Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments

Laurent Coron, Vazken Andréassian, Charles Perrin, Julien Lerat, J. Vaze, Marie Bourqui, F. Hendrickx

Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments

L. Coron,[1,2] V. Andréassian,[2] C. Perrin,[2] J. Lerat,[3] J. Vaze,[3] M. Bourqui,[1] and F. Hendrickx[1]

Received 7 December 2011; revised 26 March 2012; accepted 15 April 2012; published 26 May 2012.

[1] This paper investigates the actual extrapolation capacity of three hydrological models in differing climate conditions. We propose a general testing framework, in which we perform series of split-sample tests, testing all possible combinations of calibration-validation periods using a 10 year sliding window. This methodology, which we have called the generalized split-sample test (GSST), provides insights into the model's transposability over time under various climatic conditions. The three conceptual rainfall-runoff models yielded similar results over a set of 216 catchments in southeast Australia. First, we assessed the model's efficiency in validation using a criterion combining the root-mean-square error and bias. A relation was found between this efficiency and the changes in mean rainfall (P) but not with changes in mean potential evapotranspiration (PE) or air temperature (T). Second, we focused on average runoff volumes and found that simulation biases are greatly affected by changes in P. Calibration over a wetter (drier) climate than the validation climate leads to an overestimation (underestimation) of the mean simulated runoff. We observed different magnitudes of these models deficiencies depending on the catchment considered. Results indicate that the transfer of model parameters in time may introduce a significant level of errors in simulations, meaning increased uncertainty in the various practical applications of these models (flow simulation, forecasting, design, reservoir management, climate change impact assessments, etc.). Testing model robustness with respect to this issue should help better quantify these uncertainties.

Citation: Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx (2012), Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, *Water Resour. Res.*, *48*, W05552, doi:10.1029/2011WR011721.

1. Introduction

1.1. Challenges in Predicting Hydrological Response Under Variable Climatic Conditions

[2] Quantifying the impacts of climate change on streamflow has been an increasing concern in the past few years and has been the focus of many studies throughout the world [*Caballero et al.*, 2007; *Vicuna and Dracup*, 2007; *Steele-Dunne et al.*, 2008; *Chiew et al.*, 2009; *Görgen et al.*, 2010]. The modeling steps associated with this task include (1) selecting emission scenarios, (2) running global circulation models (GCMs), (3) downscaling the GCM's output to a scale that can be used for hydrology, and (4) running hydrological models that simulate the rainfall-runoff (RR) transformation at the catchment scale. Step 4 is often considered to contribute less than the other steps to the overall uncertainty [*Wilby and Harris*, 2006; *Prudhomme and Davies*, 2009; *Kay et al.*, 2009; *Arnell*, 2011; *Teng et al.*, 2011]. However, the uncertainty associated with the estimation of parameters of hydrological models cannot be neglected [*Wilby*, 2005; *Vaze et al.*, 2010b; *Merz et al.*, 2011]. This is even truer when RR models are run under climatic conditions significantly different from calibration conditions (e.g., projections of future conditions versus current conditions). Indeed, many unknowns remain concerning the actual transposability over time of model parameters under contrasted conditions. Although this transposability is a critical issue in the context of climate change impact studies where nonstationary conditions are explicitly considered, it also has implications in other more operational model applications (like forecasting, design, etc.), in which model robustness is essential to provide reliable results. Besides, fluctuations in climate also exist in historical time series (such as Hurst-Kolmogorov behaviors [see *Koutsoyiannis*, 2011]), which may question parameters transferability whenever a model is used to simulate flows on a period whose climatic conditions are different from those in model calibration.

[1]LNHE, EDF R&D, Chatou, France.
[2]Hydrosystems and Bioprocesses Research Unit, Irstea, Antony, France.
[3]Black Mountain Laboratories, CSIRO Land and Water, Acton, ACT, Australia.

Corresponding author: L. Coron, LNHE, EDF R&D, 6 quai Watier, F-78401 Chatou CEDEX, France. (laurent.coron@edf.fr)

0043-1397/12/2011WR011721

1.2. Parameter Dependency on Calibration Period Climate

[3] Because of the lack of knowledge and data on the true functioning of the hydrological system, all hydrological models remain to some extent conceptual and empirical [*Murphy et al.*, 2006]. As a result, deriving physically

meaningful values for the model's parameters via calibration remains a challenging task. In their discussion on model "pathologies," *Coron et al.* [2011] reviewed various situations where parameter estimation can be hampered and induce values that transfer poorly to other periods. Common examples are cases where input quality and/or availability evolve as well as issues related to low levels of parameter identifiability. Problematic situations may also emerge from the climatic dependency of model parameters.

[4] Existing hydrological models have been developed using either top-down (conceptual models) or bottom-up (physically based) approaches. However, both types of models suffer from the same problems when it comes to the calibration of their parameters. At calibration stage, the optimal set can vary over time in accordance with seasonal and/or long-term climatic variations. *Wagener et al.* [2003] applied a five-parameter lumped model to an English catchment and showed that summer and rain storm periods require different optima for the parameters controlling rapid water transfer. *Choi and Beven* [2007] sampled the times series of a South Korean catchment according to a hydrological similarity measure. Calibrating TOPMODEL parameters on each cluster, they found that optimal sets on some clusters were not convenient for use on others. *Rosero et al.* [2010] found that parameters from the Noah land surface model, which should in principle be controlled only by physical site characteristics (e.g., soil and vegetation type), were also strongly influenced by climatic conditions. These findings are not restricted to a limited number of particular catchments. Although the work from *Rosero et al.* [2010] was based on only nine catchments, *Vaze et al.* [2010b] and *Merz et al.* [2011] conducted studies over 61 Australian and 273 Austrian catchments, respectively, and observed similar dependencies. *Merz et al.* [2011] established a link between the HBV parameters representing snow and soil moisture processes with climatic characteristics such as air temperature and PE. Cases of apparent independence may also be observed (see the work by *Niel et al.* [2003] on 17 African catchments). Most of these past results indicate that the assumption of parameter stability over time is strong. Parameter values can vary seasonally because of differences in dominant hydrological processes controlling runoff generation in different seasons but may also change on longer time scales in relation to climate variability (e.g., modifications of annual groundwater balance, vegetation change, etc.). Recently, *de Vos et al.* [2010] made an interesting proposition to further investigate the reasons for this disturbing dependency: they suggested clustering time series according to climatic similarities and allowing parameters to vary over these clusters during calibration. Rather than an alternative optimization method, they presented it as a tool for investigating model functioning and thus identifying the possible needs for improvements.

[5] A parallel can be made between transferring parameters over time and space. For example, *Merz et al.* [2011] observed temporal trends on parameter values (due to climate evolution) that were comparable to variations over space when moving between regions with different climates. A similar trade of space for time was made by *Singh et al.* [2011], who used a regional approach to evaluate extrapolation skills of parameter sets by transferring them to other catchments in warmer climatic zones.

[6] Note that, to ease reading, the term climate is abusively used hereafter to designate the main characteristic of precipitation, temperature, etc., over a time-limited period (often 10 years long here), whereas climate usually refers to characteristics over an extended period of time (typically several decades).

[7] A typical testing procedure to investigate parameter dependency on climate and related consequences on model efficiency is the differential split-sample test (DSST) proposed by *Klemeš* [1986]. This is a specific case of the split-sample test (SST), where calibration and validation periods are chosen according to their climatic differences. The parameter dependency on the calibration periods is analyzed through the evolution of the model's performance on this test. Examples of applications of DSST include the studies by *Refsgaard and Knudsen* [1996], *Donnelly-Makowecki and Moore* [1999], *Xu* [1999], *Seibert* [2003], *Wilby* [2005], *Chiew et al.* [2009], *Vaze et al.* [2010b], and *Bastola et al.* [2011]. Most of these authors observed decreases in model performance (i.e., larger model errors) after transferring parameter sets between climatically contrasted periods. They concluded that a model's suitability for climate change impact studies depends on the judicious choice of the calibration period. In that context, some recommend the use of extended periods to ensure sufficiently diverse climatic and flow conditions during calibration to give a representative picture of their natural variability. Others suggest the use of calibration periods whose climatic conditions are closer to future ones, often corresponding to the recent records. However, this second option prevents from benefiting from the full information available to quantify modeling uncertainty.

1.3. A Need for Further Investigation

[8] Increasingly, hydrologists use RR models over wide ranges of climatic conditions not necessarily encountered during the calibration stage, and need to have an estimate of the uncertainties associated with their simulations. Several studies have emphasized the limitations in the transfer of parameters between climatically contrasted periods, some of which were mentioned above. However, very few investigations have been conducted using a methodology that would provide general conclusions on this issue, i.e., on the basis of a large number of catchments and using different models. The two main studies by *Vaze et al.* [2010b] and *Merz et al.* [2011] showed that differences in climate between calibration and validation could significantly affect model performances. For instance, *Vaze et al.* [2010b] found that transferring parameters to a drier climate was particularly problematic and concluded that such transfer should not be made for changes in mean rainfall greater than 15%.

[9] The above mentioned studies require complementary work (1) to develop more general testing procedures able to provide comparable results under various conditions and over a wide range of parameter transfer conditions, thus resulting in more robust conclusions on parameter transferability, (2) to apply such procedures on a variety of cases for a better quantification of model robustness under a changing climate, and (3) to identify the situations where parameters are not transferable and, if possible, explain why.

[10] This study extends the work of *Vaze et al.* [2010b] by enlarging the catchment set and proposing a new

generalized methodology to evaluate the validity of RR models for use under nonstationary climatic conditions. Some of the conclusions of these authors are confirmed, but new insights are also provided (e.g., on catchment-specific behaviors).

[11] This paper is organized as follows. The catchment set and hydrological models used are presented in section 2. The methods used to evaluate the parameter transferability are described in section 3. The results are presented in section 4, starting from the analysis of the entire data set and then distinguishing different behaviors. A discussion around the methodology and the results is provided in section 5. Conclusions are summarized in section 6.

