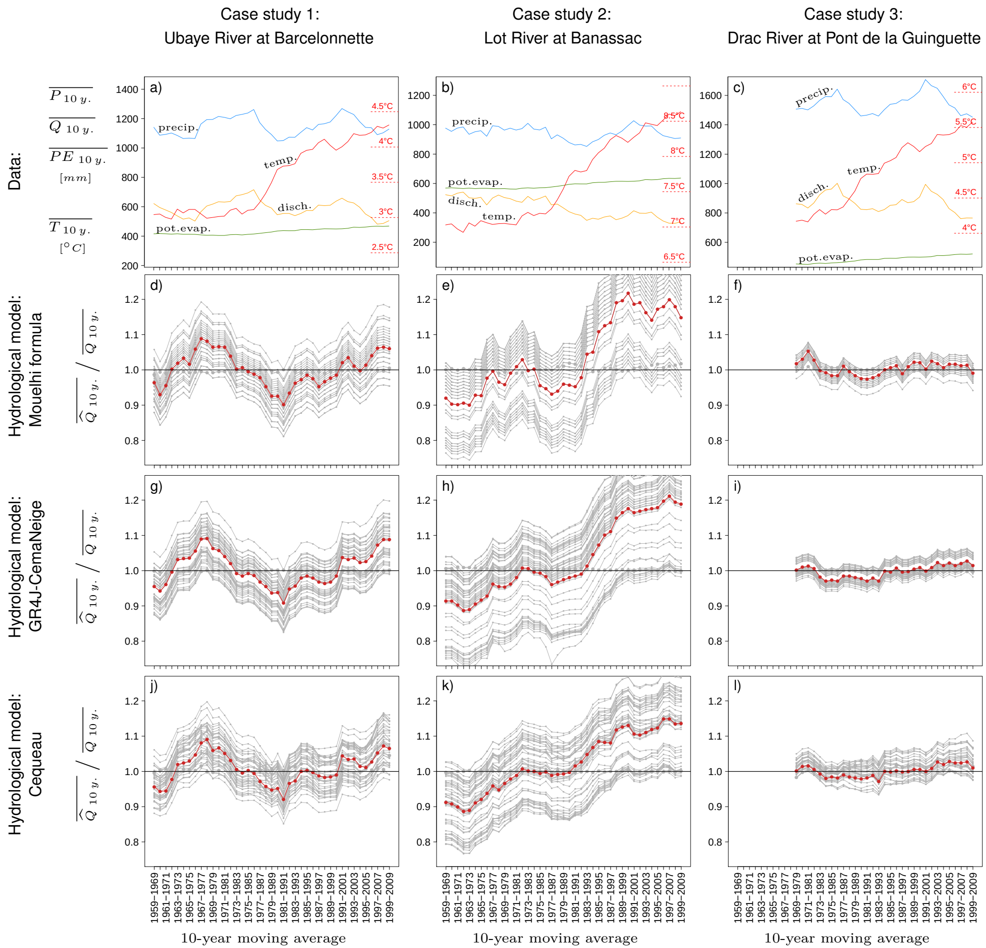


Fig. 5. Examples of behavioural similarities observed on three catchments with the three models tested (for **d** to **l**, the various $\omega_{\theta_{\mathrm{SP}[i]}}$ curves are in grey and the single $\omega_{\theta_{\mathrm{TP}}}$ curve is in red).

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

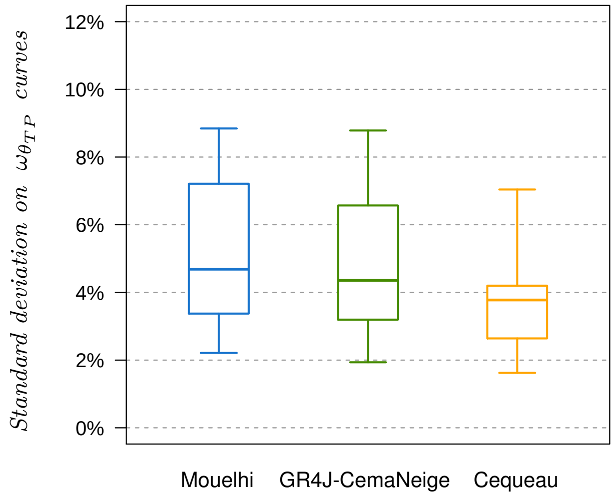


Fig. 6. Standard deviations of the 10-yr moving average on volume bias obtained during calibration over the full record (summary for the three models over 20 catchments through the $\sigma[\omega_{\theta_{TP}}]$).

