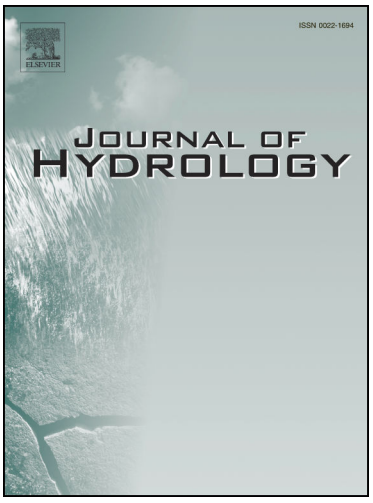
Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds

Thibault Mathevet, Hoshin Gupta, Charles Perrin, Vazken Andréassian, Nicolas Le Moine

Journal Pre-proofs



Research papers

Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds

Thibault Mathevet, Hoshin Gupta, Charles Perrin, Vazken Andréassian, Nicolas Le Moine

PII: S0022-1694(20)30158-X
DOI: https://doi.org/10.1016/j.jhydrol.2020.124698
Reference: HYDROL 124698

To appear in: *Journal of Hydrology*

Received Date: 24 October 2019
Revised Date: 22 January 2020
Accepted Date: 14 February 2020

Please cite this article as: Mathevet, T., Gupta, H., Perrin, C., Andréassian, V., Le Moine, N., Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds, *Journal of Hydrology* (2020), doi: https://doi.org/10.1016/j.jhydrol.2020.124698

Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds

Thibault Mathevet[1], Hoshin Gupta[2], Charles Perrin[3], Vazken Andréassian[3], Nicolas Le Moine[4]*

[1]*EDF-DTG (Électricité de France), 134 rue de l'Etang, 38950 Saint Martin le Vinoux, France*

[2]*Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, Arizona, USA*

[3]*Université Paris-Saclay, INRAE, UR HYCAR, Antony, France*

[4]*SorbonneUniversité, UMR Metis, Paris, France*

** Visiting research scholar at Hydrology and Atmospheric Sciences, University of Arizona, in 2014.*

Corresponding author: Thibault Mathevet (thibault.mathevet@gmail.com)

Abstract

To assess the predictive performance, robustness and generality of watershed-scale hydrological models, we conducted a detailed multi-objective evaluation of two conceptual rainfall-runoff models (the GRX model, based on the GR4J, and the MRX model, based on the MORDOR model), of differing complexity (with respectively, 5 and 11 free parameters in the rainfall-runoff module, and

4 and 11 free parameters in the snow module). These models were compared on a large watershed sample consisting of 2050 watersheds worldwide. Our results, based on the three components of the Kling-Gupta Efficiency metric (KGE), indicate that both models provide (on average) similar levels of performance in evaluation when calibrated with KGE, for water balance (mean bias lower than 2%), time-series variability (mean variability bias lower than 2%) and temporal correlation (mean correlation around 0.83). Further, both models clearly suffer from lack of robustness when simulating water balance, with a significant increase of the proportion of biased simulations over the evaluation periods (absolute bias lower than 2% in calibration and lower than 20% in evaluation for 80% of the watersheds). Simulation performance depend more on the hydro-meteorological conditions of a given period than on the complexity of the model structure. We also show that long-term aggregate statistics (computed on the overall period) can fail to reveal considerable sub-period variability in model performance, thereby providing inaccurate diagnostic assessment of the predictive model performance. Typically the median absolute bias is lower than 8% in evaluation, but the median maximum bias can be as high as 50% within a subperiod, for both models, when calibrated with KGE.

Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds

Thibault Mathevet[1], Hoshin Gupta[2], Charles Perrin[3], Vazken Andréassian[3], Nicolas Le Moine[4]*

[1]EDF-DTG (Électricité de France), 134 rue de l'Etang, 38950 Saint Martin le Vinoux, France

[2]Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, Arizona, USA

[3]Université Paris-Saclay, INRAE, UR HYCAR, Antony, France

[4]SorbonneUniversité, UMR Metis, Paris, France

** Visiting research scholar at Hydrology and Atmospheric Sciences, University of Arizona, in 2014.*

Corresponding author: Thibault Mathevet (thibault.mathevet@gmail.com)

Keywords:

Hydrological modeling, Large-sample hydrology, Calibration, Evaluation, Diagnostics, Kling-Gupta Efficiency

Highlights:

- Large-sample hydrology allows robust statistical analysis of the performance of conceptual rainfall-runoff models

- Models of different complexity provide similar levels of performance during calibration and evaluation
- Models suffer from lack of robustness when simulating long-term water balance

Abstract

To assess the predictive performance, robustness and generality of watershed-scale hydrological models, we conducted a detailed multi-objective evaluation of two conceptual rainfall-runoff models (the GRX model, based on the GR4J, and the MRX model, based on the MORDOR model), of differing complexity (with respectively, 5 and 11 free parameters in the rainfall-runoff module, and 4 and 11 free parameters in the snow module). These models were compared on a large watershed sample consisting of 2050 watersheds worldwide. Our results, based on the three components of the Kling-Gupta Efficiency metric (KGE), indicate that both models provide (on average) similar levels of performance in evaluation when calibrated with KGE, for water balance (mean bias lower than 2%), time-series variability (mean variability bias lower than 2%) and temporal correlation (mean correlation around 0.83). Further, both models clearly suffer from lack of robustness when simulating water balance, with a significant increase of the proportion of biased simulations over the evaluation periods (absolute bias lower than 2% in calibration and lower than 20% in evaluation for 80% of the watersheds). Simulation performance depend more on the hydro-meteorological conditions of a given period than on the complexity of the model structure. We also show that long-term aggregate statistics (computed on the overall period) can fail to

reveal considerable sub-period variability in model performance, thereby providing inaccurate diagnostic assessment of the predictive model performance. Typically the median absolute bias is lower than 8% in evaluation, but the median maximum bias can be as high as 50% within a subperiod, for both models, when calibrated with KGE.

1 Introduction

Rainfall-Runoff (RR) models are widely used for a broad range of research and operational objectives, from hypothesis testing to improving process understanding to streamflow prediction for flood design. Whatever the application, hydrologists and modelers share a particular interest in: i) the efficiency, robustness and realism of model structures (and their consequent simulations); ii) the generality (transposability) of model structures across locations (i.e. ability to be efficient in a variety of hydroclimatic contexts); and iii) methods for parameter identification (*Gupta et al., 2014*). To achieve these objectives, a variety of strategies for model development and specification have been pursued, ranging from detailed site-specific investigations to more general studies. The term robustness is often used to describe some expected model properties in a broad sense. Here, robustness is understood as the capability of a model to hold a certain level of performance in changing conditions, i. e. independently from the input/output information used for calibration. Robustness is usually assessed by comparing the difference of

evaluation metrics under changing conditions (typically from calibration to evaluation periods, but also from dry to wet conditions, etc.).

The investigations discussed in this paper are rooted in the past experience of the authors with RR model intercomparison studies (*Perrin et al., 2001, 2003, 2008; Le Moine et al., 2007; Pushpalatha et al., 2011, 2012; Coron et al., 2012, 2014*), as well as investigations intodiagnostic model identification procedures (*Gupta et al 2008, 2009, 2012; Yilmaz et al., 2008; Martinez and Gupta, 2010, 2011; de Vos et al., 2010; Pokhrel et al., 2012*).

1.1 Why Large-Sample Hydrology?

The field of RR modeling is seeing an increasing number of studies based in the use of data sets containing large samples of watersheds. *Gupta et al. (2014)* point out that the use of such large-sample data sets has four main benefits: i) **improved understanding**, based in rigorous testing and comparison of competing RR model hypothesis and structures, via a uniform and controlled testing scheme; ii) **improved robustness of generalization**, by facilitating a robust statistical analysis of model performance, thereby reducing the undue influence of outliers and case-specific studies; iii) **facilitation of classification, regionalization and model transfer**, by providing diversity of hydrometeorological context; and iv) **support for the estimation of uncertainties**, by establishing a realistic range of RR model predictive

performance and uncertainties under a diverse range of hydrometeorological contexts.

In particular, the use of a large sample of watersheds makes it possible to effectively compare competing RR model structures (either different versions of a single structure, or different structures) and to develop a realistic assessment of their predictive capabilities (*Andréassian et al., 2009*) when applied at other (out-of-sample) locations. More general conclusions can be drawn, and their statistical significance can be tested (*Mathevet et al., 2006*). Further, it enables testing whether conclusions are dependent on the choice of watersheds or not (i.e. conditions such as watershed location, dominant processes, hydroclimatic context, etc.).

Since the *Gupta et al. (2014)* synthesis of large-sample studies, other studies have been published, using samples of hundreds or thousands of watersheds around the world, such as in Europe (*Andréassian et al., 2014; Donnely et al., 2016 ; Rojas-Serna et al., 2016 ; Lane et al., 2019*), in New-Zealand and Australia (*Mc Millan et al., 2016a ; Zheng et al., 2018*), in the USA (*Newman et al., 2015 ; Essou et al., 2016 ; Addor et al., 2017 ; Melsen et al., 2018; Pool et al., 2019; Mizukami et al., 2019*) or in Chile (*Alvarez-Garreton et al., 2018*). These studies reveal that watershed samples set-up long ago are still used and that some new samples are being built, pursuing research objectives ranging from regional studies to general studies on RR modeling. Within the scope of the Panta Rhei research initiative of

the International Association of Hydrological Sciences (*McMillan et al., 2016b*), guidelines for creating and sharing large-sample data sets have beenproposed (*Addor et al., 2019*).

1.2 Insights from previous studies

Gupta et al. (2014) listed 94 large-sample studies conducted over the last 25 years worldwide. Here we mention the results of only a few studies. *Perrin et al. (2001)* compared 20 conceptual RR (CRR) model structures (at a daily time-step), on more than 400 watersheds, with performance evaluation based mainly on the Nash-Sutcliffe Efficiency (NSE, *Nash and Sutcliffe, 1970*). Following a classical split sample calibration and evaluation procedure, they reported that the more complex CRR models suffered from lack of robustness (significant decrease of model performances on independent evaluation periods), and that simpler CRR structures having 4 to 6 free parameters tended to provide the best results, regardless of the performance evaluation metric used.

Similarly, *Mathevet et al. (2006)* evaluated four CRR model structures (at an hourly time-step) on more than 300 watersheds, with different performance evaluation metrics based on NSE. They suggested that the NSE formulation does not allow a comprehensive statistical evaluation of mean CRR model performance on a sample of watersheds. They introduced a modified bounded version of the NSE to mitigate some numerical problems linked to large negative values, and suggested a

framework for testing whether observed differences in CRR model performance are significant or not.

More recently, *Fenicia et al. (2011)* and *Kavetski et al. (2011)* promoted the concept of flexible CRR model structures, in contrast to the fixed CRR model structures tested by *Perrin et al (2001)*, *Mathevet et al. (2006)* and other earlier studies. In the flexible model framework, user-specified hypotheses are introduced using various arrangements of reservoirs, lag functions, junctions and constitutive functions. This facilitates the dialog between the understanding of dominant watershed processes by field hydrologist and the conceptualization by the modeler.

Following from this*, Van Esse et al. (2013)* tested the flexible model concept by evaluating 30 CRR model structures (1 fixed and 29 flexible) on more than 200 watersheds at the hourly time-step, using a variety of different performance metrics. Interestingly, their study showed that although the flexible approach performs better than the fixed approach, it had a higher chance of inconsistent results when calibrated on two different periods. When analyzing results on watersheds where the two approaches produced consistent performance over multiple time periods, their average performance was almost equivalent. This finding highlights how difficult it is to predict model performance on other periods, and that conceptually different structures can yield similar levels of performance.

Last, *Coron et al. (2012)* compared three CRR model structures on more than 200 watersheds at the daily time-step, using various performance metrics. They found that, although the three CRR models produced different levels of performance, they exhibited rather homogenous behavior when calibrated and evaluated on contrasting climatic periods. Again, the CRR models seem to be plagued by a significant lack of robustness, particularly in terms of water balance. These findings were corroborated by those of *Coron et al. (2014)*, using three CRR model structures of different complexity on a sample of 20 watersheds, with a focus on the annual and long-term water balance. Results showed that the three CRR models have strong behavioral similarities in terms of water balance simulation: the mean annual streamflow time-series simulated by the models were more strongly correlated than any single simulation with the observations.