2. Catchment Set and Models

2.1. Study Area

[12] The choice of the study area was greatly influenced by the scope of this work. We needed sufficiently variable climatic conditions to make it possible to select contrasted periods for model testing in the most extreme conditions. In this perspective, we used a set of 216 catchments in southeast Australia, where climate variability is notoriously greater than in many other places in the world. The initial set was composed of 228 catchments, but 12 catchments were not used in our testing procedure because of insufficient data availability (see section 3.4). This catchment set has been described in detail by *Vaze et al.* [2010a]. The catchments are located on a large zone from south Queensland to west Victoria along the Great Dividing Range [Figure 1]. The range of climate and physical characteristics of the data set is summarized in Table 1. Rainfall is the most important driver of runoff in Australia and is much more variable both temporally and spatially than the other climate variables [*Vaze et al.*, 2010a]. Only 15% of the rainfall becomes runoff on average for the catchment set. Significant variations in rainfall and considerable variations in streamflow can occur between years, as shown by the interannual variability coefficients (see Table 1). Variability also exists between longer periods, as shown in Figure 2, which plots the series of relative mean PE, rainfall and streamflow values over a 10 year sliding window. The three graphs illustrate how the mean climate over a decade can differ from the climate over the entire record (32 years in most cases). While ranges of $\pm 10\%$ in mean rainfall and $\pm 3\%$ in mean PE between 10 year subperiods can be observed for most catchments, this results in relative variations on streamflow that can be much larger (up to 50% in absolute value). Note that on most catchments, the 1980s were wetter than average, while the end of the period was much drier. This is a quite interesting contrast for the objectives of our study. Across the data set, there is also a spatial variation of average conditions in terms of rainfall, PE and runoff. Catchments located east and south of the Great Dividing Range are wetter than the catchments located inland. Moreover, catchments are summer dominated in the

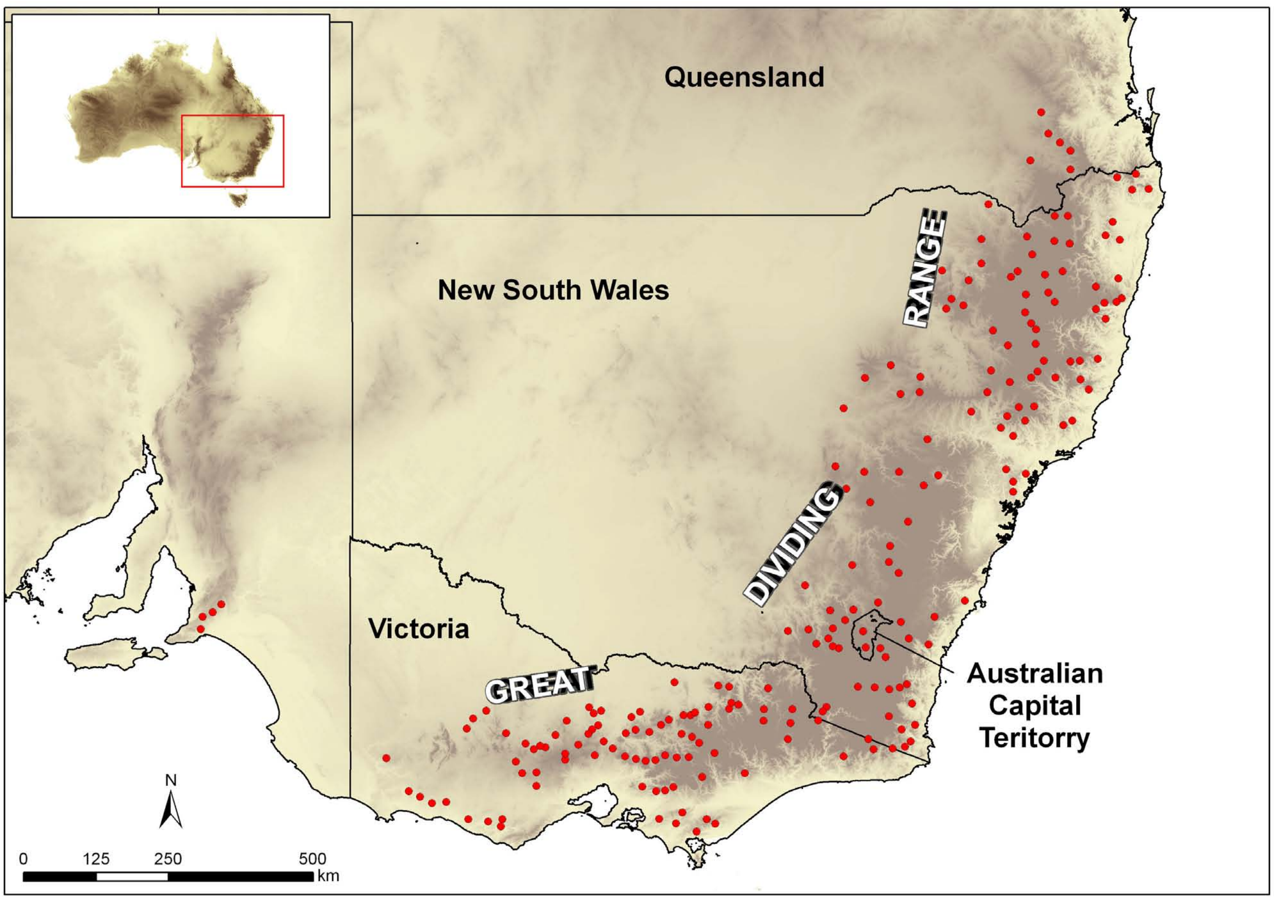

Figure 1. Locations of the 216 catchments.

Table 1. Percentiles of the Distributions of a Few Catchment Characteristics on the Entire Set of 216 Catchments[a]

	Statistics Over the Entire Catchment Set				
	5th Percentile	25th Percentile	Median	75th Percentile	95th Percentile
Catchment surface (km^2)	70	160	330	630	1240
Mean annual potential evaporation PE_{ma} (mm)	1070	1120	1200	1290	1410
Mean annual rainfall P_{ma} (mm)	570	720	860	1100	1400
Mean annual runoff Q_{ma} (mm)	30	60	120	250	500
Aridity index P_{ma}/PE_{ma} (%)	45	59	73	92	121
Rainfall-runoff yield Q_{ma}/P_{ma} (%)	4	9	15	23	36
Interannual variability of PE (%)	2.2	2.5	2.7	2.8	3.1
Interannual variability of P (%)	13	17	20	23	29
Interannual variability of Q (%)	29	46	70	91	114

[a]Values of interannual variability correspond to coefficients of variation calculated on 10 year periods.

north (i.e., most rainfall and runoff occur in summer) and winter dominated in the south, while interannual variability is greater in the north than in the south.

[13] Daily data of rainfall and PE were derived from the SILO Data Drill, which provides daily data for 0.05° surface grids (~5 × 5 km) across Australia (http://www.longpaddock.qld.gov.au/silo/). These estimates are interpolated from point measurements made by the Australian Bureau of Meteorology. Daily PE is computed using Morton's wet environment algorithms [*Morton*, 1983]. Daily streamflow data for the 216 catchments were obtained from relevant state government agencies and were checked for errors. For a majority of catchments, continuous records of rainfall, PE and runoff were available for the 1974–2006 period. Most of the catchments range in size between 100 and 1000 km^2, with a median value around 330 km^2. They are mostly unregulated with no major storage or irrigation schemes. These data were partly used in recent Australian projects such as the Murray-Darling Sustainable Yields project [*Chiew et al.*, 2008], the South-Eastern Australian Climate

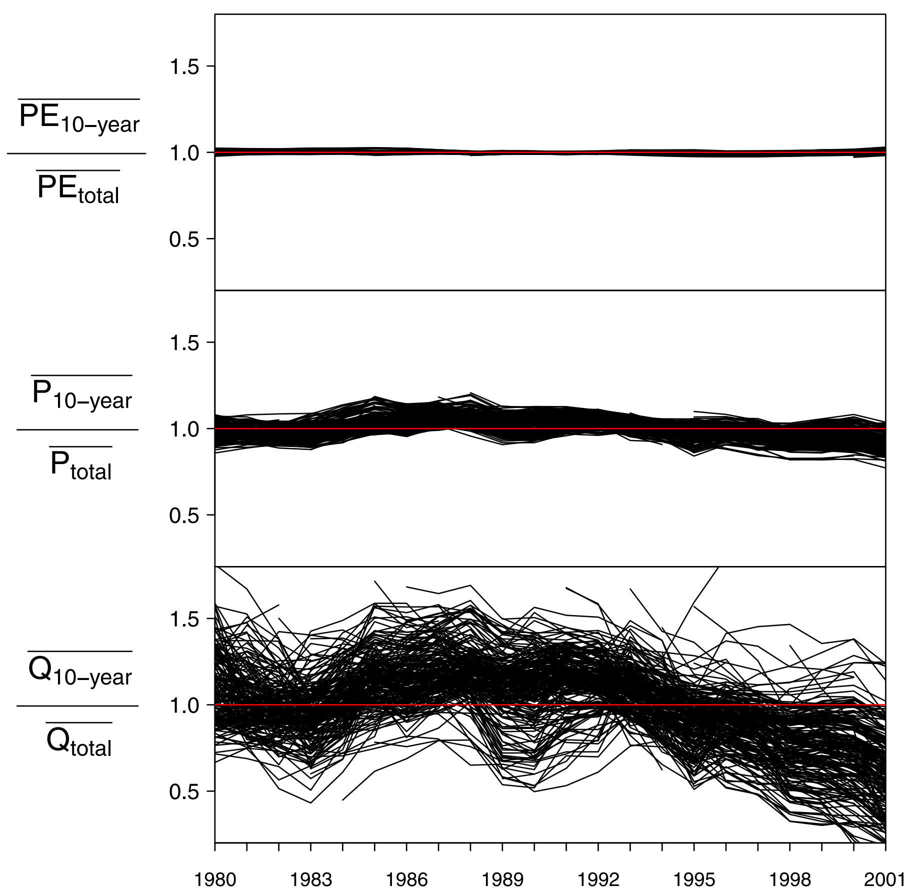

Figure 2. Relative long-term climate variability of potential evapotranspiration (PE), rainfall (P), and streamflow (Q) over the catchment set. For each catchment, a line corresponds to the series of mean values over a 10 year sliding window. Values are expressed relative to the average value on the total record (usually 1974–2006). Each value is plotted at the central year of the 10 year window.

Initiative (http://www.seaci.org/) and the study for the New South Wales Office of Water on climate impact on runoff [*Vaze and Teng*, 2011].

2.2. Hydrological Models

[14] Three daily lumped continuous reservoir-type RR models were used in this study: GR4J, MORDOR6 and SIMHYD. These models had already been applied to large data sets in previous studies [*Chahinian et al.*, 2006; *Oudin et al.*, 2008; *Perrin et al.*, 2008; *Chiew et al.*, 2009]. Table 2 gives an overview of the characteristics of these models as well as the references where detailed descriptions are provided. In spite of their parsimony (only a few free parameters), they showed a good level of efficiency in past applications and correspond to different representations of the RR transformation. GR4J is based on two stores and four parameters to calibrate, while MORDOR6 and SIMHYD both have four stores and six and seven free parameters, respectively.

3. Crash Test Methodology and Analysis Method

[15] Our general objective was to study the transfer of parameter sets between climatically contrasted periods. In the vein of the discussion of *Andréassian et al.* [2009] on model evaluation, we implemented a "crash test" methodology for models to be used in changing climatic conditions (typically such as climate change impact studies), i.e., a testing methodology putting models in extremely demanding application conditions, in order to explore their application limits. Loss of robustness caused by inappropriate parameter transfers are analyzed through the variations in model performance.

3.1. Generalized Split-Sample Test

[16] The differential split-sample test (DSST) discussed in section 1.2 allows us to evaluate models in contrasted climatic conditions. The method usually follows three steps: (1) A small number of subperiods are selected according to one climatic characteristic (e.g., mean rainfall or temperature for the catchment). (2) The calibration-validation test is applied on these periods. (3) The validation performances are compared to evaluate whether they vary significantly when climatic characteristics differ between calibration and validation periods.