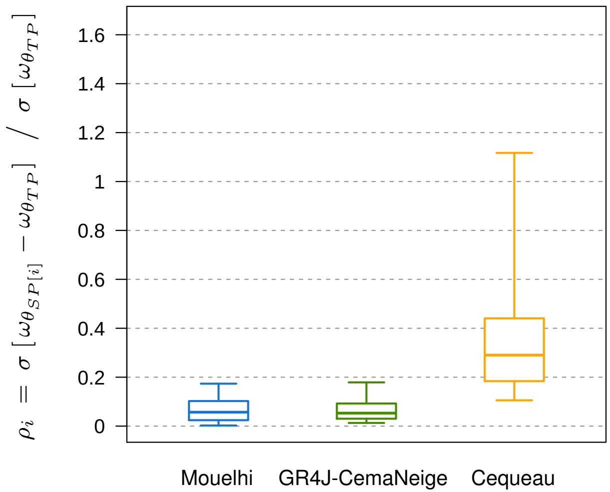


Fig. 7. Behavioural similarities observed between sub-period and full record calibrations in terms of 10-yr moving average on volume bias (summary for the three models over 20 catchments through the ρ_i ratio, see Eq. 6).

HESSD
10, 11337–11383, 2013

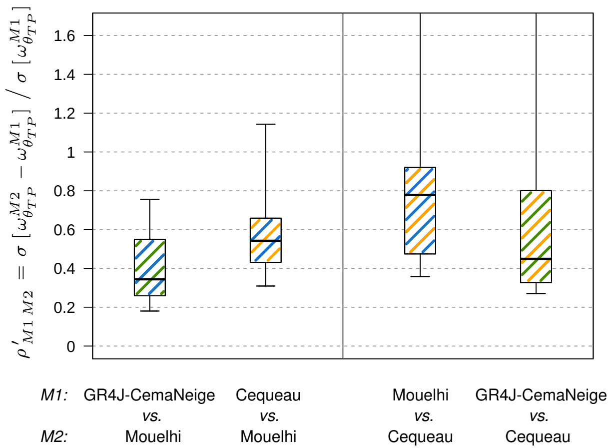


Fig. 8. Behavioural similarities observed between different models in terms of 10-yr moving average on volume bias. Calibrations over the full record (summary over 20 catchments through the $\rho'_{M_1 M_2}$ ratio, see Eq. 7).