In regards to model performance evaluation, the aforementioned studies typically used evaluation criteria based on the NSE calculated on streamflow (Q) or its various non-linear transformations (sqrt(Q), ln(Q), 1/Q, etc.). The limitations of the NSE criterion have long been recognized and discussed (e.g., see *Schaefli and Gupta 2007*, among others). The Kling-Gupta Efficiency (KGE; *Gupta et al. 2009*) was proposed as an alternative that enables a more consistent assessment of model performance by focusing on a few basic required properties of any model simulation, (i) bias in the

mean, (ii) bias in the variability and (iii) cross-correlation with the observational data (measuring differences in hydrograph shape and timing).

1.3 Scope of the paper

The results of previous model intercomparison studies (including those reported above) indicate that competing CRR model structures, when properly implemented, can often provide similar results on large samples of watersheds, in terms of both hydrograph shape and level of performance, regardless of their competing structural hypotheses and degrees of complexity.

The main objective of this paper is to more deeply investigate this issue by conducting an intercomparison of two CRR models (with fixed structures), following an advanced evaluation protocol applied over a large worldwide sample of watersheds. The goal is to better investigate the general simulation behavior and statistical performance of CRR model structures of differing complexity, by taking advantage of the robust statistical properties achievable via a large-sample hydrological study.

The research questions we investigate are:

- Question 1: How statistically comparable (based on a detailed evaluation procedure) are the simulation performances of two models?

- Question 2: Is the simulation performance of the models essentially identical when provided with the same observational information?
- Question 3: Are differences in model performance dependent on watershed characteristics or on hydrometeorological processes?

To investigate these questions, we implemented an intercomparison framework based on:

a) Two model structures of differing complexity that have previously been demonstrated to have good performance (the GRX model, based on the GR4J model, and the MRX model, based on the MORDOR model),

b) A worldwide sample of more than 2000 watersheds with data at the daily time-step,

c) A multi-objective evaluation process based on the mean squared error decomposition (*Gupta et al., 2009*), and

d) A two-way split-sample testing procedure, as proposed by *Klemeš (1986)*.

The paper continues with Section 2 presenting the experimental design of the intercomparison framework, Section 3 presenting the general results of the study, and Section 4 discussing the results. Our conclusions are detailed in Section 5 along with a discussion of some issues highlighted by this study.

2. Experimental design

2.1 A large worldwide sample of watersheds

The construction of a large sample of watersheds is difficult and time-consuming, and suffers from a number of accessibility issues (*Viglione et al., 2010*). For this study, we compiled a sample of 2,050 watersheds, based mainly on existing datasets used in previous studies. While this sample is not evenly distributed across hydroclimatic regimes and continents, it is (to our knowledge) the largest and most comprehensive sample used in such a study to date.

The main part (80%) of this large sample comes from studies in France (INRAE sample, 1188 watersheds; *Le Moine et al., 2008; Mathevet et al., 2012; Valéry et al., 2014 a & b; Nicolle et al., 2013; Coron et al., 2014*), the USA (MOPEX sample, 320 watersheds; *Duan et al., 2006*), and Australia (CSIRO sample, 356 watersheds; *Vaze et al., 2010; Teng et al., 2012; Coron et al., 2012*). Part of the French sample was compiled by EDF (Electricité de France, French electricity producer) for model development and calibration studies (*Mathevet et al., 2012*), and is mainly based on watersheds where operational hydrometeorological forecasts and hydrological studies are being conducted. The remaining part (20%) of watersheds is from other countries, including Sweden and Switzerland (93 and 31 watersheds respectively; *Valéry et al., 2014 a & b*), the UK (TDMWG sample, 60 watersheds; *Croke et al., 2006*), Laos and Italy (1 and 1 respectively; compiled by EDF).

As shown by **Table 1, Table 2** and **Figure 1**, this worldwide sample covers a variety of hydrometeorological regimes, with a median watershed area of 255 km² and 50% of watershed area between 100 and 1000 km². However, due to difficulties of data-access, the sample is mostly composed of watersheds in France, USA and Australia, with some complementary watershed samples (in Sweden, Switzerland, UK, Laos, Italy) selected to enrich the sample diversity. The French and USA samples tend to cover a wide range of hydroclimatic conditions. A cluster analysis on major hydroclimatic characteristics indicates that six clusters represent the main features of the sample (Table 2), ranging from arid to temperate and cold conditions. Two clusters represent more than the half of the watersheds, temperate with warm summer (T+WS), and arid with desert and steppe (A) (following the Köppen-Geiger climate classification, *Peel et al., 2007*).

In spite of the importance of data quality, particularly when dealing with large samples, there is a practical challenge in performing detailed quality checks on such datasets (*Andréassian et al, 2009; Gupta et al., 2014*). Since this study involves model evaluation under realistic conditions, and because we are mainly investigating general tendencies in regards to model performance/behavior, we will assume that this statistical distribution of watersheds is sufficient to allow relatively robust conclusions about model results and simulations.

The data used are mean daily rainfall, air temperature and streamflow time series. Climatic data are averages at the watershed scale, but the way these averages were computed is variable between the national sub-samples of watersheds (from point observations to spatial reanalyses of precipitations and air temperature). Potential evapotranspiration (PE) was computed using the temperature-based formula proposed by *Oudin et al. (2005)*. Due to the multiple sources of data composing the national sub-samples, data quality is an issue difficult to address. We expect that guidelines for creating and sharing large sample of data sets (Addor et al., 2019) will improve the overall quality of data sets in the future. However, this sample is representative of what is generally available for hydrological studies and modeling.

2.2 2 Two conceptual Rainfall-Runoff models

We selected two, lumped, CRR model structures:

- the *GRX* model (slightly modified from *Le Moine, 2008* and *Pushpalatha et al., 2011*);
- the *MRX* model (modified from *Garçon, 1996* and *Garavaglia et al., 2017*).

These models have already been intensely tested, beginning with the 2004 MOPEX Workshop in Paris (*Andréassian et al., 2006*). A number of studies have shown their structures to be relatively efficient with comparable performance during both simulation and extrapolation (*Mathevet, 2005; Chahinian et al., 2006; Le Moine,*

2008; Velazquez et al., 2010; Garavaglia, 2011; Pushpalatha et al., 2011; Seiller et al., 2012; Coron et al., 2012; Valéry et al., 2014 a & b). Both the *GR* family of models and the *MRX* model have been progressively improved over the years (for *GR* see *Pushpalatha et al., 2012; Coron et al., 2014*; for *MRX* see *Mathevet et al., 2012; Le Lay et al., 2015; Garavaglia et al., 2017*), and are widely used in France for research, engineering, and operational applications, by consultants and national and EDF hydrometeorological forecast centers. Figure 2 illustrates the structures and free parameters of the GRX and MRX hydrological models;. The model equations can be found in *Le Moine, (2008)* and *Garavaglia et al. (2017),* respectively.

2.2.1 *GRX* Rainfall-Runoff model

The *GRX* model used in this study is derived from the *GR4J* model (*Perrin et al., 2003*), developed at a daily time-step, following a multi-objective framework (*Le Moine, 2008; Pushpalatha et al., 2011*). The model was developed in an empirical manner (i.e. trial and error), and has previously been tested on more than a thousand watersheds in France. It has five free-parameters, two stores (one for runoff production and one for routing), a direct component (10% of effective rainfall), a main component (90% of effective rainfall), and a dynamical water gain/loss exchange function (so that exchanges can be bi-directional) applied to the two flow components. Potential evapotranspiration is estimated using the *Oudin* formula

(*Oudin et al., 2005*), and water balance is controlled by the runoff production store and by the water gain/loss exchange function.

2.2.2 *MRX* Rainfall-Runoff model

The *MRX* model used in this study is a modified version of the model used for engineering and operational applications (flood, drought and inflow forecasts, hourly to daily time-steps, deterministic or probabilistic forecasts) at EDF. It was initially developed by *Garçon (1996)* over a few Alpine watersheds, and was then implemented on hundreds of French watersheds. The model has also been applied outside France in a number of intercomparison studies (*Mathevet et al., 2006; Andréassian et al., 2006; Valery et al., 2014b*). The structure of the original model has been updated recently (*Garavaglia et al., 2017),* mainly for operational purposes (this new structure is not used in this study).

The model has eleven free parameters, four stores (one for production, one for production and routing, two for routing), a fast routing component, a slightly delayed routing component, and a slow routing component. Water balance is controlled by two stores and by a correction factor of the potential evapotranspiration function, which is based on air temperature. Although the structure of *MRX* is rather complex compared to *GRX* (as shown by Figure 2), past studies have shown *MRX* to be rather robust, and that fixing some of its parameters to default values results in only minor reduction in performance (*Mathevet, 2005; Garavaglia et al., 2017*).

2.2.3 Snow accumulation and melt model (*SAM*)

Snow processes are significant in approximately 25% of the watersheds of our sample (with more than 10% of total precipitation falling as snow). However, due to limitations in the availability of meta-data (e.g., Digital Elevation Model), only the snowy watersheds from France, Sweden and Switzerland were included (i.e., snowy watersheds from the original MOPEX sample were removed).

To focus attention on differences between the two RR model structures, the same *SAM* component (taken from *MRX*) was used here for both models. It is characterized by: i) a transition range of temperatures for determining the solid fraction of precipitation; ii) the thermal state of the snowpack (controlling snowpack inertia and snowmelt); iii) a degree-day factor (controlling snowmelt); and iv) a snow depth/altitude repartition function. To determine the transition range of temperature and the snow depth/altitude repartition, the hypsometric curve of the watershed is required.

The parameterizations of the *SAM* components were adapted to the levels of complexity of the *GRX* and *MRX* models, with four-parameter and eleven-parameter *SAM* versions respectively. Previous tests (not detailed here, see *Valéry et al., 2014b*) indicate that the use of the simplified *SAM* version has a limited impact on the *GRX* model performance, i.e. that the loss of efficiency compared to the more complex SAM version is not significant enough to support this extra-complexity.

2.3 Evaluation metrics

We used a number of metrics that provide summary assessments of model performance, including the *Kling-Gupta Efficiency* (*KGE*; Eq. 1) and its three components (Eqs. 2-4) based on the decomposition of the *Mean Squared Error* criterion (*Gupta et al., 2009; Gupta and Kling 2011*). The classical *Nash-Sutcliffe Efficiency* (*NSE*; Eq.5) served as a comparative reference with previous studies (albeit a questionable basis for benchmarking – see *Schaefli and Gupta, 2007*):

$$KGE(Q) = 1 - \sqrt{(\beta(Q) - 1)^2 + (\alpha(Q) - 1)^2 + (r(Q) - 1)^2} \quad \text{Eq.1}$$

where, □ is the mean bias, □ is the variability bias and ***r*** is the linear correlation.

$$\beta(Q) = \frac{\widehat{\mu_Q}}{\mu_Q} \quad \text{Eq. 2}$$

$$\alpha(Q) = \frac{\widehat{\sigma_Q}}{\sigma_Q} \quad \text{Eq. 3}$$

$$r(Q) = \frac{\Sigma_{i=1}^{n}(Q_i - \overline{Q}).(\hat{Q}_i - \hat{\overline{Q}})}{\sigma_Q.\widehat{\sigma_Q}} \quad \text{Eq. 4}$$

$$NSE(Q) = 1 - \frac{\frac{1}{n}\Sigma_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\frac{1}{n}\Sigma_{i=1}^{n}(Q_i - \overline{Q})^2} \quad \text{Eq.5}$$

where Q and $\hat{Q}$ are the observed and estimated streamflows respectively, μ_Q and $\widehat{\mu_Q}$ are the corresponding means, σ_Q and $\widehat{\sigma_Q}$ are the corresponding standard deviations, and ***n*** is the number of time steps.

In addition, we used the modified *KGE(Q)* proposed by *Kling et al. (2012),* which avoids the commonly observed correlation between □ and □ , by replacing □□ by the coefficient of variation bias □:

$$\gamma = \frac{\hat{\sigma}_Q}{\sigma_Q} \Big/ \frac{\hat{\bar{Q}}}{\bar{\bar{Q}}} = \alpha / \beta \quad \text{Eq. 6}$$

□, □ and □ assess the mean and variability biases in the long-term water balance and variability components of model simulations,while ***r*** assesses the temporal dynamics of the model. As discussed by *Gupta et al (2009)*, the aggregate measures *NSE(Q)* and *KGE(Q)* mix these components to provide summary performance assessments that can be used in single-criteria model calibration and evaluation. In contrast, the terms of the *KGE(Q)* decomposition (□, □□ ***and r***) provide a comprehensive assessment that is more valuable than the aggregated value. We consider that this decomposition is easy to interpret since it clearly identifies three basic required properties of model simulations.