[17] This procedure has two limitations for obtaining robust and generalizable conclusions. First, it requires knowing in advance which climatic characteristics most likely play a key role in limiting the parameter set transfer. If the influence of different characteristics is tested, it is often difficult to compare the results because the subperiods used are different. Indeed they are selected according to the climatic property studied: for example, the driest period may differ from the warmest one. Second, the number of transfer tests is usually small, as often only two or three contrasted periods can be identified. This limits the possibility of drawing general conclusions and discovering the main drivers of parameter transferability from the results themselves. Indeed, it might be hard to distinguish the effect of the climate difference from other aspects potentially influencing parameter transfer.

[18] To overcome these limitations, we propose a generalization of the standard SST and DSST schemes. The objective is to test the model in as many and as varied climatic configurations as possible, including similar and contrasted conditions between calibration and validation. The problem was approached the other way round compared to what is usually done: numerous tests of parameter transfer were carried out and the results were analyzed to determine afterward whether the variations in the transfer quality were related to climatic aspects. This approach will be called the generalized split-sample test (GSST) hereafter.

[19] The GSST procedure simply consists of a series of calibration-validation tests on subperiods of equal length, considering all possible configurations. This procedure is based on the following steps (see Figure 3).

[20] 1. A sliding window of the chosen length (5 years on the graph) is used to define subperiods. Between two periods, the window is moved by 1 year (i.e., one hydrological cycle), thus allowing the subperiods to overlap. In Figure 3, these subperiods are the dark gray bars, while light gray represents the remaining part of the time series.

[21] 2. The hydrological model(s) are calibrated on each subperiod using a previously selected function. This provides one parameter set θ per period. At this step, any objective function or calibration algorithm can be used.

[22] 3. For each calibration subperiod, the optimized parameter set is used to perform all the possible validation tests on independent subperiods. Validation subperiods overlapping with the calibration one are not considered to ensure strict independence of calibration and validation conditions (see Figure 3). Moreover, a reference flow series is simulated for the calibration period using the parameter set obtained after calibrating the model on that period. Note that the number of validation tests will not be the same for all calibration periods. But this is not a problem as all results will be analyzed together.

Table 2. Overview of the Characteristics for the Three Models Tested[a]

	GR4J	MORDOR6	SIMHYD Plus Routing
Number of free parameters	4	6	7
Structure overview: production	A rainfall interception by PE, a nonlinear SMA store, an intercatchment GW exchange function	A rainfall correction factor, a nonlinear SMA store, a lower evaporative store	An interception store, a SMA nonlinear store
Structure overview: transfer	Two unit hydrographs, a nonlinear routing store	A direct flow component, an infiltration store, a linear routing store, a unit hydrograph	A groundwater store, a nonlinear routing store, a convolution delay
Sources and first publications	*Perrin et al.* [2003]	*Mathevet* [2005] (original MORDOR version from *Garçon* [1996])	*Chiew et al.* [2002] (with simplifications from *Tan et al.* [2005])

[a]PE, potential evapotranspiration; SMA, soil and moisture capacity; GW, groundwater.

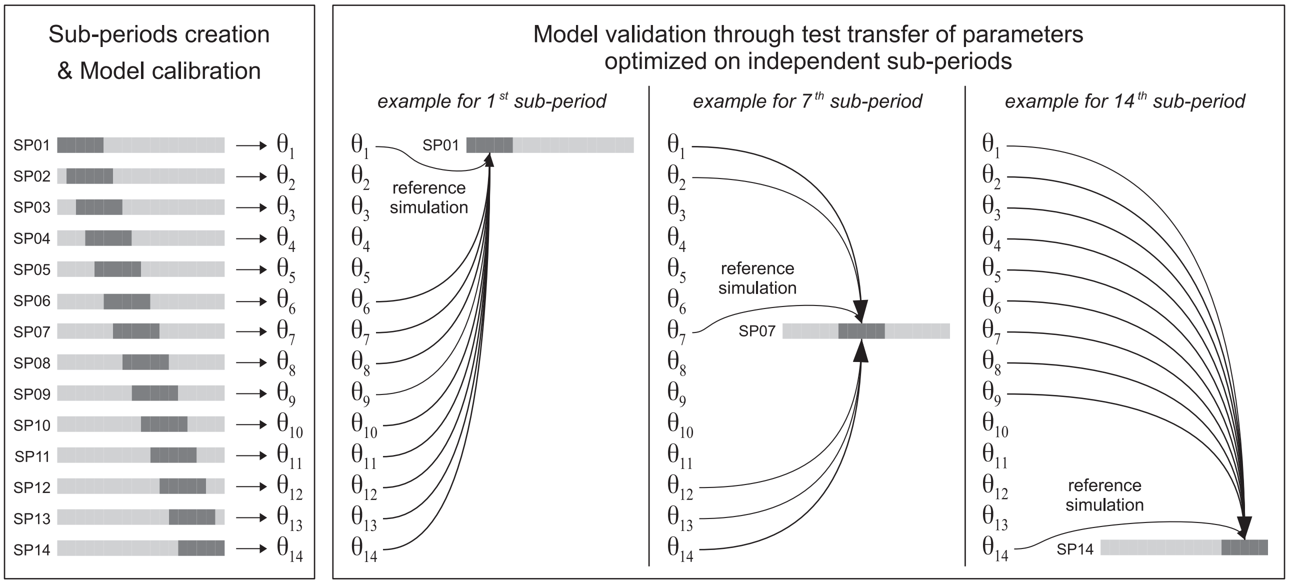

Figure 3. Illustration of the proposed generalized split-sample test (GSST) procedure (example with 18 years available and 5 year subperiods).

[23] To our knowledge, the SST, in which periods are not a priori selected but all combinations are tested, had not been used in this way before. The GSST procedure is meant to overcome the limitations previously mentioned: (1) It leads to a large number of test cases for the analysis, providing a continuum of the possible range of climatic differences existing in the observed series; for instance, with an 18 year long time series, sampling into 5 year periods and 3 year periods generates 90 and 182 SSTs, respectively. (2) It is less subjective because no choice is made before the tests. (3) The influence of changes in any climate characteristics on model robustness can be studied (ranging from mean interannual variables to indices characterizing the mean seasonal contrasts within the year). Indeed, all possible configurations are tested and lead to a unique list of validation performances. Switching between rainfall and temperature in the analysis simply means expressing these performances relative to one or the other climatic characteristic (the common practice where periods are selected with respect to one characteristic or the other leads to results which are not directly comparable).

[24] It could be argued that if the number of tests is increased there will be considerable redundancy in the tests carried out since the subperiods are not independent. Actually, our intention was to multiply the number of SSTs to study the entire range of climatic differences available between periods, even if each of them does not radically differ from all the others. The important point is that the calibration and validation periods are actually independent, which is in agreement with the original SST scheme.

3.2. Which Criteria Can Quantify the Extrapolation Capacity of a Given Parameter Set?

[25] When a model is used to simulate discharges, errors will arise: (1) for reasons which were already noticeable during calibration (data and model structure errors, identifiability issues, etc.) and (2) by the move from the calibration period to another period leading to the use of less than optimal parameters for this application period [*Merz et al.*, 2011]. In this study, we investigate this second aspect under a wide range of conditions. Separating these two sources of error is essential to achieving an informative evaluation of the extrapolation capacity of hydrological models: a model may work well in calibration but show poor transposability over time.

[26] Let us consider a split sample test where a parameter set θ is transferred from a period D ("donor", i.e., calibration) to a period R ("receiver", i.e., validation). With these notations, the root-mean-square error (RMSE) the Nash-Sutcliffe efficiency (NSE) [*Nash and Sutcliffe*, 1970] and the bias on total volumes can be written as

$$\mathrm{RMSE}_{D\to R} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\hat{Q}_{R,k}[\theta_D] - Q_{R,k}\right)^2} \tag{1}$$

$$\mathrm{NSE}_{D\to R} = 1 - \frac{\sum_{k=1}^{n}\left(\hat{Q}_{R,k}[\theta_D] - Q_{R,k}\right)^2}{\sum_{k=1}^{n}\left(\overline{Q}_R - Q_{R,k}\right)^2} \tag{2}$$

$$\mathrm{BIAS}_{D\to R} = \frac{\sum_{k=1}^{n}\hat{Q}_{R,k}[\theta_D] - \sum_{k=1}^{n}Q_{R,k}}{\sum_{k=1}^{n}Q_{R,k}} \tag{3}$$

in which $Q_{R,k}$ is the observed discharge at time step k on period R, $\hat{Q}_{R,k}[\theta_D]$ the simulated discharge at time step k on period R using the parameter set θ optimized on D, and n is the total number of time steps in period R.

[27] The advantage of the relative formulation of bias is to provide values that are comparable between periods and catchments. In their study on the time stability of parameters, *Merz et al.* [2011] plotted model bias in validation and

in calibration on the same figure to show that their trend on model error was indeed caused by the parameter transfer.

[28] Another way to emphasize the performance losses caused by the parameter transfer is to study the evolving performance from calibration to validation. $\mathrm{RMSE}_{D\to R}$ or $\mathrm{NSE}_{D\to R}$ values for different D periods but a single R period can be directly compared since all errors are calculated on the same time steps. Differences or ratios can be computed to highlight the quality of a given parameter set compared to another. However, the RMSE is dependent on the mean volume and will tend to be greater for periods (or catchments) showing larger discharges. NSE is built around a ratio between the squared model error and the variance of observed flows. Under certain conditions, one can assume that changes in variance or volumes between periods have a limited impact on the comparison results. Limitations appear when the periods compared show contrasted climatic properties and hence contrasted flow levels. It becomes even more complicated when results from different catchments are analyzed together. Contrary to what is often done in the literature, we decided not to use differences in NSE to conclude on the influence of changes in climate on parameter transferability. Instead, we defined the following model robustness criteria (MRC):

$$\mathrm{MRC}_{D\to R} = \frac{\varepsilon_{D\to R}}{\varepsilon_{R\to R}} - 1 \tag{4}$$

where ε is the objective function to be minimized during calibration. The main idea is that the quality of a given parameter set is assessed relative to a reference set, obtained through calibration. $\varepsilon_{D\to R}$ is one estimate of the model error on period R using the parameters calibrated on period D (e.g., $\varepsilon_{D\to R} = \mathrm{RMSE}_{D\to R}$). It varies depending on the ability of the parameter set optimized on period D to simulate discharges on period R. $\varepsilon_{R\to R}$ should be the smallest value of ε achievable on period R with the model. $\varepsilon_{D\to R}$ and $\varepsilon_{R\to R}$ are comparable since they are computed on the same "receiver" period. MRC should theoretically be positive. Its interpretation is straightforward. It takes a value of 0 if the parameter set optimized on D gives the same fit it would have if it was calibrated on R. The higher the value, the less suitable the parameter set for the receiving period R. For example, a MRC value of 0.2 means that there has been a 20% error increase due to the use of a transferred parameter instead of the optimal one. Note that a negative value would mean that the parameter set optimized on period D performs better on period R than the reference set optimized over period R. This would be the indication of a problem in parameter optimization on period R, where the global optimum had not been identified properly. Here this happens in a very limited number of calibration runs (less than 1%), indicating that this has only a marginal influence on our results.