HESSD
10, 11337–11383, 2013

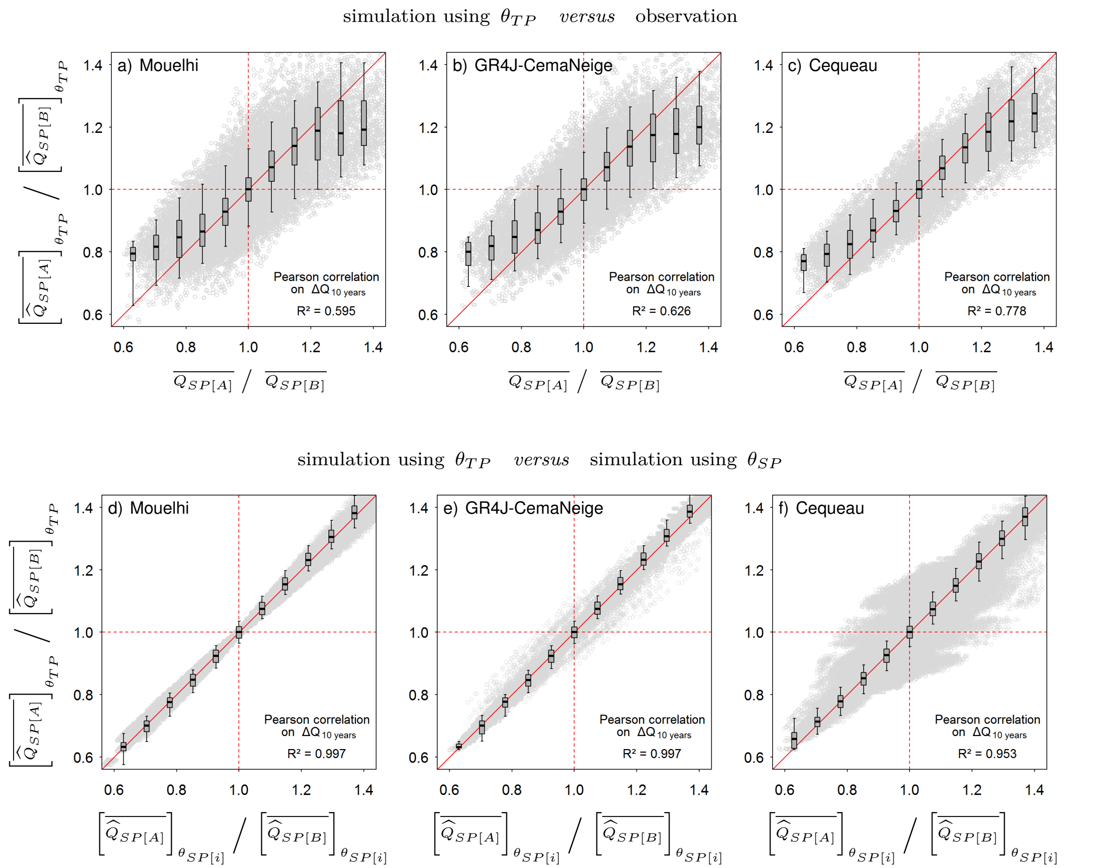


Fig. 9. Comparisons of relative changes on 10-yr-mean flows, observed and simulated (aggregation of results from 20 catchments).

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

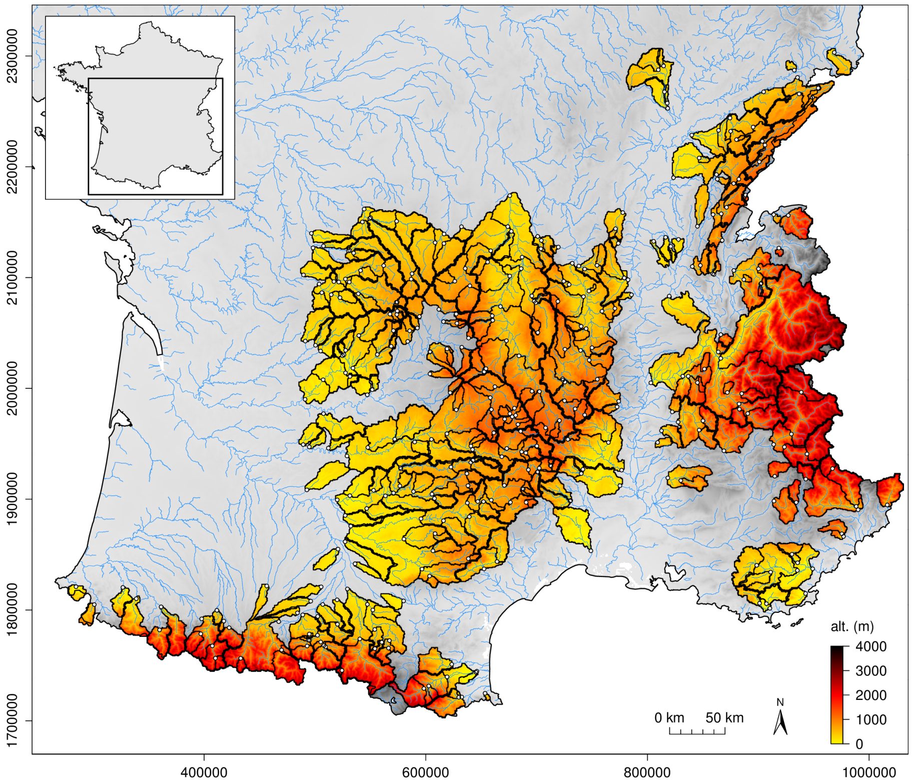


Fig. A1. Locations of the 365 catchments used in the additional testing with the Mouelhi formula and GR4J-CemaNeige model.

HESSD
10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.