To provide more information about model simulation properties, different hydrological signatures (*Euser et al., 2013; Westerberg and McMillan, 2015*) can be used as evaluation metrics, particularly for operational applications and engineering studies, such as hydrological forecasts, general and extreme hydrological studies (*Mathevet et al., 2012; Garavaglia et al., 2017*). For the sake of simplicity, the

analyses reported in this text focuses on the streamflow time series, while complementary results related to hydrological regime and high or low flows are presented in the supplementary materials of this study.

2.4 Testing procedure

We implemented a classical split sample model calibration and evaluation procedure (*Klemeš, 1986*). For each watershed, the time period was divided into two independent time periods of equivalent length. Calibration was performed on each period (first half, then second half), followed by evaluation on the other period (second half, then first half). The mean length of the calibration and evaluation periods is 14 years. This means that performance assessment was based on 2050 x 2 = 4100 sets of criterion values computed on both the calibration and evaluation periods. A one-year 'warm-up' period prior to the beginning of each test period was used to minimize state initialization errors. Parameter optimization was carried out using an automatic calibration based on a Genetic Algorithm (*Mathevet, 2005*), independently using each of the two aggregate performance metrics *KGE(Q)* and *NSE(Q)* as objective function.

3. Results

3.1 Boxplots analyses

Figure 3 shows the sample distributions of different performance criteria represented as boxplots (for all 2050 watersheds and for both calibration and evaluation periods) for each of the two models when either *KGE(Q)* or *NSE(Q)* was used as the calibration criterion.

Decomposition terms: When using ***KGE(Q)*** for calibration (***Figure 3a***), the calibration period boxplots of □, □ and □ show that model simulations have almost no long-term bias in either mean (water balance) or standard deviation (variability) of the sample distribution. However, this behavior is not the same during evaluation. While evaluation performance over the sample remains, on average, unbiased in the mean and variance, individual models for specific watersheds do show increases in mean and/or variability bias so that the widths of the boxplots increase (as shown by the range of the 0.25/0.75, 0.1/0.9 or 0.05/0.95 quantiles), indicating lack of robustness. The typical range of bias for □, □ and □ is rather high, being around ±20% for 80% of the watersheds. Clearly, in going from calibration to evaluation period, neither model maintains the quality of long-term water balance and variability, and both models show very similar decrease in performance.

In contrast, the sample distributions for the correlation coefficient ***r*** indicate that both models have very similar and quite high performance (median value around 0.88),

that is robust with almost no difference between calibration and evaluation periods. So, the models seem to have a similar ability to reproduce hydrograph timing and shape. This is not entirely surprising given that the timing and shape are largely determined by the temporal patterns of precipitation and evapotranspiration (i.e. temperature). However, since watershed storage, release and routing also play an important role, this result indicates that such properties are reasonably well represented, at the aggregate level, by both models.

Distributions of KGE and NSE: The results clearly show that the two composite criteria filter out (i.e., do not reveal) some important information. *KGE* distributions in calibration are very similar to $\boldsymbol{r}$ distributions, which makes sense since □□□and□□ are effectively zero (the samples are unbiased) so that $KGE \approx r$ (see Eq. 1 in *Gupta and Kling, 2011*). However, the distributions of *KGE* and *NSE* show a general (on average) decrease of performance during evaluation, due to the (on average) decline in performance in water balance and variability. These results indicate that the major cause of lack of robustness (significant decrease of model performances from calibration to evaluation periods) of RR models may be the difficulty in representing the long-term water balance over different time periods, while the dynamical behavior may actually be rather well represented.

Of course, an alternative interpretation could be that long-term mean and variability are not adequate as informative diagnostics of model performance, and that the linear

correlation coefficient is not a sufficiently powerful metric of temporal difference in shape and timing of the hydrographs. We will come back to this point later in the Discussion.

The comparison of ***Figure 3a*** with ***3b*** illustrates the influence of calibration metric (objective function) on model performance. Consistent with the theoretical analysis of *Gupta et al. (2009)*, the use of *NSE* as calibration metric leads to only a small bias in the mean (water balance), but a significant negative bias in the variability during calibration, due to the fact that variability and correlation interact in an undesirable manner. However, sample variability for both □□□and□□□ increases significantly during evaluation, while the correlation performance is slightly worse.

Since the differences between calibration and evaluation are essentially identical for the two metrics, these results indicate that the choice of calibration metric has significant effects on the characteristics of performance, while having limited impact on the robustness. However, using *KGE* (rather than *NSE*) as calibration metric improves long-term water balance and variability while maintaining relatively good performance in terms of long-term correlation. This result is also shown by *Mizukami et al. (2019)*, on a sample of 492 watersheds in the USA, comparing two hydrological models, calibrated with NSE and KGE metrics.

Figure 3c shows distributions of the differences in performance between the two models. The results are shown in two ways – inter-model differences (*GRX* minus

MRX) for calibration and evaluation periods, and inter-period differences (evaluation minus calibration) for the two models. These results indicate that median inter-model differences in performance are negligible during calibration and fairly small during evaluation. In contrast, inter-period differences are larger, and have much wider distributional ranges. The fact that inter-model differences are small compared to inter-period differences suggests two things:

1) Differences in complexity between the two models do not result in significant differences in model performance (as assessed by the performance metrics used).

2) There is either a) possible changes in time in the data quality and/or characteristics of the watersheds, b) model structural inadequacy so that parameters would need to be different for different periods, or c) inadequate calibration of the models (e.g., due to failure of the metrics to provide meaningful performance assessment).

Here, given no information was available on changes in the data quality or watershed characteristics, a combination of structural inadequacy and calibration inadequacy could be assumed and further investigated.

3.2 Scatterplot analyses

Next, scatterplots of results from the 2050 watersheds are used to look for correlations and interdependencies. For simplicity, the figures show only mean bias (□) and correlation (***r***), while ***Table 3*** lists the scatterplot linear correlations for different combination of models, periods and parameter sets. Results for variability and standard deviation bias (□ and □) are not presented, being similar to those of mean bias. Results concerning *KGE* and *NSE* are similar to those for mean bias and correlation.

During calibration (*Figure 4a* left), there is essentially no correlation in *GRX* versus *MRX* scatterplot of mean bias □ because the two models show very limited mean bias, although *MRX* is unable to achieve a proper water balance (strong negative bias) on several watersheds while *GRX* does quite well. However, on the evaluation periods (*Figure 4a* right), mean bias ***β*** performance is highly correlated (***r*** = 0.85) with bias values well organized around the 1:1 line. This supports the hypothesis that bias in water balance is determined largely by data period, regardless of calibration period performance or model structure.

This finding is further supported by the analysis of evaluation biases between periods for the two models (***Figure 4b***). The evaluation period biases are negatively correlated (***r = -0.73*** for *GRX* and ***r = -0.63*** for *MRX*), indicating that water balance

over-estimation during one period is associated with water balance under-estimation during the other period, regardless of model.

In terms of linear correlation ***r*** statistic, results (*Figure 4*c) show high correlation between the two models during both calibration and evaluation (***r = 0.88***) and at a level that is essentially equivalent to the mean sample correlation between the observations and the simulations (i.e., the mean value of the ***r*** statistic). However, *MRX* displays a small tendency towards better performance (in terms of ***r***), which is likely due to its more complex structure. This result supports the idea that the metrics used here are not sufficient to properly assess model behavior and thereby to distinguish between capabilities of the models.

The difference in performance during the calibration and evaluation periods provides an assessment of model robustness (***Figure 5a***). Clearly, both models suffer from strong lack of robustness with respect to water balance, with lack of bias during calibration (vertical line around 0.0) resulting in a wide range of bias during evaluation with virtually no relationship between them (except for *MRX* on the same limited number of watersheds mentioned earlier).

Conversely, there is high correlation and relatively good robustness of model performance (i.e., limited differences between calibration and evaluation) in terms of correlation statistic r (***Figure 5b***).

Figure 5c shows the results plotted in a different way with both axes showing the correlation statistic r computed for a particular evaluation period, but with each axis corresponding to the parameter set obtained by calibration on a different period. The very high (>0.9) correlation indicates lack of dependence of ***r*** statistic performance on period selected for calibration. Clearly, both models (when calibrated) are able to reproduce the patterns of temporal dynamics relatively well (as measured in an aggregate sense by ***r***) and this ability is not dependent on the period chosen for calibration.

3.3 Consistency of model performance

The results above show considerable similarity in values of the performance metrics for the *GRX* and *MRX* models. Since the models have different structures and levels of complexity, we next assess their consistency in performance, i.e. their relative ability to produce better simulations during calibration and evaluation. ***Table 4*** shows the proportion of watersheds where *GRX* is consistently better than *MRX*, and vice versa. In regards to water balance □ and variability □ and □, each model shows consistent superiority on both 'calibration' and 'evaluation' periods in approximately 23% to 28% of the cases. It means that in ~50% of the cases, the '*best'* model is different from one period to the other. In regards to ***r***, *KGE* and *NSE*, the *MRX* model is superior in 43 to 48% of the cases compared to *GRX* (22 to 27% of the cases). Meanwhile, the '*best'* model differs in only 30% of the cases. These

results indicate that *MRX* is able to take a rather slight advantage of its more complex structure to provide (on average) some improvement over *GRX* (with the caveats mentioned earlier regarding certain outlier cases, and the diagnostic ability of the metrics used here).

3.4 Overall summary of major results

From the large-sample evaluation of model performance presented above, we can draw the following conclusions:

1) Both models suffer from a strong lack of robustness in the simulation of water balance and streamflow variability. The water balance bias varies on the range ±10% for 50% of the watersheds, and on the range ±20% for 80% of the watersheds.

2) The performance of both models (assessed on independent evaluation periods) is highly correlated (***r*** ranging from 0.75 to 0.92). This means that model performance correlation (between simulations provided by the two models) is at the same level as the correlation between each of the model simulations and the observations, suggesting that there is no significant difference in overall abilities of the two models across the range of watersheds used for testing.

Further, it seems that differences in hydroclimatic conditions between calibration to evaluation periods play a more important role on the differences in performance from calibration to evaluation than differences in model structures do.

4. Discussion

4.1 To what extent does model performance depend on hydro-meteorological characteristics of the watershed?

To assess the impact of hydro-meteorological characteristics of a watershed on model performance, we analysed the distribution of ***b*** and ***r*** metric performance on the evaluation period for the six region types listed in ***Table 2***. It is clear from ***Figure 6*** that arid watersheds are significantly different from temperate to cold watersheds, with mean bias varying approximately on the range ±35% for 80% of the arid watersheds, while varying on the smaller range ±10-20% for 80% of the temperate to cold watersheds. For both models, the ability to simulate long-term water balance progressively improves as watershed humidity and runoff yield increase.

Arid watersheds also show significantly poorer linear correlation ***r*** performance for both models, with considerable variability in the results (lack of inter-watershed robustness). Again, performance improves as watershed humidity and runoff yield increase (for temperate to cold watersheds). In the latter hydroclimatic conditions, the more complex *MRX* model shows a small tendency towards better performance than *GRX*.

This trade-off between model performance and hydro-meteorological characteristics of the watershed is also shown within the USA by *Newman et al. (2015).* That study is particularly interesting because the hydro-meteorological variability is rather strong (and comparable to our study) across of the USA. Analyzing spatial distribution of performance of the SAC-SMA model calibrated on 671 watersheds, *Newman et al. (2015)* stated that the main factors influencing variation of model performance were aridity and precipitation intermittency. Previous studies have also confirmed that such models perform better in wetter watersheds (*Liden and Harlin, 2000 ; McMillan et al., 2016a; Lane et al., 2019*).

Following *Coron et al. (2012)*, we looked for a link between the lack of evaluation period water balance robustness, and the variability of hydroclimatic characteristics between calibration and evaluation periods. ***Figures 7 (a, b, c)*** show scatterplots between water balance bias and the differences or ratios between long-term mean air temperature (T), precipitation (R) and runoff (Q) during evaluation and calibration for the *MRX* model (being similar, results for *GRX* are not shown). Water balance bias shows no relationship with air temperature (or PE; results not shown) and precipitation for this sample of 2050 watersheds, whereas a slight relationship is seen with 'runoff ratio variability between calibration and evaluation' ***(Figure 7c)***. This absence of clear relationship reinforces the findings of *Coron (2013)* who

showed that whereas such a relationship is possible, it can be extremely variable between watersheds.