[29] The formulation of MRC overcomes most of the difficulties mentioned previously in comparing performances. The only requirement to allow comparing MRC values obtained under various conditions is that the ratio $\varepsilon_{D\to R}/\varepsilon_{R\to R}$ must be independent from the period or catchment characteristics (in terms of volumes or dynamic). For example, $\varepsilon = \mathrm{RMSE}$ can be used, whereas $\varepsilon = -\mathrm{NSE}$ cannot since a variance term would remain in MRC and results from various catchments could therefore not be mixed. When ε is the mean square error, MRC is a modified version of –NSE in which the benchmark model at the denominator has been changed [see *Lerat et al.*, 2012]. In the NSE formulation, this benchmark is the mean observed flow value over the R period while in MRC, it is the flow simulated by the tested model (e.g., GR4J) using the parameter set optimized on R. Provided ε shows the expected properties mentioned above, MRC is fully comparable over various conditions in terms of climate, catchment scale or dynamic. MRC values should not be significantly affected by the imperfect fit of the model to a specific period (caused by data and model structure errors). Conversely, the influence of using transferred rather than optimized parameters is highlighted. Therefore the climatic extrapolation capacity of a parameter set can be quantified and the results analyzed together on a large number of case studies.

3.3. Methodology for Analyzing the Results

[30] Variations in MRC values were analyzed relative to the differences in climate between the calibration and validation periods, aiming to investigate the potential link between the quality of parameter transfers and the variation in climate from calibration to validation. Changes in climate were expressed as ratios (e.g., 10% less rainfall) or differences (e.g., $+1°$C). We built graphs where each MRC value was plotted against the corresponding change of the selected climate variable. An example is shown in Figure 4 for the MORDOR6 model and rainfall variations. Figure 4a shows several parameter transfer tests carried out on a single receiving period. Because all values on the x and y axes are relative, the results for all the other receiving periods can be plotted on the same graph. The plotting procedure was then repeated for all the catchments: this provides a large cloud of points, as shown in Figure 4b. To extract the information contained in the graph, the cloud was divided into vertical slices with the same number of points (instead of slices of equal width, which would be less robust). In each slice, the distribution of MRC values is summarized by a box plot (showing the 0.05, 0.25, 0.5, 0.75, and 0.95 percentiles) (see Figure 4c). Because of the lesser density on the left and right sides of the graph, the slices have different widths and the box plots are therefore not evenly spaced. Nevertheless, every point has a corresponding box plot and the voids appearing on the graph are always covered by the neighboring box plot. The vertical spread for each box plot indicates the range of performance loss obtained for the corresponding test conditions. Comparing the relative position of the box plots indicates whether a change in climate from calibration to validation causes loss of robustness due to inappropriate parameter transfers.

[31] To better understand the interpretation of this graph, let us take the example of the box plots shown in Figure 4c. The climate variable used here is mean rainfall. The box plot obtained for about +30% in rainfall (right-hand side of the graph) represents the cases of donor periods that are wetter than the receiver period (i.e., parameter transfer from wet to dry). This box plot is compared to the box plot for 0% change in rainfall. The latter indicates the "usual" performance loss when parameters are transferred under similar climate conditions (median of about 20%). The higher values shown by the box plot on the right indicate

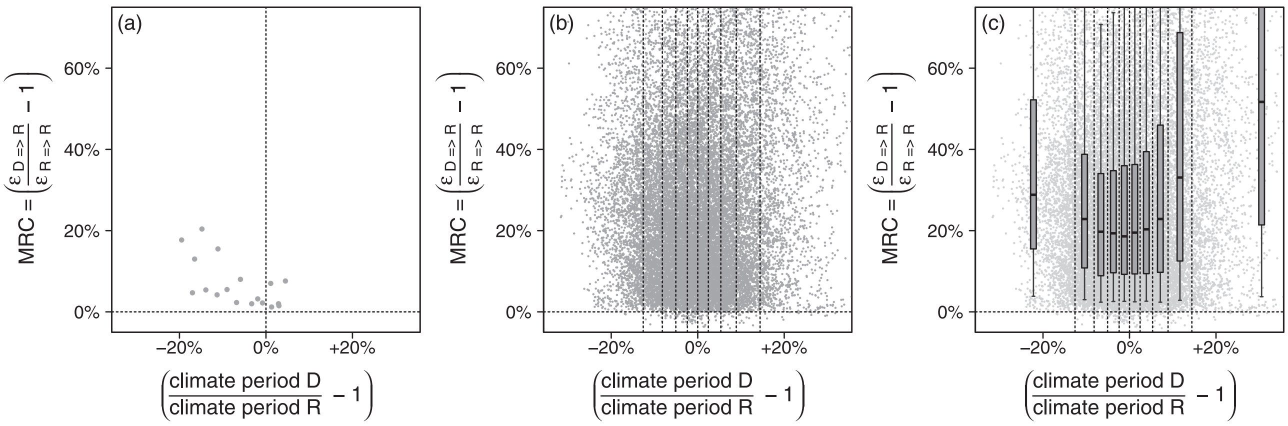

Figure 4. Procedure followed to illustrate the results. The relative loss of performance is plotted against the relative evolution of climate conditions. (a) Dotty plot for a single period, (b) dotty plot for all periods of all catchments, and (c) summary of dotty plots as box plots.

that using a parameter set calibrated on a period 30% wetter than the validation climate is likely to reduce the quality of the model's simulations (here the box plot median increases from 20% to 50%). In a way, this representation puts into perspective the loss of efficiency due to a common transfer and the loss stemming from the climatic extrapolation conditions.

3.4. Crash Test Conditions

[32] The models' parameters were calibrated to minimize the following objective function:

$$\varepsilon = \mathrm{RMSE}[\sqrt{Q}](1 + |\mathrm{Bias}|) \quad (5)$$

[33] The combination of $\mathrm{RMSE}[\sqrt{Q}]$ and bias gives weight to dynamic representation as well as water balance. The models' ability to simulate mean runoff is of particular importance in the context of climate change impact studies. Using square-root-transformed flows to compute the RMSE reduces the influence of high flows during calibration and was found to give a good compromise between alternative criteria [*Oudin et al.*, 2006a]. Given the small number of free parameters in the tested models, we used a prior systematic inspection of the parameter space followed by a simple steepest descent local search procedure to determine the most likely zone of convergence. This approach proved efficient for such parsimonious models compared to more complex search algorithms [see *Edijatno et al.*, 1999; *Mathevet*, 2005].

[34] Choosing the subperiod length used in the sampling methodology (see section 3.1) is a difficult task (see, e.g., the discussions by *Yapo et al.* [1996] and *Anctil et al.* [2004]). The calibration period should be long enough to allow for correct parameter determination. At the same time, using overly long periods may play against the study's objectives, as it would reduce the contrast between periods. Also, the number of independent test periods per catchment decreases when the subperiod length increases. We repeated the work with 10 year long and 5 year long periods and will present here the results obtained with the 10 year long calibration periods. Because of the length of available records (more than 30 years) and the high variability of the Australian climate (compared to other parts of the world), using 10 year periods still provides significant differences in mean climate (see section 2.1 and Figure 2) and therefore does not change the conclusions.

[35] The number of tests that could be made on the catchments following the GSST procedure depended on data availability. Subperiods with more than 10% missing values were excluded from the tests. As a result, 12 catchments were not considered from the initial set because of insufficient record length or excessive gaps in data. Therefore, a total of 216 catchments were used in the tests. Fifty tests or more were made for 183 catchments (85% of the set). A maximum of 156 split-sample tests was reached for 134 catchments (62% of the set), corresponding to permutations of 10 year periods over the 1974–2006 period.

3.5. Climate Variables Investigated

[36] Various climate variables can be used in the analysis. Examples include mean annual or seasonal rainfall, PE or temperature, the number of extreme rain or drought events, aridity index, etc. Here we based the analysis on the common variables P, PE, and T with averages computed over the test period. For each split sample test, we determined the changes in mean P, PE and T between calibration and validation periods and plotted these ΔP, ΔPE, and ΔT against each other. Scatter plots were summarized in box plots using the representation introduced in section 4.3. Figure 5 shows the existence of correlations between variables: the link between ΔT and ΔPE is not surprising; their anticorrelation with ΔP indicates that an increase in temperature (or PE) on these catchments generally coincides with a decrease in precipitation.

4. Results

4.1. Calibration Results

[37] We first provide an overview of calibration performance to evaluate the quality of the reference parameter sets. The box plots in Figure 6 show the distribution of calibration performance in terms of NSE calculated on root-square discharges (Figure 6, left) and bias on total volume (Figure 6, right). The calibration results for our tests are shown in black, corresponding to the objective function ε defined in

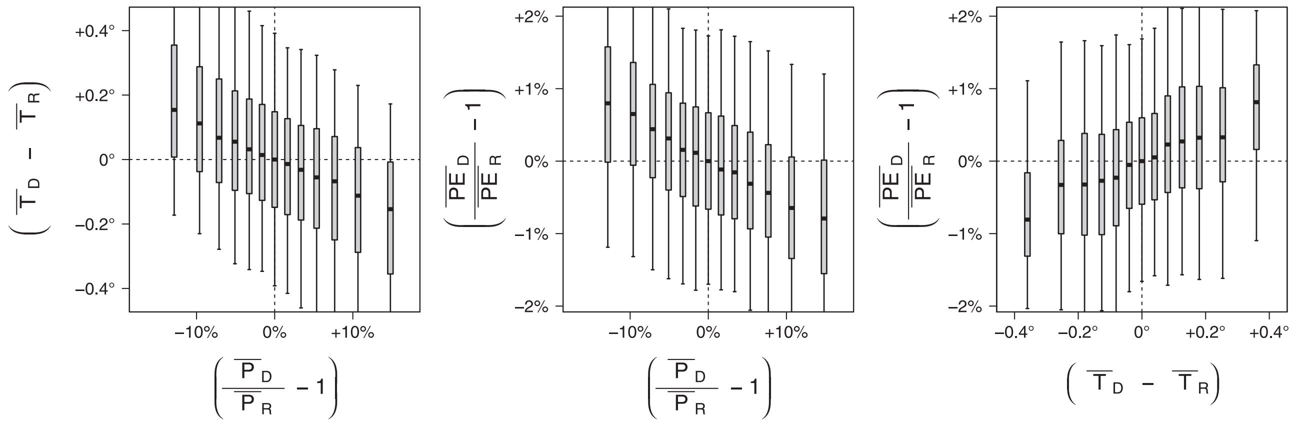

Figure 5. Correlation between variations in P, PE, and T for all the split-sample tests over the entire catchment set.

equation (5). As a source of comparison, we plotted the results obtained for $\varepsilon = \mathrm{RMSE}[\sqrt{Q}]$ in gray. We discussed above how the comparison of NSE values between periods or catchments relies on strong assumptions, not always valid in our context. These results only aim at checking that the models perform reasonably well in calibration. We note the benefit from adding the bias constraint in the objective function and the limited consequences on $\mathrm{NSE}[\sqrt{Q}]$ values, which is in agreement with the results found by *Viney et al.* [2009] and the theoretical comments made by *Gupta and Kling* [2011].