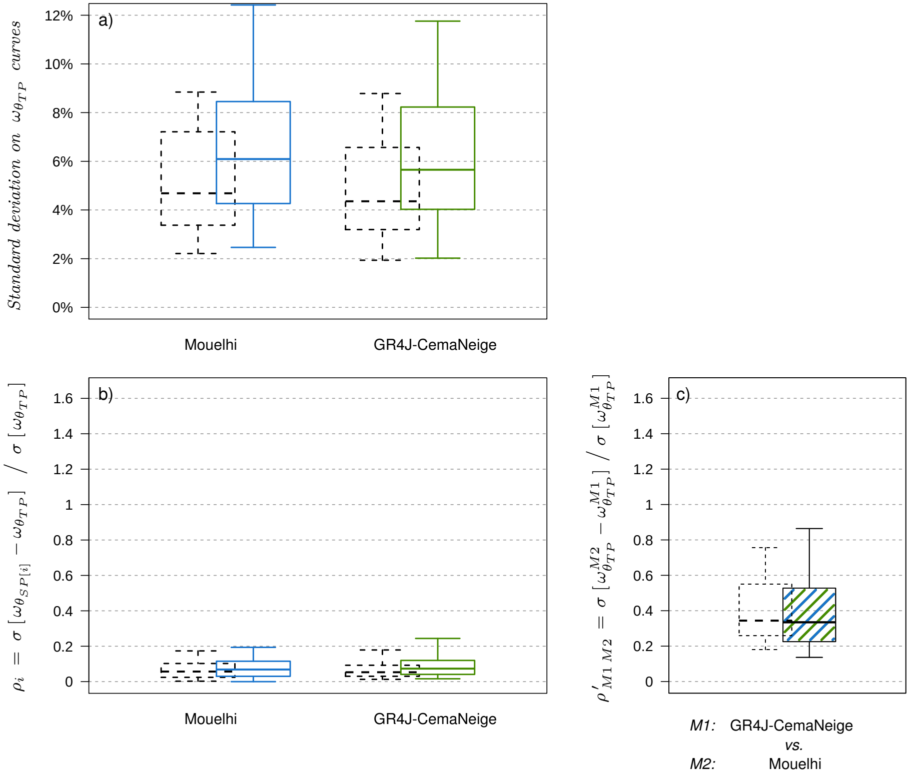


Fig. A2. Distributions of $\sigma[\omega_{\theta_{TP}}]$, ρ_i and $\rho'_{M_1 M_2}$ values obtained for the set of 365 catchments (solid coloured lines) and comparison with the previous results obtained on 20 catchments (dashed black lines).

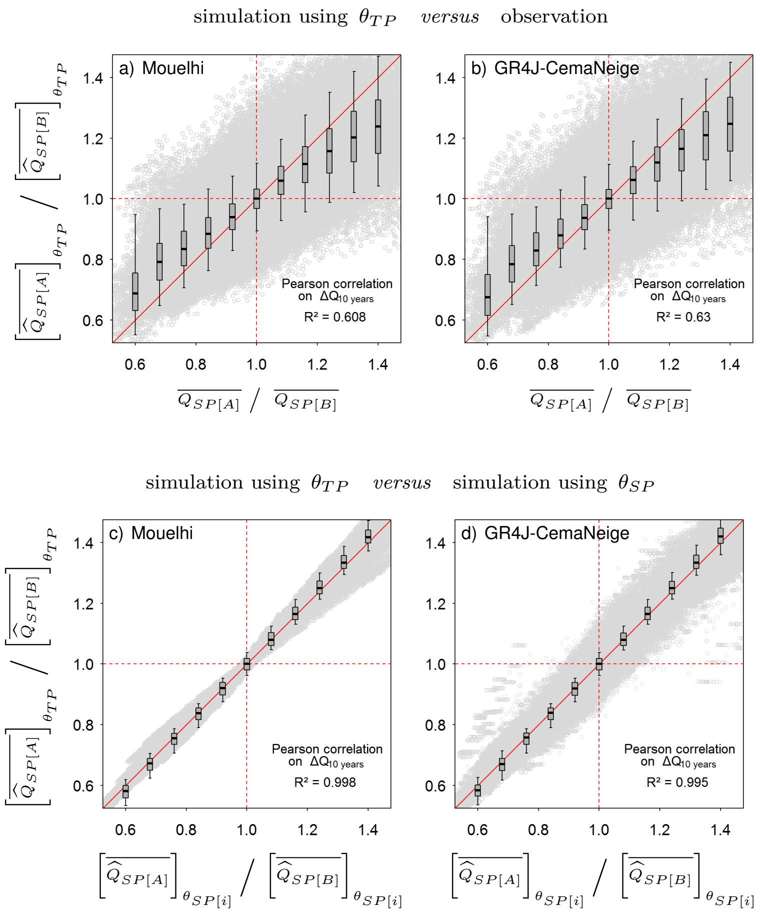


Fig. A3. Comparisons of relative changes on 10-yr-mean flows, observed and simulated (aggregation of results from 365 catchments).

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

HESSD

10, 11337–11383, 2013

On the lack of robustness of hydrologic models

L. Coron et al.