We further investigated the slight relationship between water balance bias ***b*** and the runoff ratio between calibration and evaluation. Arid ***(Figure 7d)*** and temperate watersheds with hot summers ***(Figure 7e)*** exhibit some degree of relationship (correlations from 0.35 to 0.46). The lack of robustness is likely due to complex tradeoffs between (i) climate non-stationarity that causes runoff variability (runoff ratios between calibration and evaluation ranging from 0.2 to 0.25), (ii) model structural inadequacies in the context of arid and non-perennial watersheds, and (iii) over-calibration of the model parameters to water balance (see *Coron et al., 2012*).

4.2 How adequate are long-term aggregate model performance metrics as diagnostic tools?

Here, we present results suggesting that long-term aggregate model performance metrics such as *KGE* and *NSE* or their component terms (□□□□□□□, ***r***) might not be adequate for diagnostic assessment of model performance or for intercomparison of RR model structures. In particular, we examine the lack of model performance robustness in regards to simulating water balance, by examining the properties (***Figures 8 and 9***) of the time series of cumulative residuals during both calibration and evaluation.

For the La Borrèze at Lachapelle-Auzac watershed (southern France), ***Figure 8a*** tracks the temporal trajectory of the long-term mean volume bias of the *GRX* model, and thereby reveals the maximum value of volume bias achieved during the period (i.e., the absolute value of the difference between the maximal and the minimal cumulative residual). Clearly, even if the overall mean volume bias for the calibration period is close to zero, the value can be different from zero at various intermediate times during the calibration period (as shown on the figure by the period maximum volume bias). Meanwhile the mean volume bias trajectory for the evaluation period can be quite different.

We analyzed the distribution of '*inter-period'* (calibration to evaluation) and '*inter-model'* (*GRX* to *MRX*) correlations of the cumulative residuals time-series obtained for all of the watersheds (***Figure 8b***). The '*inter-period'* cumulative residual time-series are more strongly correlated (median $r \sim 0.9$) than the '*inter-model'* cumulative residual time-series (median $r \sim 0.7$). This indicates that the properties and temporal variations of the mean volume bias trajectory depend more on the model structure than on the period used for model calibration. Further, the results for *MRX* are somewhat more robust than for *GRX*, perhaps due to the fact that the latter uses a more simple approach to determining water balance (i.e., the *GRX* structure requires calibration of only a runoff production store capacity and a water gain/loss exchange function).

As shown in ***Figure 8c*****,** while the period mean volume bias (or ***Bp***) can be close to zero for the calibration and evaluation periods, the period maximum volume bias (***Bx***) can be quite large. So, even if the overall calibration period bias can be quite small, the bias on shorter intervals of the calibration periods can reach as high as ~30%. Further, whereas the sample average overall bias is larger (~10%) during evaluation than during calibration, the sample average bias on shorter intervals can reach as high as ~50% and be highly variable (ranging from 10% to 175% for 80% of the watersheds).

Figure 9 shows plots of the mean and variability of the correlation metric ***r*** when computed over *n* independent 90-day sub-periods (where *n* is the number of independent 90-day sub-periods in the full period). These distributions are also compared to the distributions obtained using the full period. *MRX* tends to be slightly better than *GRX*, with slightly higher mean and slightly lower variability (indicating somewhat better temporal robustness) on both the calibration and evaluation periods (***Figure 9*****a**). Further, the 90-day sub-periods show slightly worse mean correlation performance (on average) than on the full periods. Similarly, the average difference between *MRX* and *GRX* model performance increases when evaluated on shorter sub-periods (***Figure 9b***).

These results support our earlier hypothesis that the use of '*entire-period*' aggregate performance metrics may not provide sufficient discrimination to properly assess

model behavior and to thereby distinguish between the capabilities of different models. It therefore seems necessary to investigate and develop metrics that better reflect both sub-period and overall period model performance, so as to reduce the tendency to over-estimate the true predictive performance of a model (*Andréassian et al., 2009*). Recent research on taking into account changing conditions in model calibration and evaluation protocols (*Gharari et al., 2013; Thirel et al., 2015; Fowler et al., 2018*) or attempts to reduce calibration sensitivity to a limited part of the observations (*Brigode et al., 2015; Wang et al., 2012*) might reduce sub-period sensitivity of model performances.

4.3 How well do the models perform on other hydrological signatures (regime, high flows, low flows)?

GRX and MRX model performance was evaluated on some hydrological signatures (hydrological regime, high flows or low flows). For sake of brevity, the detailed results are presented in the supplementary material. From that large-sample evaluation of model performance on hydrological signatures, we can draw the following conclusions:

1) Both models are able to reach a high level of performance in terms of the regime and flood hydrological signatures. On these two hydrological signatures, there is a similar lack of robustness of model performance,

certainly due to the lack of robustness in the simulation of the water balance and streamflow variability.

2) Both models have a low level of performances in terms of the drought hydrological signature.
3) The difference of behavior between regime and flood signatures, and drought signature is not surprising, since the drought signature appears more demanding than the two other signatures, when calibrating models using an objective function based on the streamflow sample only (Pushpalatha et al, 2012).
4) Differences in performances of GRX and MRX seem to be slightly more pronounced on hydrological signatures (regime, flood and drought) than on the streamflow time series. This result could be interpreted as a better ability of MRX to take advantage of its complexity, compared to GRX, when evaluated on independent evaluation metrics that have not been used for model calibration. In other words, comparing models structures using the same evaluation metrics to those used during parameter optimization may reduce our ability to discriminate between model structures.

5. Conclusions and perspectives

This paper has reported a worldwide large-sample intercomparison (based on 2050 watersheds from eight countries) of two CRR model structures (*GRX* and *MRX*), using a multi-objective evaluation process and two-way split-sample testing. The watershed sample represents a diversity of hydro-meteorological and measurement contexts thereby lending confidence and statistical robustness to the analyses and inferences drawn therefrom. Overall, our results support the following answers to the research questions posed:

Question 1: How statistically comparable (based on a detailed evaluation procedure) are the simulation performances of two models?

- The *GRX* and *MRX* models, although of differing structural complexity provide similar levels of long-term (aggregate) performance during calibration and evaluation (as assessed by the various metrics computed).
- Both models suffer from a lack of robustness when simulating water balance and streamflow variability, although simulation of streamflow timing and rate of change is quite good (as indicated by the long-term linear correlation between observed and simulated time series).
- The use of *KGE* as an objective function tends to reduce long-term process model bias (on average), which is not the case with *NSE*.

Question 2: Is the simulation performance of the models essentially identical when provided with the same observational information?

- The more complex *MRX* model tends to provide slightly better and more robust reproduction of short-term processes than the simpler *GRX* model (as indicated by the distributions of short-term linear correlations between observed and simulated time-series).
- Model evaluations using hydrological signatures that are different from the optimization objective function show that differences in model performances are more actually significant. The more complex *MRX* model tends to provide better and more robust performance on hydrological regime and high flows, while both models perform poorly on low flows.

Question 3: Are differences in model performance dependent on watershed characteristics or on hydrometeorological processes?

- Model performance variations from one period to another appear to be mainly due to temporal variations in the hydroclimatic conditions of the period.
- However, the properties of the trajectory of mean volume bias seems to depend more on model structure than on the hydroclimatic characteristics of the period used for model calibration.

In regard to the relatively poor model performance on arid/dry watersheds, and to the dependence of performance on hydroclimatic conditions of a given period, these results highlight the model structural inadequacy of the two tested models (and perhaps more generally of current CRR models) on arid and non-perennial watersheds, and the tendency to over-calibrate model parameters to specific climatic periods. Our results further illustrate the difficulty in properly simulating the dynamics of the watershed behavior over both the long and the short terms.

Further, our results clearly show that sub-period variability in model performance can be quite high (especially for water balance), and that aggregate long-term (full period) statistics may tend to over-estimate true predictive performance of a hydrologic model. While these results should be viewed as preliminary, they suggest that there may be value in computing and examining distributions of the various model performance metrics over sub-period samples, instead of relying upon a single period-average deterministic value. This could greatly improve model diagnosis by helping to reveal situations involving model structural inadequacy, changes in hydro-meteorological processes and/or problems in data.

6. Acknowledgments

Authors would like to acknowledge: Météo France, SCHAPI-Banque Hydro, EDF, Laurent Coron, Nicolas Le Moine and Audrey Valery for the French data sets, Jai Vaze and Francis Chiew for the Australian data sets (CSIRO), John Schaake and

Qingyun Duan for the American (MOPEX) data set, Audrey Valery for the Swiss (Météo Suisse and OFEV) and Swedish (SMHI) data sets, Berit Arheimer for the Swedish (SMHI) data sets, and Barry Croke and Ian Littlewood for the English data sets (TDMWG). We thank the two anonymous reviewers and the Associate Editor Roger Moussa for their constructive feedbacks on this work, which helped improving the quality of the article.

Hoshin Gupta received partial support from the Australian Research Council through the Centre of Excellence for Climate System Science (grant number CE110001028), and from the EU-funded project 'Sustainable Water Action (SWAN): Building Research Links Between EU and US' (INCO-20011-7.6 grant number 294947).

Thibault Mathevet received appreciated support from Annick Gingras-Genois and Federico Garavaglia from EDF-DTG to visit the Department of Hydrology and Water Resources of the University of Arizona. Shervan Gharari, Mari A. Sans Fuentes, Rositsa Yaneva, Natalia Limones, and Exo Roast Coffee baristas are warmly acknowledged for helping to make the time in Tucson both fascinating and amazing.

7. References

Addor, N., A. J., Newman, N., Mizukami, and M. P., Clark, 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies,

Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017.

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., Mendoza, P. A., 2019. Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrological Sciences Journal, Special Issue: Hydrological Data: Opportunities And Barriers, DOI: 10.1080/02626667.2019.1683182

Alvarez-Garreton, C., P. A., Mendoza, J. P., Boisier, N., Addor, M., Galleguillos, M., Zambrano-Bigiarini, A., Lara, C., Puelma, G., Cortes, R., Garreaud, J., McPhee, and A., Ayala, 2018. The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrol. Earth Syst. Sci., 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018.

Andreassian, V., S. Bergstrom, N. Chahinian, Q. Duan, Y.M. Gusev, I. Littlewood, T. Mathevet, C. Michel, A. Montanari, G. Moretti, R. Moussa, O.N. Nasonova, K.M. O'Connor, E. Paquet, C. Perrin, A. Rousseau, J. Schaake, T. Wagener, and Z. Xie, 2006. Catalogue of the models used in MOPEX 2004/2005, IAHS Publication n°307, 41-93.

Andreassian, 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, 1757-1764, doi:10.5194/hess-13-1757-2009.

Andreassian, V., F. Bourgin, L. Oudin, T. Mathevet, C. Perrin, J. Lerat, L. Coron, and L. Berthet, 2014. Seeking genericity in the selection of parameter sets: Impact on hydrological model efficiency, Water Resour. Res., 50, 8356–8366, doi:10.1002/2013WR014761.

Brigode, P., E. Paquet, P. Bernardara, J. Gailhard, F. Garavaglia, P. Ribstein, F. Bourgin, C. Perrin and V. Andréassian, 2015. Dependence of model-based extreme flood estimation on the calibration period: case study of the Kamp River (Austria), Hydrological Sciences Journal, 60:7-8, 1424-1437, DOI: 10.1080/02626667.2015.1006632

Chahinian, N., V. Andréassian, Q. Duan, V. Fortin, H. V. Gupta, T. Hogue, T. Mathevet, A. Montanari, G. Moretti, R. Moussa, C. Perrin, J. Schaake, T. Wagener, and Z. Xie, 2006. Compilation of the MOPEX 2004 results, IAHS Publication n°307, 313-338.

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

Coron, L., 2013. Les modèles hydrologiques conceptuels sont-ils robustes face à un climat en évolution ? Diagnostic sur un échantillon de bassins versants français et australiens. Thèse de doctorat. AgroParisTech. 234 p., Paris, France.