4.2. Which Climate Property Causes Problems for Parameter Transfer?

[38] We applied the GSST procedure over 10 year periods for 216 catchments. As in Figure 4, the variations in model robustness criteria (MRC defined in equation (4)) are plotted against the differences in climate between calibration and validation for all catchments. These graphs are provided in Figure 7. To facilitate the analysis of Figure 7, the individual points are not shown. The nine graphs correspond to the three models and three climate variables considered (mean P, PE, and T). The shape of each scatterplot is then analyzed to determine which climatic characteristics influence parameter transferability and to what extent this affects model performance.

[39] First, we note the relatively wide vertical range of the box plots. This indicates that performance losses can be low or high depending on the period or catchment, independently from the mean climate evolution. The center part of each graph gives an overview of the level of error obtained when parameters are transferred under similar climate conditions. We observe a median (thick black line within the box plot) at about 18–20% for GR4J, MORDOR6 and SIMHYD. These values mean that transferring parameters to another period with a similar mean climate leads to an 18%–20% increase in model error ($\varepsilon = \mathrm{RMSE}[\sqrt{Q}](1 + |\mathrm{Bias}|)$) on average compared to calibration. This performance loss is a combination of two aspects. First, there is an incompressible loss of performance when going from calibration to validation due to the model's approximations (inputs, parameters and structure). Second, average conditions such as the total rainfall can remain stable between the two periods, whereas meaningful differences exist (e.g., daily variability in rainfall and runoff).

[40] We can now evaluate whether an additional loss occurs when the climate conditions between the "donor" and "receiver" periods differ significantly. In spite of the vertical spread, trends are visible on the three left graphs in Figure 7 (case of mean rainfall). Considering the large number of points in the scatterplot, these trends on the box

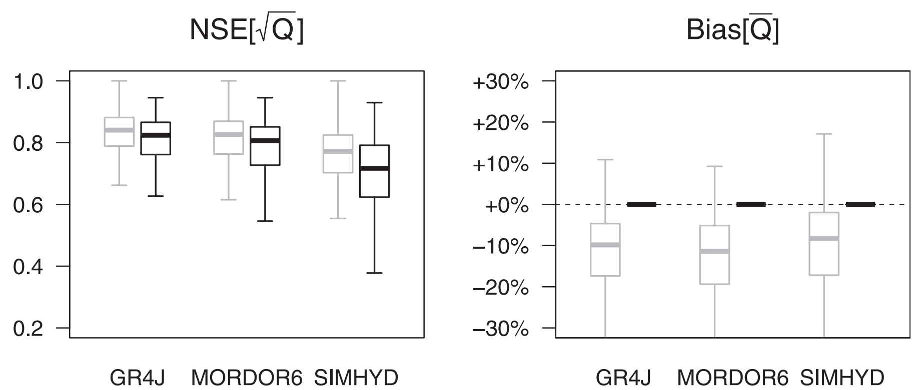

Figure 6. Calibration performance on the entire catchment set with two objective functions: $\varepsilon = \mathrm{RMSE}[\sqrt{Q}]$ (in gray) and $\varepsilon = \mathrm{RMSE}[\sqrt{Q}](1 + |\mathrm{Bias}|)$ (in black).

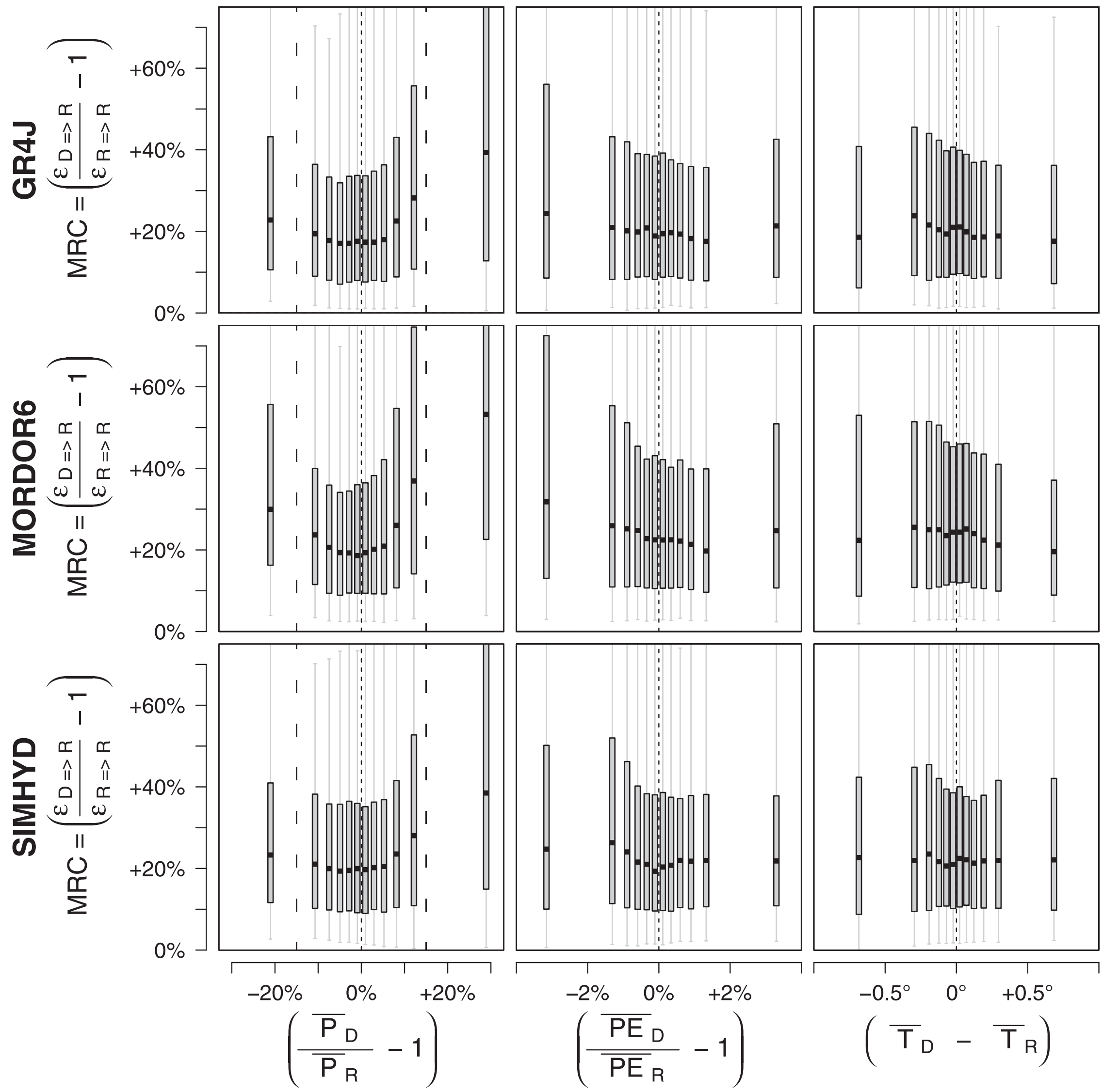

Figure 7. Study of models deficiencies: Performance loss due to the parameter transfer plotted against changes in mean P, PE, and T for the 216 catchments, with $\varepsilon = \mathrm{RMSE}[\sqrt{Q}](1 + |\mathrm{Bias}|)$.

plot medians, but also on the 25th and 75th percentiles, are definitely not the result of isolated cases and should be considered significant, although they are small. This indicates that a link can be established between model performance (estimated through the objective function) and the difference in mean rainfall between calibration and simulation periods. The performance loss is not symmetric for positive and negative changes in mean rainfall. We can observe a greater loss when donor periods are wetter than receiver periods. In addition, the trend seems to be the strongest for MORDOR6 and the weakest for SIMHYD. For example, a +15% difference in mean rainfall leads to an average increase in *MRC* from about 18 to 30% for GR4J, 18 to 40% for MORDOR6 and 20 to 30% for SIMHYD. Interestingly, similar trends are not visible for PE and T, in spite of the correlation observed between ΔP, ΔPE and ΔT (although we notice slightly larger performance losses when $\Delta\mathrm{PE} < 0$). These differences likely result from the fact that the environment in southeast Australia is water limiting (as opposed to energy limiting), which may cause a higher sensitivity of flows to rainfall compared to PE and temperature (sensitivity measured through the objective function).

[41] The results obtained here are in general agreement with the findings reported by *Vaze et al.* [2010b], although the methodology has been modified to provide better robustness on conclusions (more tests and a different analysis procedure). Amplitudes cannot be compared directly but the general shapes are similar: a change in mean rainfall reduces parameter transferability and this transferability seems better from dry to wet than vice versa.

4.3. Consequences of These Deficiencies on Volume Predictions

[42] In this section, we attempt to specifically quantify the ability of models to correctly predict mean runoff over a period in the context of varying climatic conditions. Mean runoff is a basic but nonetheless crucial indicator for

water resources management under current conditions as well as potential future ones.

[43] Bias values have the advantage of being directly comparable. Therefore, we plotted the bias in validation against the climate differences between calibration and validation (results not presented here). We found trends for all three climatic characteristics (P, PE, and T), although they were stronger for changes in rainfall. All models showed a tendency to overestimate flows in the validation period when the calibration period was wetter and cooler and to underestimate flows when the calibration period was dryer and warmer. Nevertheless, no clear conclusion could be drawn on the individual role played by each climate variable because of the dependency observed between ΔP, ΔPE, and ΔT (see Figure 5). In spite of its many advantages, the representation method used to build Figure 7 has one drawback: only one climate characteristic can be analyzed at a time.

[44] To overcome this limitation and to make sure that the results can be compared with studies carried out in alpine regions (i.e., energy-limited conditions), we used a representation where two climate variables could be displayed at a time. We could then determine if only one or both variables influence the validation bias when they vary. Figure 8 shows the median bias on simulated flows plotted for all possible combinations of two climate variables between ΔP, ΔPE, and ΔT. Note that this representation illustrates the models deficiencies in relation with calibration

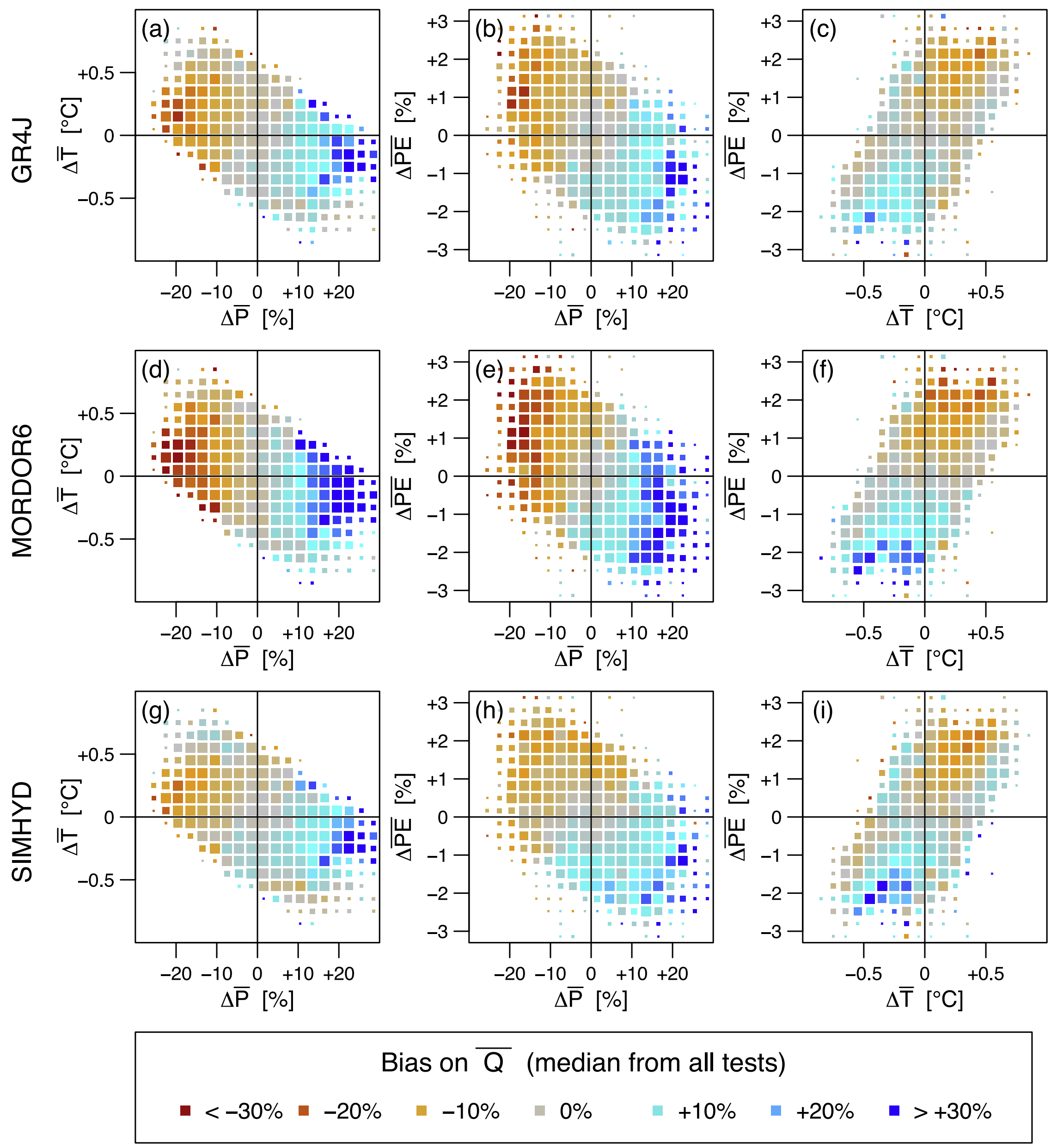

Figure 8. Study of models deficiencies: bias on simulated discharges as a function of changes in P, PE, and T during parameter transfer (median bias from an aggregation of results on 216 catchments; the changes in climate are computed with the climate of the receiver period as reference).

conditions. It should not be interpreted as a plot of the actual flow sensitivity to variations in climate. Compared to the graphs previously presented, this representation also has limitations: (1) only one percentile of the box plots can be shown at a time; (2) the cloud gridding does not ensure an equal number of points in each cell but equal distribution along the *x* and *y* axes. To limit the impact on interpretations of this second limitation, the cell sizes are modified to reflect the number of points available. The size is proportional to the number of points under 50 values and is fixed above 50 values (e.g., few points are available for $\Delta T > 0$ and $\Delta PE < 0$ or for $\Delta T > 0$ and $\Delta P > 0$).

[45] The median bias values obtained for all three models over the 216 catchments show the following variations. If we consider absolute values of bias, we note a symmetrical pattern along the *y* axis for Figures 8a, 8d, and 8g. Indeed, for constant T, the absolute bias increases with changes in P, while the contrary is not true (i.e., no symmetry along the *x* axis). As a result, we can conclude that on average for the data set, changes in P influence the bias more than changes in T. This is to be expected as all the 216 catchments used in this study are in a water-limited area where relative changes in P have larger impacts on runoff than changes in T or PE. The situation is not as clear on Figures 8b, 8e, and 8h (ΔPE and ΔP). Changes in P and PE have a combined effect on the validation bias, although the greatest changes in P ($\pm 15\%$) seem to have more impact than the greatest changes in PE ($\pm 2\%$). Finally, ΔPE and ΔT also have some effect on the validation bias. None of them seems to be more influential than the others (with respect to the points available). A parallel can be made with the results from *Oudin et al.* [2006b], who showed that biased rainfall inputs have a larger impact on models efficiency than biased PE inputs.

[46] The comparison of results between GR4J, MORDOR6 and SIMHYD indicates that even though the intensity varies depending on the model's structure, the general behaviors are similar. This constitutes a difference with the findings of *Vaze et al.* [2010b], who obtained contrary results depending on the model used. Here we observed the smallest and strongest absolute bias for the SIMHYD and MORDOR6 models, respectively, when P changes between calibration and validation. This relation between the change in climate from calibration to validation and the simulation bias was also observed by *Merz et al.* [2011], although the study area was completely different (Australia versus Austria). Further investigations are needed to determine whether these results are related to the model and objective function used in model calibration. On average, we observed that a 20% absolute bias was introduced when mean rainfall differed by 10%–20% and PE differed by 1%–2% between calibration and validation.

4.4. Are Models Deficiencies Similar on All Catchments?

[47] A significant link between climate difference during parameters transfer and bias on simulated discharges was found when the results from the entire data set were analyzed together. However, we do not have information on the homogeneity of this link between different catchments. Interpretations made on a single catchment are always difficult to generalize. However, the use of GSST provides an average of 125 SSTs per catchment and calibrations were made on 10 year periods. Therefore, conditions can be considered sufficient to make a rough analysis at the catchment scale of the consequences that parameter time transfer may have on model efficiency.

[48] For each catchment, a library of validation tests is available and can be classified according to differences in climate between calibration and validation. We can estimate a median value of simulation bias for a specific change in climate (e.g., +10% in rainfall, −0.5°C, etc.). A series of maps can then be built for a quick overview of the spatial variability of parameter transferability issues (one map per climate difference). Here we present the results for changes in mean rainfall as they have the greatest impact on bias for our catchment set. Because the results are quite similar for the three models tested, only the results for GR4J are shown here. The maps in Figure 9 are for $\Delta P = -10\%, 0\%, +10\%$, where each catchment is represented by one symbol. The triangle direction (upward/downward) shows the sign, while the color indicates the intensity of the bias obtained during simulations. Black crosses correspond to cases where this level of rainfall difference was not available for the catchment during the GSST procedure.

[49] In accordance with the previous findings, for a great majority of catchments, we found that simulation bias was close to zero when the mean rainfall was similar between calibration and validation. Mean flows were overestimated when rainfall was greater during calibration than validation and vice versa. Some catchments were exceptions to these generic results with two possible situations: (1) both ΔP and ΔPE affect the simulation bias (see in Figure 8). For example, the combination of $\Delta P = 0$ and $\Delta PE < 0$ can lead to a positive bias or the combination of $\Delta P < 0$ and $\Delta PE > 0$ to a negative bias; (2) other climatic characteristics may differ and affect the results, although these changes are not detected when using mean P, T, and PE; (3) other issues which are catchment specific and unrelated to climate variations may induce robustness loss due to parameter transfer. Despite data verification, cases of changes in input quality and availability or changes in rating curves between periods are never completely avoidable.

[50] Generally, we found that the magnitude of models deficiencies differs between catchments. For some catchments, transferred parameters were suitable for other periods, even with contrasted climate. For other catchments, a clear relation was established between simulation quality (particularly bias on total volume) and the differences in climate between calibration and validation. In spite of various attempts, we could not relate the intensity of these changes with catchment or climate characteristics. However, we noted that the largest models deficiencies occurred on catchments with relatively low annual runoff yield. Besides, as mentioned in section 1.2, temporal changes of nonclimatic characteristics (e.g., land use, quality of input data, etc.) may lead to performance losses, with impacts possibly increasing with the time elapsed between calibration and validation periods. This may contribute to explain the cases of strongest models deficiency, potentially caused by the combined negative impact of differences in climate conditions and in other aspects between periods.

[51] Two examples of the possible situations are presented in Figure 10. These catchments should not be

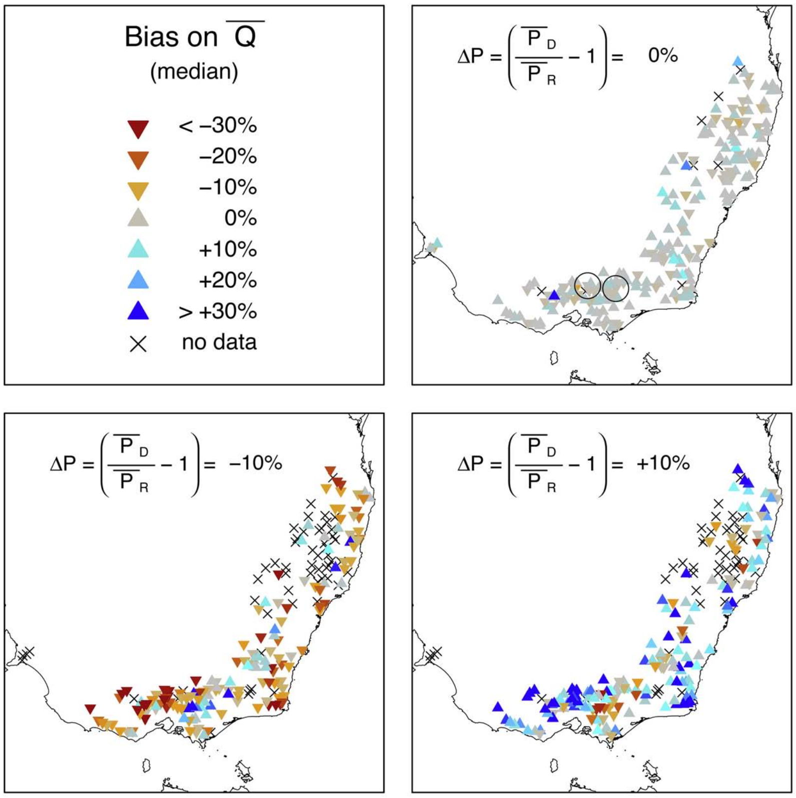

Figure 9. Study of models deficiencies: maps of median bias on discharges simulated by GR4J for ΔP $= -10\%, 0\%, +10\%$

considered as representing average conditions obtained over two subgroups. Indeed, we obtained a large variety of results depending on the catchment analyzed. In some cases, the simulation quality varied independently from any climatic characteristic, while distinct correlations were found on others. The results for the Rose River at Matong North (179 km^2) and the Pranjip Creek River at Moorilim (818 km^2) catchments are used as an illustration. The location of these two catchments is shown by black circles on the first map in Figure 9. They were chosen because they illustrate some of the contrasted situations we encountered, with similar climate variations. Figure 10 (left) shows the variations of *MRC* with respect to changes in rainfall between calibration and validation. The results for all the calibration-validation combinations are plotted on the graph. The performance losses are clearly stable for the first catchment, while they are greatly influenced by changes in rainfall for the second one. For both catchments, the solid black dots are the results of parameter transfers to the receiver period 1978–1987. The corresponding simulations are illustrated in Figure 10 (right) as mean monthly flows over this 10 year period. The solid and dashed lines correspond to the observation and the reference simulation obtained from calibration, respectively. The envelope with horizontal shading (vertical shading) shows the range of simulated values when the calibration period was drier (wetter) than the validation period. All simulations are relatively similar in the case of the Rose River (the two envelopes even overlap sometimes). Contrary to this, the range of simulated values is very large for the Pranjip River and the curves of the different simulations are positioned in accordance with the climate difference. When classified from most overestimating to most underestimating, the curve order is indeed almost identical to the classification of climate differences between calibration and validation.