Coron, L., V. Andreassian, C. Perrin, M. Bourqui, and F. Hendrickx, 2014. On the lack of robustness of hydrologic models regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments, Hydrol. Earth Syst. Sci., 1818, 727-746. doi:10.5194/hess-18-727-2014

Croke, B., I. Littlewood and D. Post, 2006. Rainfall - streamflow - air temperature datasets (and catchment information) available internationally to assist with PUB Decade top-down modelling, In: Voinov, A., Jakeman, A., Rizzoli, A. (eds). Proceedings of the iEMSs Third Biennial Meeting: "Summit on Environmental Modelling and Software". International Environmental Modelling and Software Society, Burlington, USA, July 2006

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

Donnelly, C., J. C.M. Andersson and B. Arheimer (2016) Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across

Europe, Hydrological Sciences Journal, 61:2, 255-273, DOI: 10.1080/02626667.2015.1027710

Duan, Q., J. Schaake, V. Andréassian, S. Franks, G. Goteti, H.V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O.N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener and E.F. Wood, 2006. Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, J. Hydrol., 320 (1–2), 3-17, doi:10.1016/j.jhydrol.2005.07.031.

Essou, G.R.C., R. Arsenault, F.P. Brissette, 2016. Comparison of climate datasets for lumped hydrological modeling over the continental United States J. Hydrol., 537, pp. 334-345, 10.1016/j.jhydrol.2016.03.063

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

Fenicia, F., D. Kavetski, and H. H. G. Savenije, 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resources Research, 47(11), W11510, DOI 10.1029/2010wr010174.

Fowler, K., G., Coxon, J., Freer, M., Peel, T., Wagener, A., Western, R., Woods, and L., Zang, 2018. Simulating runoff under changing climatic conditions: A framework for model improvement. Water Resources Research, 54, 9812–9832. https://doi.org/10.1029/2018WR023989

Garavaglia, F., 2011. Méthode SCHADEX de prédétermination des crues extrêmes. Méthodologie, applications, études de sensibilité. Thèse de Doctorat, Université de Grenoble.

Garavaglia, F., Le Lay, M., Gottardi, F., Garçon, R., Gailhard, J., Paquet, E., and Mathevet, T., 2017. Impact of model structure on flow simulation and hydrological realism: from a lumped to a semi-distributed approach, Hydrol. Earth Syst. Sci., 21, 3937-3952, https://doi.org/10.5194/hess-21-3937-2017, 2017.

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

Gharari, S., M., Hrachowitz, F., Fenicia, and H. H. G., Savenije, 2013. An approach to identify time consistent model parameters: Sub-period calibration. Hydrology and Earth System Sciences, 17(1), 149–161. https://doi.org/10.5194/hess-17-149-2013

Gupta, H. V., Wagener, T. and Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Process., 22: 3802–3813. doi: 10.1002/hyp.6989

Gupta, H. V., H. Kling, K. K. Yilmaz and G. F. Martinez-Baquero, 2009. Decomposition of the Mean Squared Error & NSE Performance Criteria: Implications for Improving Hydrological Modelling, J. Hydrol., Vol. 377, pp. 80-91, doi: 10.1016/j.jhydrol.2009.08.003.

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.

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz and M. Ye, 2012. Towards a Comprehensive Assessment of Model Structural Adequacy, Opinion Paper, 48(8), 1-16, W08301; doi:10.1029/2011WR011044

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V., 2014. Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18, 463-477, doi:10.5194/hess-18-463-2014.

Kavetski, D., and F. Fenicia, 2011. Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resources Research, 47(11), W11511, DOI 10.1029/2011wr010748.

Klemeš, V., 1986. Operational testing of hydrological simulation models, Hydrol. Sci. J., 31 (1), pp. 13-24, 10.1080/02626668609491024

Kling, H., M. Fuchs and M. Paulin, 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424–425, 264-277,doi:10.1016/j.jhydrol.2012.01.011.

Lane, R. A., G., Coxon, J. E., Freer, T., Wagener, P. J., Johnes, J. P., Bloomfield, S., Greene, C. J. A., Macleod, and S. M., Reaney, 2019. Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain, Hydrol. Earth Syst. Sci., 23, 4011–4032, https://doi.org/10.5194/hess-23-4011-2019, 2019.

Le Lay, M., R. Garçon, J. Gailhard and F. Garavaglia, 2015. Assessment of the water balance over France using regionalized Turc-Pike formula for operational hydrology. 2015 AGU Fall Meeting, San Francisco, USA.

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 Resources Research 43(6), W06428, doi: 10.1029/2006WR005608.

Le Moine, N., 2008. Le bassin versant de surface vu par le souterrain : une voie d'amélioration des performances et du réalisme des modèles pluie-débit ?

Thèse de Doctorat, Université Pierre et Marie Curie (Paris), Cemagref (Antony), 324 pp.

Lidén, R. and J. Harlin, 2000. Analysis of conceptual rainfall–runoff modelling performance in different climates, J. Hydrol., 238 (3–4), pp. 231-247, 10.1016/S0022-1694(00)00330-9

Nash, J. and J. Sutcliffe, 1970. 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.

McMillan, H.K., D.J. Booke and C. Cattoën, 2016a. Validation of a national hydrological model, J. Hydrol., 541 (Part B), pp. 800-815, 10.1016/j.jhydrol.2016.07.043

McMillan, H., A. Montanari, C. Cudennec, H. Savenije, H. Kreibich, T. Krueger, J. Liu, A. Mejia, A. Van Loon, H. Aksoy, G. Di Baldassarre, Y. Huang, D. Mazvimavi, M. Rogger, B. Sivakumar, T. Bibikova, A. Castellarin, Y. Chen, D. Finger, A. Gelfan, D. M. Hannah, A. Y. Hoekstra, H. Li, S. Maskey, T. Mathevet, A. Mijic, A. Pedrozo Acuña, M. J. Polo, V. Rosales, P. Smith, A. Viglione, V. Srinivasan, E. Toth, R. van Nooyen and J. Xia, 2016b. Panta Rhei 2013–2015: global perspectives on hydrology, society and change, Hydrological Sciences Journal, 61:7, 1174-1191, DOI: 10.1080/02626667.2016.1159308

Martinez, G. F., and H. V. Gupta, 2010. Toward Improved Identification of Hydrological models: A Diagnostic Evaluation of the "abcd" Monthly Water Balance Model for the Conterminous United States, Water Resour. Res., 46, W08507, doi:10.1029/2009WR008294.

Martinez, G. F., and H. V. Gupta, 2011. Hydrologic Consistency as a Basis for Assessing Complexity of Water Balance Models for the Continental United States, Water Resour. Res., doi:10.1029/2011WR011229

Mathevet, T., 2005. Quels modèles pluie-débit globaux pour le pas de temps horaire ? Développement empirique et comparaison de modèles sur un large échantillon de bassins versants. Thèse de Doctorat, ENGREF (Paris), Cemagref (Antony), France, 463 pp.

Mathevet, T., C., Michel, V. Andréassian, and C. Perrin, 2006. A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins, IAHS Red Books Series n°307, pp. 211-219.

Mathevet, T., F., Garavaglia, J., Gailhard, R., Garçon and E., Paquet, 2012. Improving model calibration and selection via a better use of evaluation metrics and streamflow sub-samples for operational applications (Poster), IAHS 90th anniversary, PUB symposium, Delft, October 23-25, 2012.

Melsen, L. A., N., Addor, N., Mizukami, A. J., Newman, P. J. J. F., Torfs, M. P., Clark, R., Uijlenhoet, and A. J., Teuling, 2018. Mapping (dis)agreement in

hydrologic projections, Hydrol. Earth Syst. Sci., 22, 1775–1791, https://doi.org/10.5194/hess-22-1775-2018.

Mizukami, N., O., Rakovec, A. J., Newman, M. P., Clark, A. W., Wood, H. V., Gupta, and R., Kumar, 2019. On the choice of calibration metrics for "high-flow" estimation using hydrologic models, Hydrol. Earth Syst. Sci., 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019.

Newman, A. J., M. P., Clark, K., Sampson, A., Wood, L. E., Hay, A., Bock, R. J., Viger, D., Blodgett, L., Brekke, J. R., Arnold, T., Hopson, and Q. Duan, 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19, 209–223, https://doi.org/10.5194/hess-19-209-2015.

Nicolle P., V., Andréassian and E., Sauquet, 2013. Blending neighbor-based and climate-based information to obtain robust low-flow estimates from short time series. Water Resources Research, vol. 49, n° 12, p. 8017-8025, doi 10.1002/2012WR012940

Oudin, L., F., Hervieu, C., Michel, C., Perrin, V., Andréassian, F., Anctil, and C., Loumagne, 2005. 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.

Paquet, E., Garavaglia, F., Garçon, R., and Gailhard, J., 2013. The SCHADEX method: A semi-continuous rainfall–runoff simulation for extreme flood estimation, J. Hydrol., 495, 23–37, 2013.Peel, M. C., B. L., Finlayson, and T. A., McMahon, 2007. Updated world map of the Köppen-Geiger climate classification, Hydrol. Earth Syst. Sci., 11, 1633-1644, doi:10.5194/hess-11-1633-2007.

Perrin, C., C. Michel, and V. Andréassian,2001, Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, J. Hydrol., 242 (3-4) : 275-301,doi:10.1016/S0022-1694(00)00393-0.

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

Perrin, C., V. Andréassian, 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.

Pokhrel P., K. K. Yilmaz and H. V. Gupta, 2012. Multiple-Criteria Calibration of a Distributed Watershed Model using Spatial Regularization and Response Signatures, J. Hydrol., 418-419, 49-60, Special Issue on DMIP-2, doi:10.1016/ j.jhydrol.2008.12.004

Pool, S., Viviroli, D., and Seibert, J., 2019. Value of a limited number of discharge observations for improving regionalization: A large-sample study across the United States. Water Resources Research, 55, 363–377. https://doi.org/10.1029/2018WR023855

Pushpalatha, R., C. Perrin, N. Le Moine, T. Mathevet, and V. Andréassian, 2011. A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, J. Hydrol., 411(1-2), 66-76, doi:10.1016/j.jhydrol.2011.09.034.

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

Rojas-Serna, C., L. Lebecherel, C. Perrin, V. Andreassian, and L. Oudin, 2016. How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments, Water Resour. Res., 52, 4765–4784, doi:10.1002/ 2015WR018549.

Schaefli, B. and H. V. Gupta, 2007. Do Nash values have value? Hydrol. Process., 21, 2075–2080. doi:10.1002/hyp.6825

Seiller, G., F. Anctil, and C. Perrin, 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrol. Earth Syst. Sci., 16(4), 1171-1189, doi:10.5194/hess-16-1171-2012.

Teng, J., F. H. S., Chiew, J., Vaze, S., Marvanek, and D. G. C., Kirono, 2012. Estimation of climate change impact on mean annual runoff across continental Australia using Budyko and Fu equations and hydrological models, J. Hydrometeorol., 13(3), 1094-1106, doi:10.1175/jhm-d-11-097.1.

Thirel, G., V. Andréassian, C. Perrin, J.-N. Audouy, L. Berthet, P. Edwards, N. Folton, C. Furusho, A. Kuentz, J. Lerat, G. Lindström, E. Martin, T. Mathevet, R. Merz, J. Parajka, D. Ruelland & J. Vaze, 2015. Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments, Hydrological Sciences Journal, 60:7-8, 1184-1199, DOI:10.1080/02626667.2014.967248

Valéry, A., V. Andréassian, and C. Perrin, 2014a. 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 1 – Comparison of six snow accounting routines on 380 catchments, J. Hydrol., 517(0), 1166-1175, doi: 10.1016/j.jhydrol.2014.04.059.

Valéry, A., V. Andréassian, and C. Perrin (2014b), 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, J. Hydrol., 517(0), 1176-1187, doi: 10.1016/j.jhydrol.2014.04.058.

van Esse, W. R., C. Perrin, M. J. Booij, D. C. M., Augustijn, F., Fenicia, D., Kavetski, and F. Lobligeois, 2013. The influence of conceptual model structure on model performance: a comparative study for 237 French catchments, Hydrol. Earth Syst. Sci., 17, 4227-4239, doi:10.5194/hess-17-4227-2013.

Vaze, J., F. H. S. Chiew, JM. Perraud, N. Viney, D. A. Post, J. Teng, B. Wang, J. Lerat, M. Goswami, 2010. Rainfall-runoff modelling across southeast Australia: datasets, models and results, Australian Journal of Water Resources, 14, 2, 101-116.

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

Viglione, A., M. Borga, P. Balabanis, and G. Blöschl, 2010. Barriers to the exchange of hydrometeorological data in Europe: Results from a survey and

implications for data policy, J. Hydrol., 394 (1-2), 63–77, doi:10.1016/j.jhydrol.2010.03.023.

Wang, Q. J., D. L. Shrestha, D. E. Robertson, and P. Pokhrel, 2012. A log-sinh transformation for data normalization and variance stabilization, Water Resour. Res., 48, W05514, doi:10.1029/2011WR010973.

Westerberg, I. K. and H. K., McMillan, 2015. Uncertainty in hydrological signatures, Hydrol. Earth Syst. Sci., 19, 3951–3968, https://doi.org/10.5194/hess-19-3951-2015.

Yilmaz K. K., H. V. Gupta and T. Wagener, 2008. A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, doi: 10.1029/2007WR006716.

Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V., and Zhang, T., 2018. On lack of robustness in hydrological model development due to absence of guidelines for selecting calibration and evaluation data: Demonstration for data-driven models. Water Resources Research, 54, 1013–1030. https://doi.org/10.1002/2017WR021470

7. Tables

	Distribution percentiles				
	0.05	0.25	0.50	0.75	0.95
Catchment area (km^2)	31	102	255	734	2785
Mean annual total precipitation (P) [mm/yr]	626	800	950	1175	1627
Part of precipitation falling as snow [-]	0	0	0.03	0.09	0.29
Mean annual air temperature [°C]	4.9	9.2	10.5	12.1	16.2
Mean annual potential evapotranspiration (PE) [mm/yr]	471	633	686	773	998
Mean annual runoff (Q) [mm/yr]	53	207	344	541	1111
Aridity index (P/PE) [-]	0.80	1.12	1.35	1.74	2.82
Runoff coefficient (Q/P) [-]	0.08	0.24	0.36	0.48	0.79
Available time series length [yr]	11	18	33	36	55

Table 1: Distributions of characteristics of the 2050 watersheds

Cluster	Mean hydroclimatic characteristics [mm/yr]	Main locations	Number of watersheds
1: A Arid with desert & steppe	P ~ 750 PE ~ 750 Q ~ 100	Central USA, Australia	586
2: T+HS Temperate with hot summer	P ~ 1200 PE ~ 1000 Q ~ 450	SE USA, Australia	155
3: T+WS Temperate with warm summer	P ~ 1000 PE ~ 700 Q ~ 400	NE USA, France, UK	785
4: C+HS Cold with hot summer	P ~ 1300 PE ~ 600 Q ~ 700	NE USA, France	305
5: C+CS Cold with cold summer	P ~ 900 PE ~ 300 Q ~ 650	Sweden	94
6: T-DS+WS Temperate without dry season and warm summer	P ~ 1600 PE ~ 550 Q ~ 1200	France, UK, Switzerland	125

Table 2: Hydroclimatic characteristics of the six watershed clusters (P = rainfall; PE = Potential evaotranspiration; Q = Streamflow). The names of the clusters come from the most representative climate class from the Köppen-Geiger climate classification (*Peel et al., 2007*)

	Linear Correlations (KGE(Q) as objective function)					
	GRX vs. ***MRX*** - Calibration	***GRX*** vs. ***MRX*** - Evaluation	***GRX*** – **inter period**	***MRX*** – **inter period**	***GRX*** – **inter parameter set**	***MRX*** – **inter parameter set**
□	0.21	**0.85**	0.10	0.25	0.05	0.24
□	0.62	**0.85**	0.04	0.02	0.20	0.23
□	0.23	**0.75**	0.21	0.39	0.15	0.45
r	**0.88**	**0.88**	**0.79**	**0.73**	**0.95**	**0.92**
KGE	**0.86**	**0.92**	0.50	0.47	0.58	0.56
NSE	**0.72**	**0.81**	0.45	0.47	0.51	0.58

Table 3: Linear correlations of model performance metrics for different combinations of models, periods and parameter sets. Inter-period means that calibration and evaluation periods are compared (period #1 in calibration vs period #2 in evaluation, and conversely). Inter-parameter set means that models calibrated on period #1 (resp. period #2) are compared to models evaluated on period #1 (resp. period #2).

	GRX > MRX during both calibration & evaluation [%]	***MRX > GRX*** during both calibration & evaluation [%]
□	26	26
□	24	28
□	28	23
r	27	48
KGE	22	43
NSE	26	48

Table 4: Percentage of cases where a given model is better than the other during both calibration and evaluation.

8. Figures

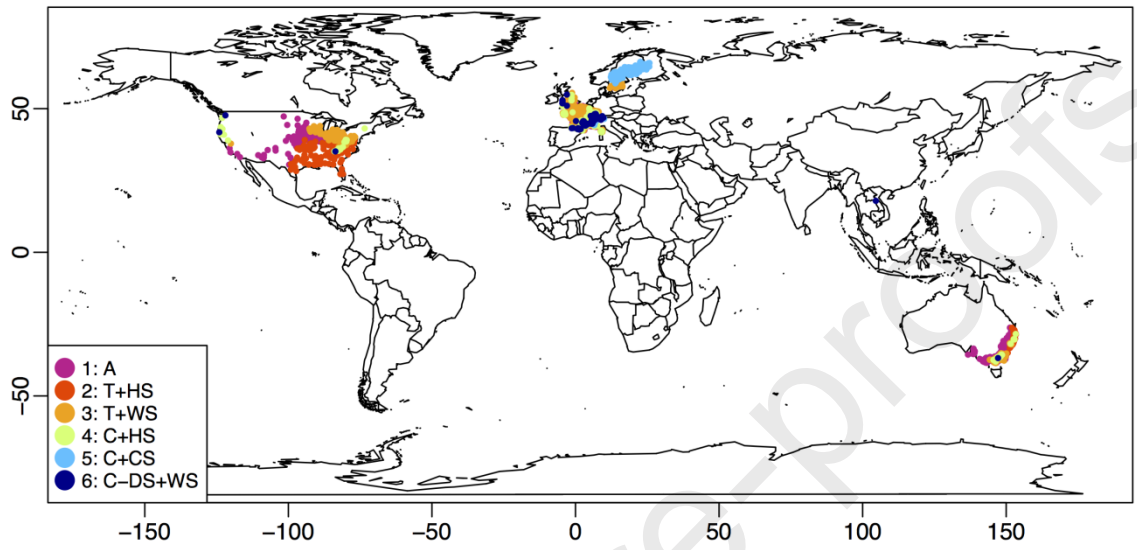


Figure 1: Locations of the 2,050 watersheds, sorted by hydroclimatic clusters (1 to 6, see Table 2 for details).

Color should be used in print

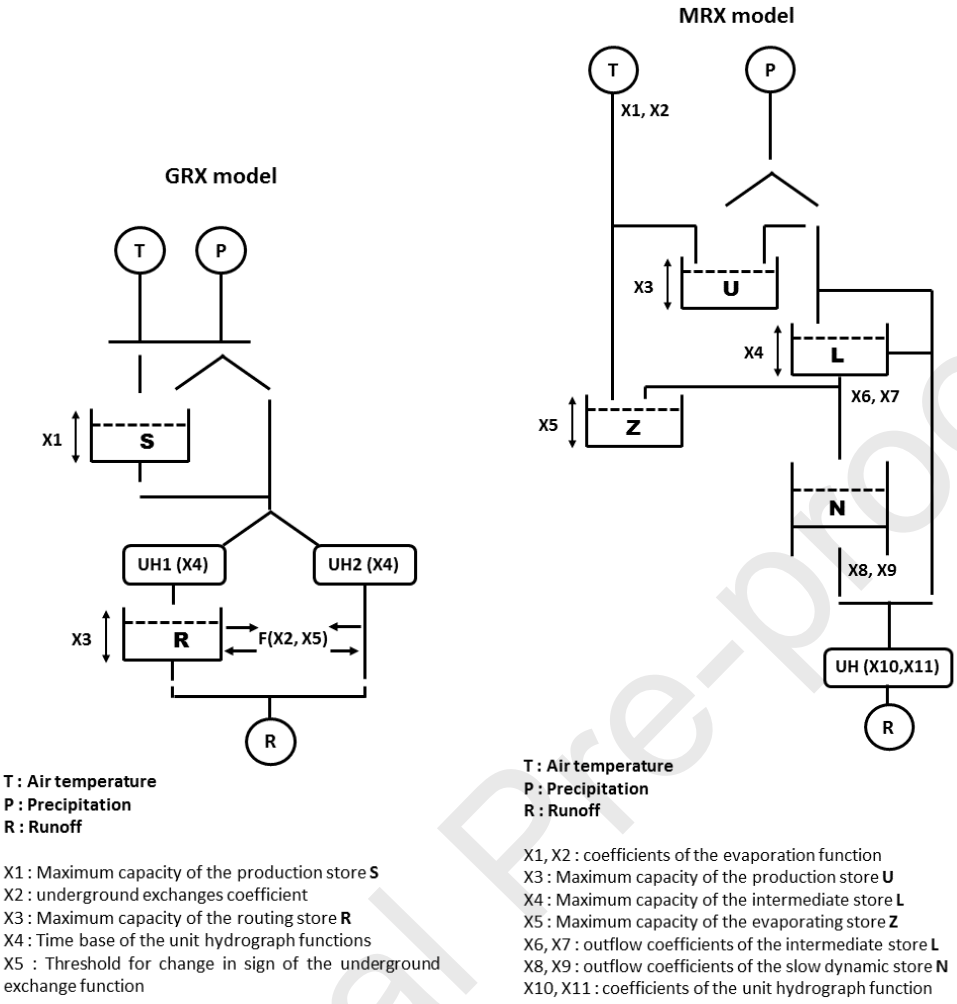


Figure 2: Structures and free parameters of the GRX and MRX hydrological models.

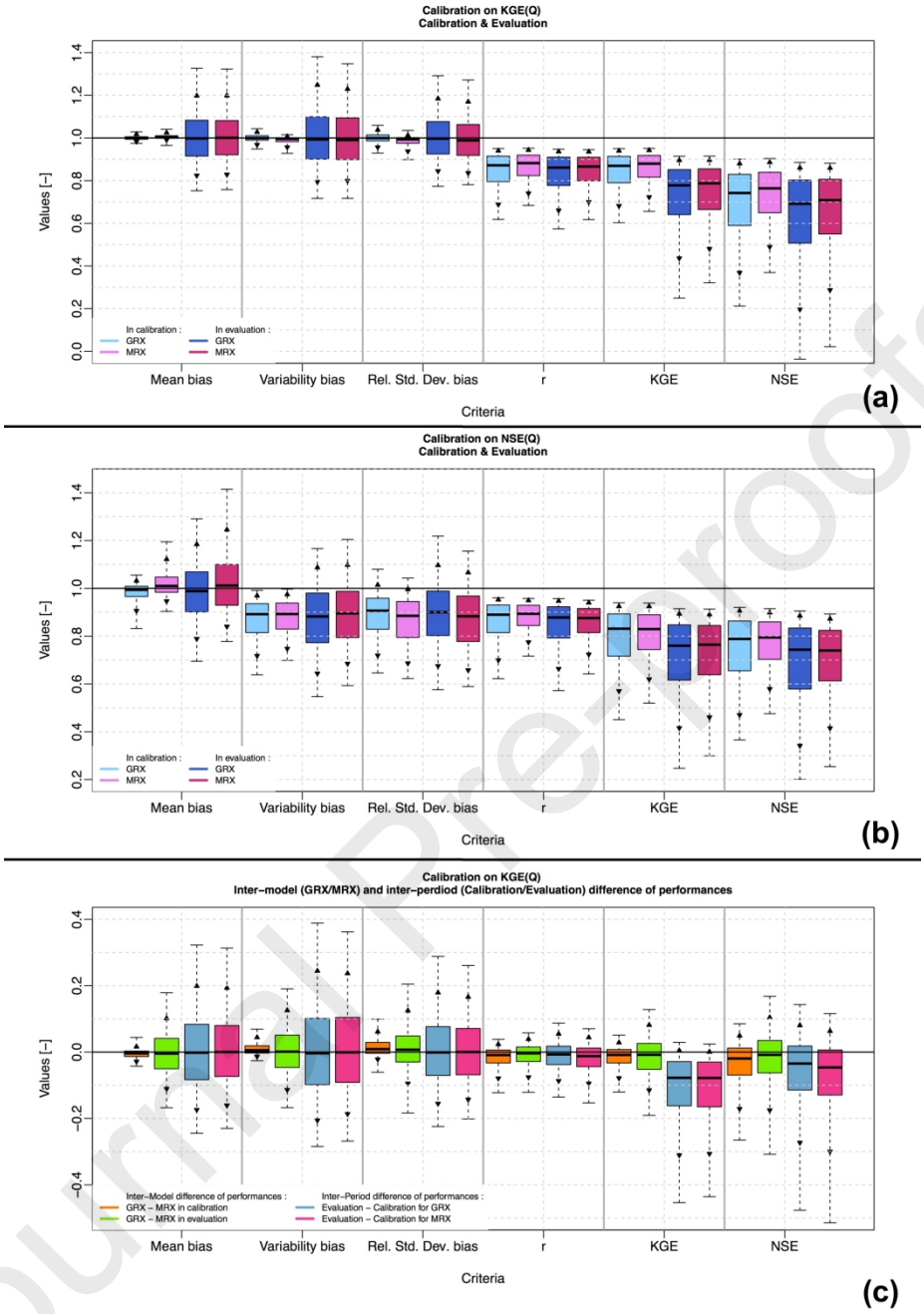


Figure 3: Comparison of distributions of ***GRX*** and ***MRX*** performance metrics (□□□□□□□**,** ***r, KGE, NSE***) during calibration and evaluation using either (a) ***KGE(Q)*** or (b) ***NSE(Q)*** as objective function. (c) Comparison of distributions of differences in ***GRX*** and ***MRX*** performance metrics (□□□□□□□**,** ***r, KGE, NSE***)

during calibration and evaluation using ***KGE(Q)*** as objective function. Boxplots represent the 5, 10, 25, 50, 75, 90 and 95 quantiles. **Color should be used in print.**