5. Discussion

5.1. Methodological Choices

[52] The objective function ε used for these results was RMSE$[\sqrt{Q}](1 + |\text{Bias}|)$. Other criteria such as RMSE[Q] $(1 + |\text{Bias}|)$ or directly RMSE$[\sqrt{Q}]$ and RMSE[Q] were tested. The results are not presented here but the overall shapes obtained in Figures 7 and 8 were similar. While here we focus more on the diagnostic part, studies investigating the reduction of robustness loss during parameter transfer may be the topic of future investigations. Among possible sources are the use of more complex objective functions considering, for example, error heteroscedasticity as suggested by *Thyer et al.* [2009] or *Schoups and Vrugt* [2010].

[53] In addition to keeping the approach simple, we aimed at obtaining an overall and robust view of the topic. As a result, we compared three conceptual models with different structures (although all were lumped) and used a relatively large catchment set. Considering the potential implications of these results, it is indeed important to determine whether the findings for one model over one catchment are isolated or similar in a number of other cases. With the same idea of maximizing the number of points for

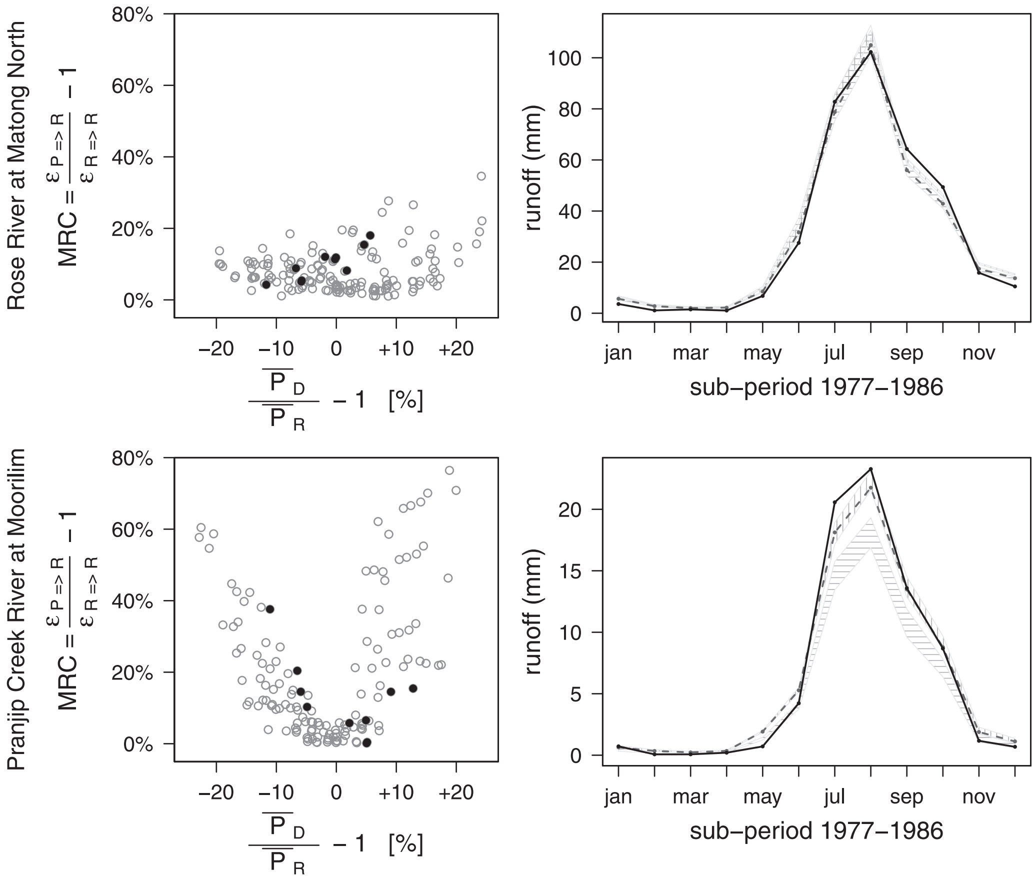

Figure 10. Examples of test results for two catchments: Rose River at Matong North (up) and Pranjip Creek River at Moorilim (bottom). (left) Performance losses during parameter transfer expressed relative to changes in mean rainfall. (right) Range of monthly flows simulated over the 1978–1987 period for various parameter transfers (horizontal and vertical shading indicates the simulation envelope for calibration on wetter or drier climate than validation).

analysis, we authorized periods to overlap in our sampling methodology. As a result, each point will not be entirely independent of the others with which it shares one calibration year or more. Because are more interested in the general picture than computing statistical relation, this does not affect the interpretation of results. Besides, forbidding overlap would require selecting a period, which would reduce the range of climatic contrasts tested for each catchment.

5.2. Results

[54] Using a parameter set under climatic conditions differing from the calibration period can lead to decreased simulation qualities. For instance, we observed performance losses when rainfall changed during parameter transfer and these losses were greater for wet to dry transfers than the other way round. These results corroborate the findings of *Vaze et al.* [2010b] and have direct implications on the use of hydrological models under contrasted climatic conditions (among which, climate change impact studies). Moreover, we observed models deficiencies in the form of an average tendency to overestimate (underestimate) discharges when parameters are optimized in a wetter (drier) climate. This is consistent with the results from *Merz et al.* [2011], although the characteristics of the two catchment sets used differ significantly. High biases were obtained for ranges of climate differences between optimization and application stages similar to possible future climate evolutions in the next decades (as projected by climate models). Since mean flow estimates are crucial indicators in water management plans for current and future conditions, these results should therefore be a particular source of concern. At the same time, we found that the magnitudes of these models deficiencies were not homogeneous but somewhat catchment dependent. This might explain why there is currently no consensus in the literature on this question, but it surely makes the situation more complex for hydrologists.

[55] Further work is needed to understand the hydrological mechanisms behind the robustness issues observed in southeast Australia because of inappropriate parameter transfer. The analysis of the link between parameter values and calibration conditions should contribute to that (this is out of the scope of this paper but will be reported in due course). The possible causes for inappropriate parameter transfer are various (see section 2). Among these, we believe that incorrect simulation of the water budget might be a major one. When transferring parameters, we indeed assume that (1) the adjustments made during parameter calibration provide a satisfactory representation of the water budget and (2) they remain valid over time (i.e., the adjustments made on one period are suitable for another). However, this might not always be true

and could cause increased bias when parameters are transferred. For instance, intercatchment groundwater flows (IGF) can play an important role in the catchment water budget. Even when they are explicitly represented in the models, the IGF formulations are extremely simplified and the associated parameter value remains difficult to identify [*Le Moine et al.*, 2007]. Similarly, the water balance can be strongly affected by errors in inputs that models tend to compensate during calibration to close the water budget. However, these errors might change over time, as the volumes in action vary, which will impact the parameters' transferability if this is not taken into account [*McMillan et al.*, 2011]. Last, the sensitivity of runoff to climate is not straightforward in southeast Australia [*Potter and Chiew*, 2011]. It seems that the RR relationship can sometimes show a form of elasticity, i.e., the prevailing hydrological processes may change between periods of different regimes [*Harman et al.*, 2011]. If this is the case, parameters calibrated under a specific climate would not be representative of the processes that are active under other conditions.

[56] Numerous unknowns remain around the cases of robustness loss sometimes observed when a model is used in a changing climate. The study of other climatic characteristics than mean conditions and the comparison between problematic cases with nonproblematic cases may provide new leads for a better understanding of the phenomena involved. Further research is obviously needed to build models actually able to cope with nonstationary conditions in all catchments. Before this can be achieved, preliminary testing seems preferable to obtain a rough estimation of the parameters' transferability on a given catchment. The procedures proposed in this paper are one way to achieve such testing.

6. Conclusions

[57] Rainfall-runoff models are essential tools for the prediction of river flows. Once calibrated under historical climatic conditions, they are sometimes fed with forcings with different climatic characteristics. This raises questions on the validity of such parameter transfer or what could be called the climatic extrapolation capacity of hydrological models. Following the path opened by *Klemeš* [1986], recent research illustrates how significant it can be, with direct implications on the prediction quality. However, there is no consensus in the literature, as most research was carried out on isolated cases and the results are not always comparable. With this in mind, we proposed a generalized testing procedure (GSST) and the associated analysis methods, with the objective of obtaining robust interpretations on this topic.

[58] When applied to 216 catchments in southeast Australia using three conceptual RR models, this methodology led to the following results. Wide ranges of performance losses were observed between calibration and validation periods. Using the same error criteria as for parameter calibration, we found a tendency toward increased simulation error, with a greater difference in mean rainfall during the parameter transfer, but nothing of similar magnitude for changes in mean PE or temperature. Bias in total volumes was affected by changes in both mean rainfall and PE. We also observed a tendency to overestimate mean runoff when the calibration period was wetter (wet to dry parameter transfer) and to underestimate mean runoff when the calibration period was drier (dry to wet parameter transfer). Even though the tendencies were observed for a majority of the catchments in the data set, we found that their intensity could greatly vary between catchments.

[59] The results obtained here corroborate previous findings obtained by others on large catchment sets to highlight the possible lack of robustness when models are used under a changing climate. They have important implications when using hydrological models as decision-making tools in a wide range of applications (flood risk management, water availability, hydropower, climate change impact studies, etc.). Therefore, we believe that research should be pursued on the improvement of methods to diagnose parameter transferability under a changing climate. Further research is needed to apply similar testing procedures with other models and on different catchment sets (e.g., catchments which are energy limited instead of water limited like those studied herein) with the same objective: determining how relevant the errors in parameter transfers due to climate differences are compared to usual transfer errors in similar conditions. The differences may provide new insights into the behavior of models and hydrological systems in changing conditions. Further research is also needed to analyze what is happening in problematic cases, i.e., what makes the parameters unsuitable on a different period and look for solutions to reduce these robustness losses by means of model structure adjustment, the choice of objective function and/or constraints during parameter optimization. All these constitute exciting challenges for the coming decade on prediction under change that will be launched by the International Association of Hydrological Sciences.

[60] **Acknowledgments.** This study was part of a collaboration between research teams from LNHE, EDF R&D (France), CSIRO Land and Water (Australia), and Irstea HBAN (France). The authors would like to thank EDF R&D for supporting this study and CSIRO Land and Water for hosting the first author of the study for a 5 month period and for making available the large data sets used. The constructive review comments made by A. Efstratiadis, J. C. Refsgaard, and an anonymous reviewer on the first version of the manuscript helped improve the text and are gratefully acknowledged.

References

Anctil, F., C. Perrin, and V. Andréassian (2004), Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models, *Environ. Modell. Software*, *19*(4), 357–368, doi:10.1016/S1364-8152(03)00135-X.

Andréassian, V., C. Perrin, L. Berthet, N. Le Moine, J. Lerat, C. Loumagne, L. Oudin, T. Mathevet, M.-H. Ramos, and A. Valéry (2009), Crash tests for a standardized evaluation of hydrological models, *Hydrol. Earth Syst. Sci.*, *13*(10), 1757–1764, doi:10.5194/hess-13-1757-2009.

Arnell, N. W. (2011), Uncertainty in the relationship between climate forcing and hydrological response in UK catchments, *Hydrol. Earth Syst. Sci.*, *15*(3), 897–912, doi:10.5194/hess-15-897-2011.

Bastola, S., C. Murphy, and J. Sweeney (2011), Evaluation of the transferability of hydrological model parameters for simulations under changed climatic conditions, *Hydrol. Earth Syst. Sci. Discuss.*, *8*(3), 5891–5915, doi:10.5194/hessd-8-5891-2011.

Caballero, Y., S. Voirin-Morel, F. Habets, J. Noilhan, P. LeMoigne, A. Lehenaff, and A. Boone (2007), Hydrological sensitivity of the Adour-Garonne river basin to climate change, *Water Resour. Res.*, *43*, W07448, doi:10.1029/2005WR004192.

Chahinian, N., et al. (2006), Compilation of the MOPEX 2004 results, in *Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment-MOPEX*, edited by V. Andréassian et al., *IAHS Publ., 301*, 313–338.

Chiew, F. H. S., M. C. Peel, and A. W. Western (2002), Application and testing of the simple rainfall-runoff model SIMHYD, in *Mathematical Models of Small Watershed Hydrology and Applications*, edited by V. P. Singh and D. K. Frevert, pp. 335–367, Water Resour. Publ., Littleton, Colo.

Chiew, F. H. S., et al. (2008), Rainfall-runoff modelling across the Murray-Darling Basin: A report to the Australian government from the CSIRO Murray-Darling Basin Sustainable Yields Project, CSIRO, Canberra.

Chiew, F. H. S., J. Teng, J. Vaze, D. A. Post, J. M. Perraud, D. G. C. Kirono, and N. R. Viney (2009), Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method, *Water Resour. Res.*, *45*, W10414, doi:10.1029/2008WR007338.

Choi, H. T., and K. Beven (2007), Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, *J. Hydrol.*, *332*(3–4), 316–336, doi:10.1016/j.jhydrol.2006.07.012.

Coron, L., V. Andréassian, M. Bourqui, C. Perrin, and F. Hendrickx (2011), Pathologies of hydrological models used in changing climatic conditions: A review, in *IUGG 2011—Hydro-climatology: Variability and Change, IAHS Publ.*, *344*, 39–44.

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

Donnelly-Makowecki, L., and R. Moore (1999), Hierarchical testing of three rainfall-runoff models in small forested catchments, *J. Hydrol.*, *219*(3–4), 136–152, doi:10.1016/S0022-1694(99)00056-6.

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

Garçon, R. (1996), Prévision opérationnelle des apports de la Durance à Serre-Ponçon à l'aide du modèle MORDOR. Bilan de l'année 1994–1995, *La Houille Blanche*, *5*, 71–76, doi:10.1051/lhb/1996056.

Görgen, K., et al. (2010), Assessment of Climate Change Impacts on Discharge in the Rhine River Basin: Results of the RheinBlick2050 Project, *CHR Rep. I-23*, Int. Comm. for the Hydrol. of the Rhine Basin, Lelystad, Netherlands.

Gupta, H. V., and H. Kling (2011), On typical range, sensitivity, and normalization of mean squared error and Nash-Sutcliffe efficiency type metrics, *Water Resour. Res.*, *47*, W10601, doi:10.1029/2011WR010962.

Harman, C. J., P. A. Troch, and M. Sivapalan (2011), Functional model of water balance variability at the catchment scale: 2. Elasticity of fast and slow runoff components to precipitation change in the continental United States, *Water Resour. Res.*, *47*, W02523, doi:10.1029/2010WR009656.

Kay, A., H. Davies, V. Bell, and R. Jones (2009), Comparison of uncertainty sources for climate change impacts: Flood frequency in England, *Clim. Change*, *92*(1–2), 41–63, doi:10.1007/s10584-008-9471-4.

Klemeš, V. (1986), Operational testing of hydrological simulation models, *Hydrol. Sci. J.*, *31*(1), 13–24, doi:10.1080/02626668609491024.

Koutsoyiannis, D. (2011), Hurst-Kolmogorov dynamics and uncertainty, *J. Am. Water Resour. Assoc.*, *47*(3), 481–495, doi:10.1111/j.1752-1688.2011.00543.x.

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

Lerat, J., V. Andréassian, C. Perrin, J. Vaze, J.-M. Perraud, P. Ribstein, and C. Loumagne (2012), Do internal flow measurements improve the calibration of semidistributed rainfall-runoff models? *Water Resour. Res.*, *48*, W02511, doi:10.1029/2010WR010179.

Mathevet, T. (2005), 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, L'École Natl. du Génie Rural, des Eaux et des For., Paris.

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

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

Morton, F. (1983), Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology, *J. Hydrol.*, *66*, 1–76, doi:10.1016/0022-1694(83)90177-4.

Murphy, C., R. Fealy, R. Charlton, and J. Sweeney (2006), The reliability of an "off-the-shelf" conceptual rainfall runoff model for use in climate impact assessment: Uncertainty quantification using Latin hypercube sampling, *Area*, *38*(1), 65–78, doi:10.1111/j.1475-4762.2006.00656.x.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I. A discussion of principles, *J. Hydrol.*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.

Niel, H., J. E. Paturel, and E. Servat (2003), Study of parameter stability of a lumped hydrologic model in a context of climatic variability, *J. Hydrol.*, *278*(1–4), 213–230, doi:10.1016/S0022-1694(03)00158-6.

Oudin, L., V. Andréassian, T. Mathevet, C. Perrin, and C. Michel (2006a), Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, *Water Resour. Res.*, *42*, W07410, doi:10.1029/2005WR004636.

Oudin, L., C. Perrin, T. Mathevet, V. Andréassian, and C. Michel (2006b), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *J. Hydrol.*, *320*(1–2), 62–83, doi:10.1016/j.jhydrol.2005.07.016.

Oudin, L., V. Andréassian, C. Perrin, C. Michel, and N. Le Moine (2008), Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, *Water Resour. Res.*, *44*, W03413, doi:10.1029/2007WR006240.

Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*(1–4), 275–289, doi:10.1016/S0022-1694(03)00225-7.

Perrin, C., V. Andréassian, C. Rojas-Serna, T. Mathevet, and N. Le Moine (2008), Discrete parameterization of hydrological models: Evaluating the use of parameter sets libraries over 900 catchments, *Water Resour. Res.*, *44*, W08447, doi:10.1029/2007WR006579.

Potter, N. J., and F. H. S. Chiew (2011), An investigation into changes in climate characteristics causing the recent very low runoff in the southern Murray-Darling Basin using rainfall-runoff models, *Water Resour. Res.*, *47*, W00G10, doi:10.1029/2010WR010333.

Prudhomme, C., and H. Davies (2009), Assessing uncertainties in climate change impact analyses on the river flow regimes in the UK. Part 1: Baseline climate, *Clim. Change*, *93*(1), 177–195, doi:10.1007/s10584-008-9464-3.

Refsgaard, J. C., and J. Knudsen (1996), Operational validation and intercomparison of different types of hydrological models, *Water Resour. Res.*, *32*(7), 2189–2202.

Rosero, E., Z.-L. Yang, T. Wagener, L. E. Gulden, S. Yatheendradas, and G.-Y. Niu (2010), 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.

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.

Seibert, J. (2003), Reliability of model predictions outside calibration conditions, *Nord. Hydrol.*, *34*(5), 477–492.

Singh, R., T. Wagener, K. van Werkhoven, M. E. Mann, and R. Crane (2011), A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate—Accounting for changing watershed behavior, *Hydrol. Earth Syst. Sci.*, *15*(11), 3591–3603, doi:10.5194/hess-15-3591-2011.

Steele-Dunne, S., P. Lynch, R. McGrath, T. Semmler, S. Wang, J. Hanafin, and P. Nolan (2008), The impacts of climate change on hydrology in Ireland, *J. Hydrol.*, *356*(1–2), 28–45, doi:10.1016/j.jhydrol.2008.03.025.

Tan, K. S., F. H. S. Chiew, R. B. Grayson, P. J. Scanlon, and L. Siriwardena (2005), Calibration of a daily rainfall-runoff model to estimate high daily flows, in *MODSIM 2005 Int. Congr. on Modelling and Simulation*, edited by A. Zerger and R. M. Argent, pp. 2960–2966, Modell. Simul. Soc. of Australia and New Zealand, Melbourne, Australia.

Teng, J., J. Vaze, F. H. S. Chiew, B. Wang, and J.-M. Perraud (2011), Estimating the relative uncertainties sourced from GCMs and hydrological models in modelling climate change impact on runoff, *J. Hydrometeorol.*, *13*, 122–139, doi:10.1175/JHM-D-11-058.1.

Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.

Vaze, J., and J. Teng (2011), Future climate and runoff projections across New South Wales, Australia: Results and practical applications, *Hydrol. Processes*, *25*(1), 18–35, doi:10.1002/hyp.7812.

Vaze, J., F. H. S. Chiew, J.-M. Perraud, N. R. Viney, D. A. Post, J. Teng, B. Wang, J. Lerat, and M. Goswami (2010a), Rainfall-runoff modelling across southeast Australia: Datasets, models and results, *Aust. J. Water Resour.*, *14*(2), 101–116.

Vaze, J., D. A. Post, F. H. S. Chiew, J. M. Perraud, N. R. Viney, and J. Teng (2010b), Climate nonstationarity—Validity of calibrated rainfall-runoff models for use in climatic changes studies, *J. Hydrol.*, *394*(3–4), 447–457, doi:10.1016/j.jhydrol.2010.09.018.

Vicuna, S., and J. A. Dracup (2007), The evolution of climate change impact studies on hydrology and water resources in California, *Clim. Change*, *82*, 327–350, doi:10.1007/s10584-006-9207-2.

Viney, N. R., J.-M. Perraud, J. Vaze, F. H. S. Chiew, D. A. Post, and A. Yang (2009), The usefulness of bias constraints in model calibration for regionalisation to ungauged catchments, paper presented at *18th World IMACS/MODSIM Congr.*, Modell. Simul. Soc. of Australia and New Zealand, Cairns, Australia.

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrol. Processes*, *17*(2), 455–476, doi:10.1002/hyp.1135.

Wilby, R. L. (2005), Uncertainty in water resource model parameters used for climate change impact assessment, *Hydrol. Processes*, *19*(16), 3201–3219, doi:10.1002/hyp.5819.

Wilby, R. L., and I. Harris (2006), A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK, *Water Resour. Res.*, *42*, W02419, doi:10.1029/2005WR004065.

Xu, C. (1999), Operational testing of a water balance model for predicting climate change impacts, *Agric. For. Meteorol.*, *98-99*, 295–304, doi:10.1016/S0168-1923(99)00106-9.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *J. Hydrol.*, *181*(1–4), 23–48, doi:10.1016/0022-1694(95)02918-4.