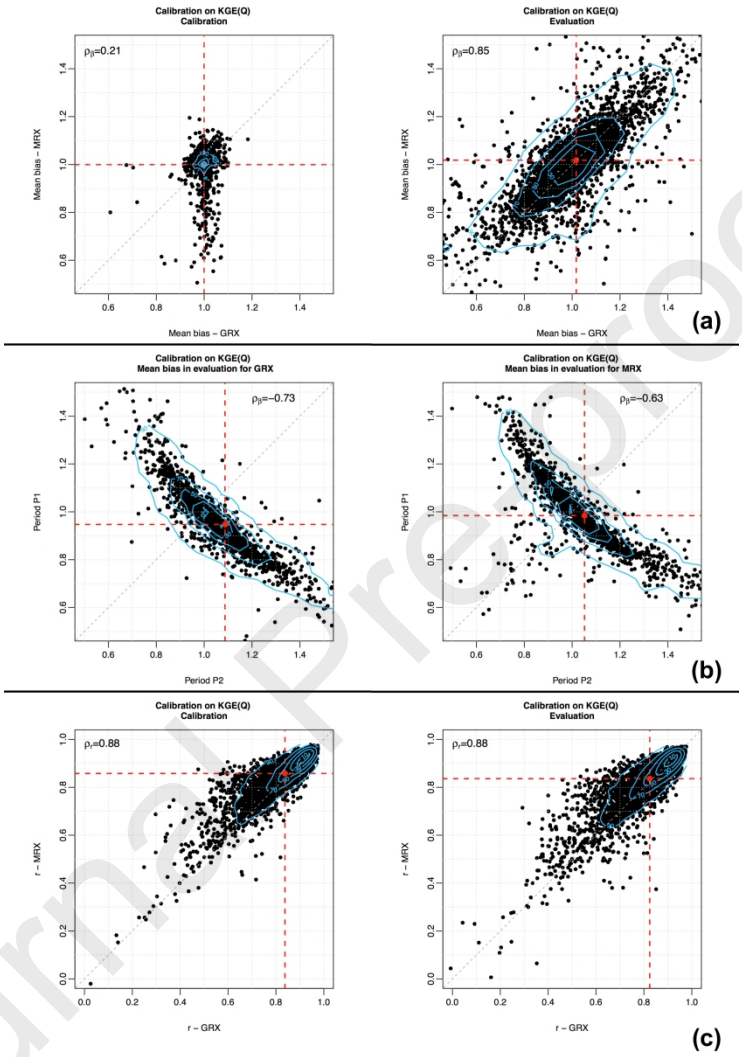


Figure 4: (a) Inter-model scatterplots (GRX versus MRX) of mean bias metric β performance during calibration (left) and evaluation (right) periods. (b) Inter-period scatterplots of evaluation period mean bias metric β performance for period P2 (parameter set calibrated on P1) versus period P1 (parameter set calibrated on P2) for GRX (left) and MRX (right). (c) Inter-model scatterplots (GRX versus MRX) of

correlation metric r performance during calibration (left) and evaluation (right) periods. Contour plots capture 10 to 90% of the points. Red dashed lines indicate the mean value. ρ □UOTErmance during calibration (left) and evaResults for variability bias and standard deviation bias are similar and not shown.

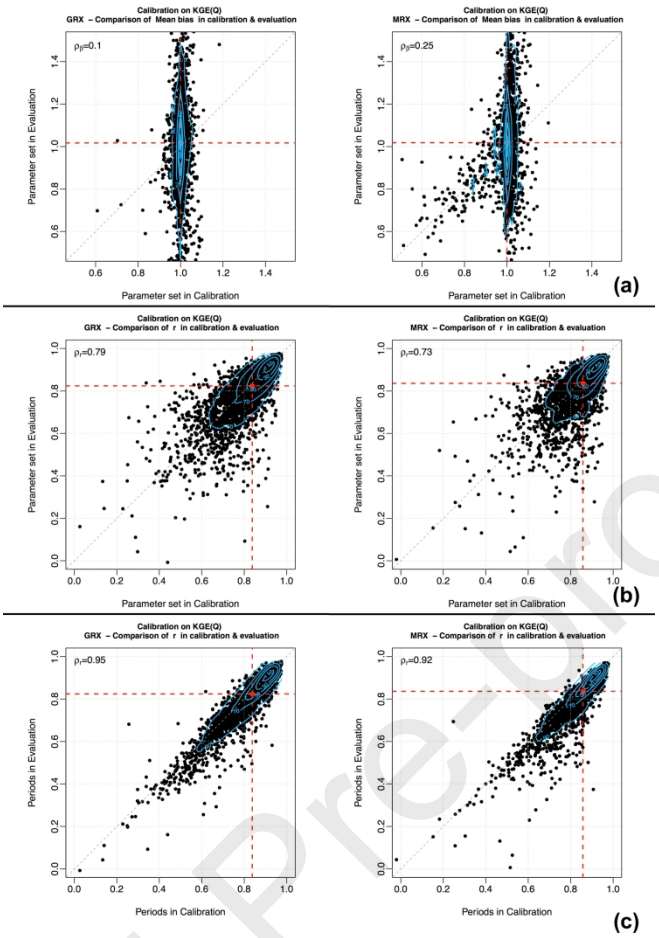


Figure 5: (a) Inter-period scatterplots (calibration versus evaluation periods) of mean bias β metric performance for GRX (left) and MRX (right). Being similar, results for variability bias α and coefficient of variation γ bias are not shown. (b) Inter-period scatterplots of correlation metric r performance for GRX (left) and MRX (right). The x-axis represents 'period two' evaluation (P2) when the model is calibrated on 'period one' (P1), and the y-axis represents 'period one' evaluation (P1) when the model is calibrated on 'period two' (P2)). (c) Inter-parameter set scatterplots of correlation metric r performance for GRX (left) and MRX (right). The x-axis represents model calibrated on 'period one' (resp. 'period two') and the

y-axis represents model evaluated on ‘period one’ (resp. ‘period two’). Contour plots capture 10 to 90% of the points. Red dashed lines indicate the mean value. ρ indicates the linear correlation statistic.

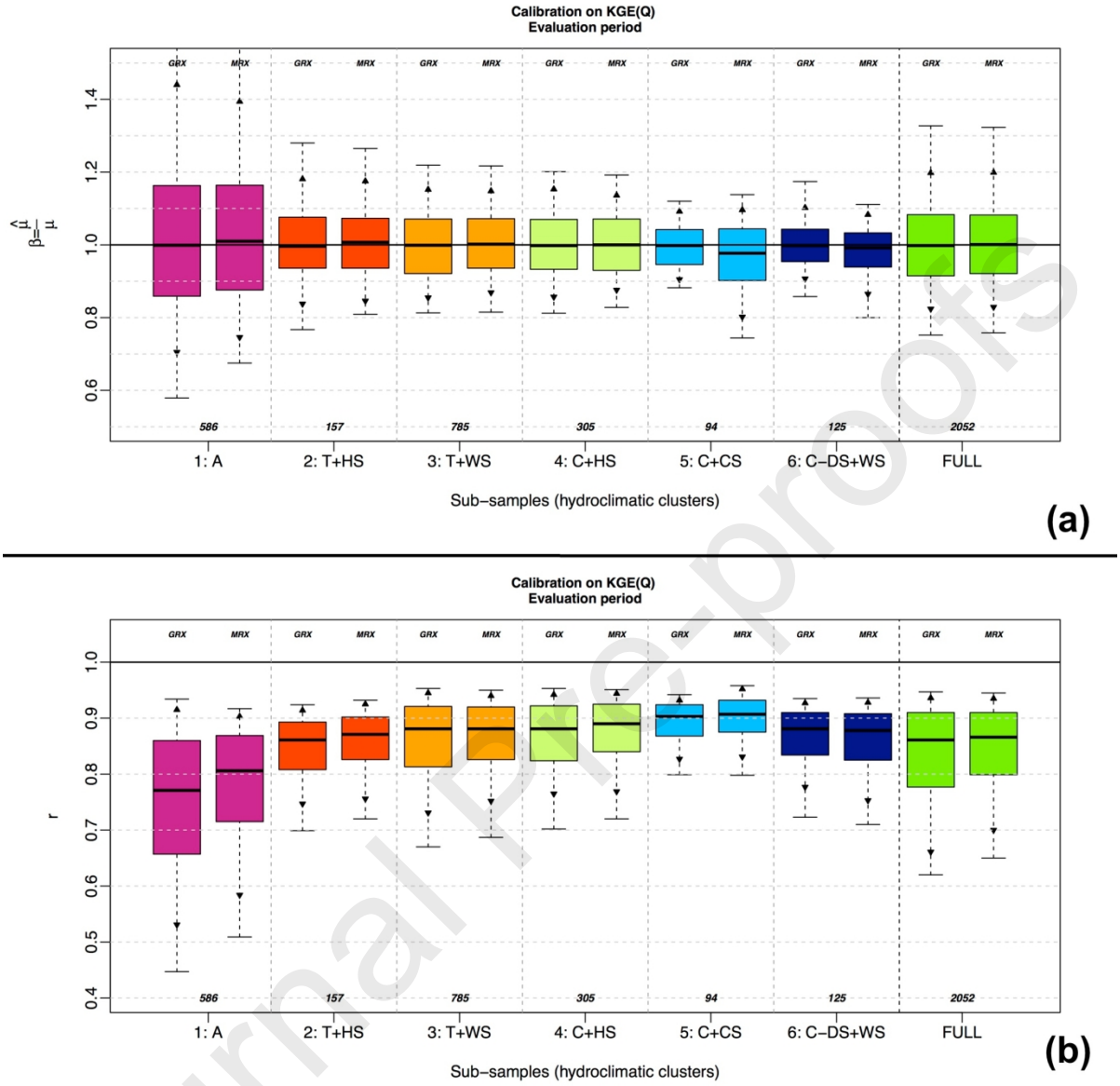


Figure 6: Distributional comparison of evaluation period model performance metrics (a) bias and (b) correlation, when KGE(Q) is used as objective function. Results are shown for six different hydro-meteorological clusters (see Table 2) and for the full sample. Boxplots represents the 5, 10, 25, 50, 75, 90 and 95 quantiles.

Color should be used in print.

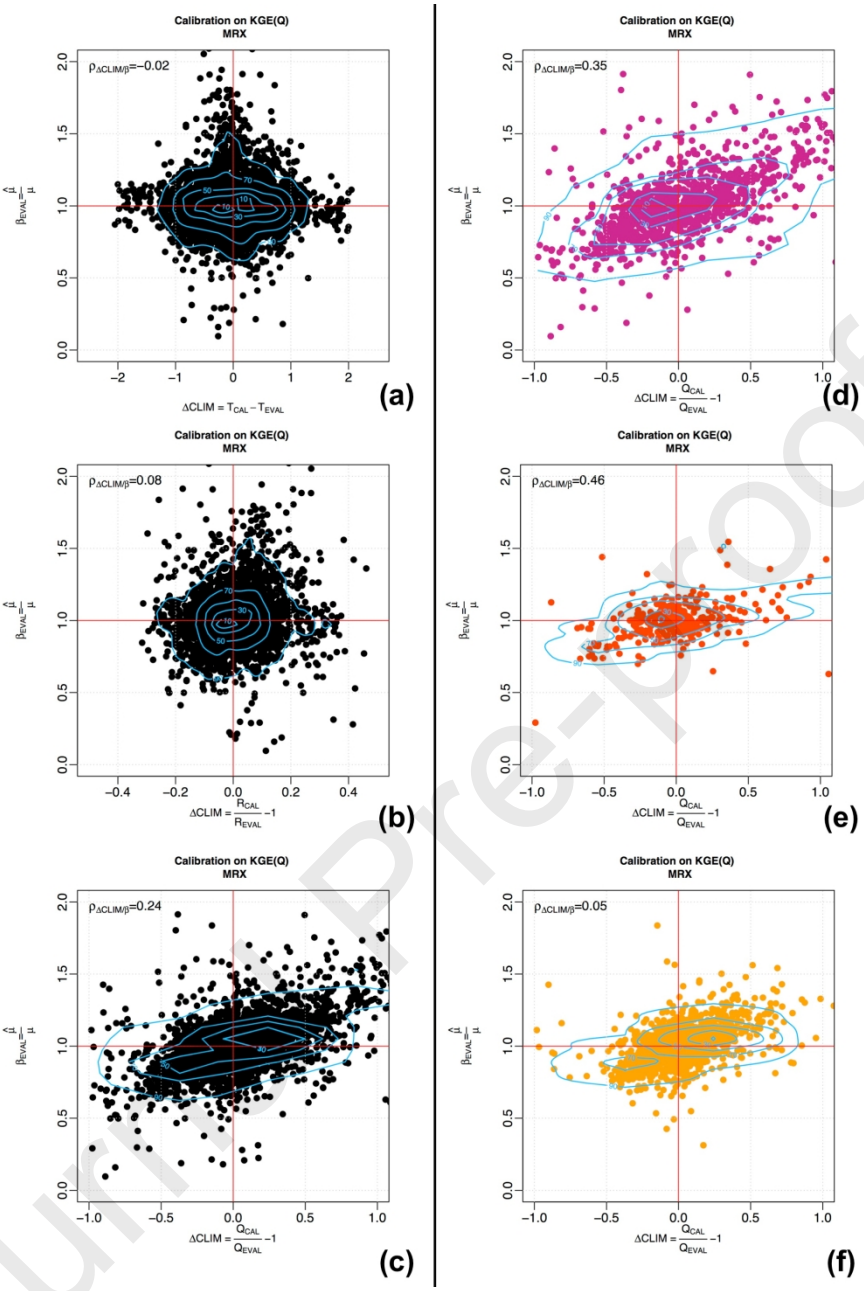


Figure 7: (a,b,c) Scatterplots of evaluation period bias □ and variability □CLIM of hydrometeorological characteristic (T, R or Q) from calibration to evaluation period. (d,e,f) Scatterplots of evaluation period bias and β variability □CLIM of runoff from calibration to evaluation period, for three different hydrometeorological clusters (1, 2 & 3). Results shown for MRX only.

Color should be used in print.

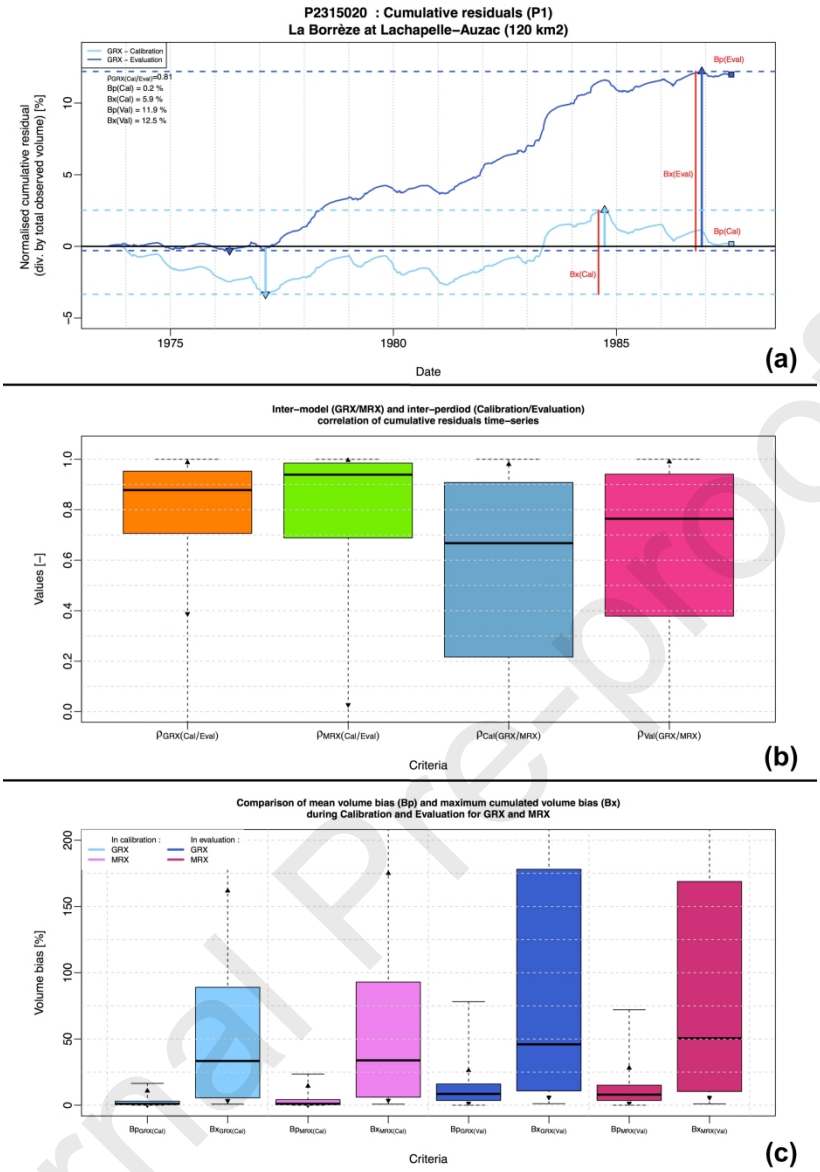


Figure 8: (a) Example showing the calibration and evaluation period normalized cumulative residual time-series for the GRX model on the La Borrèze at Lachapelle-Auzac (P2315020) watershed. Bp indicates the period mean volume bias and Bx indicates the period maximum volume bias. (b) Boxplots showing distributions of the correlation between calibration and evaluation period cumulative residuals time-series. (c) Boxplots showing distributions of the mean volume bias (Bp) and

maximum volume bias (Bx) during both calibration and evaluation. Boxplots represents the 5, 10, 25, 50, 75, 90 and 95 quantiles. **Color should be used in print.**

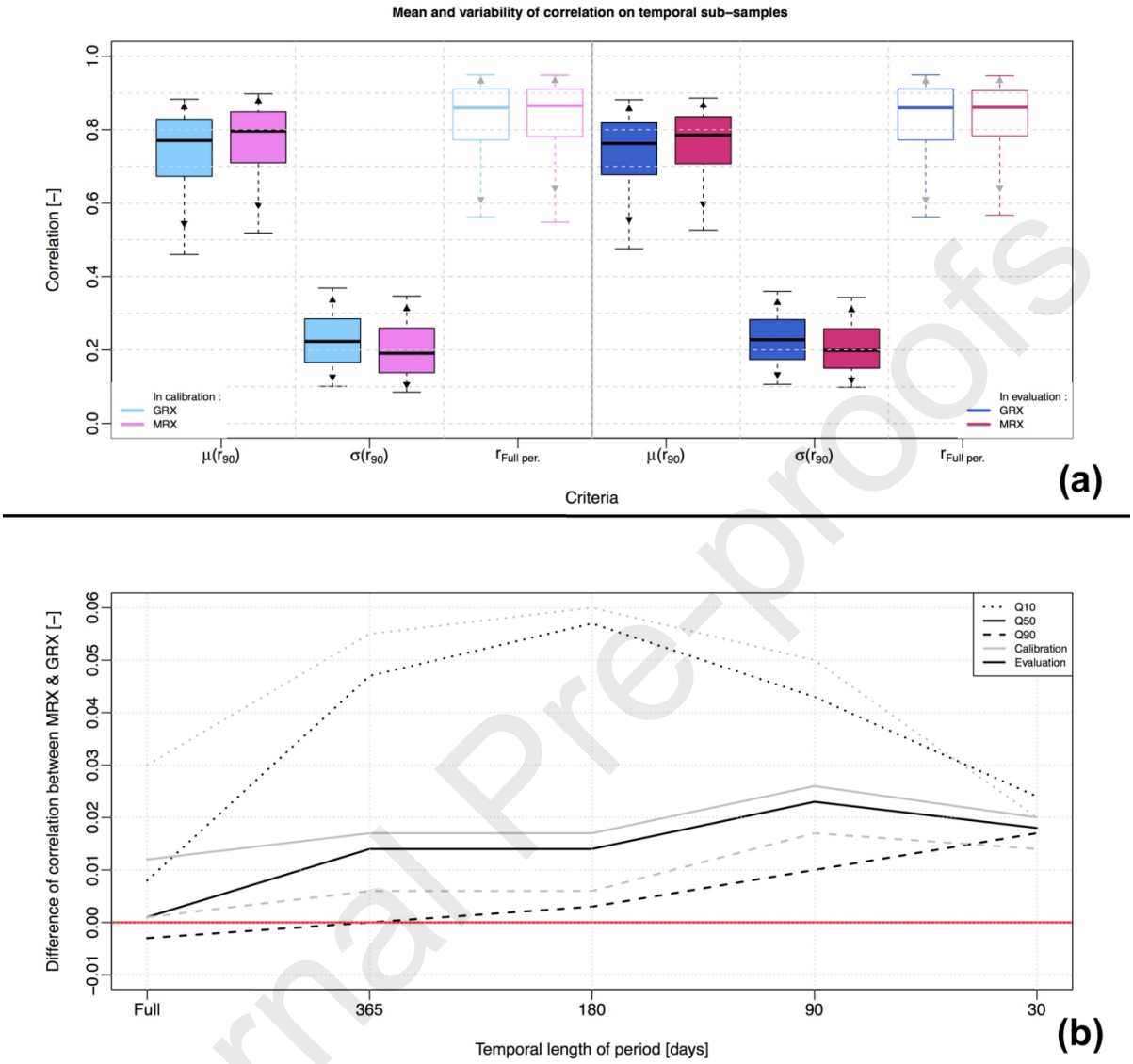


Figure 9: (a) Mean and variability of correlation over independent 90-days periods. (b) Mean and variability of correlation on 30-day to 365-day periods. **Color should be used in print.**

Assessing the Performance, Reliability and Robustness of Conceptual Rainfall-Runoff models: Analysis using a 'very large sample' of catchments

Thibault Mathevet[1], Hoshin Gupta[2], Charles Perrin[3], Vazken Andréassian[3], Nicolas Le Moine[4]

[1]Visiting research scholar at Hydrology and Atmospheric Sciences, University of Arizona, in 2014. EDF-DTG, 134 rue de l'Etang, 38950 Saint Martin le Vinoux, France

[2]Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, Arizona, USA

[3]National Research Institute of Science and Technology for Environment and Agriculture (Irstea), UR HBAN, Antony, France

[4]Université Pierre et Marie Curie, UMR Metis, Paris, France

Corresponding author: Thibault Mathevet (thibault.mathevet@gmail.com)

Abstract

To assess the predictive performance, reliability, robustness and generality of catchment-scale hydrological models, we conducted a detailed multi-objective investigation of two conceptual rainfall-runoff models, of differing complexity, on a '*very large catchment sample'* consisting of 2050

catchments worldwide. Our results indicate that both models provide (on average) similar levels of performance for water balance, variability and temporal correlation. Further, both models clearly suffer from lack of robustness when simulating water balance, with simulation performance depending more on the hydro-meteorological conditions of a given period than on the complexity of the model structure. We also show that long-term aggregate statistics (computed on the overall period) can fail to reveal considerable sub-period variability in model performance, thereby providing inaccurate diagnostic assessment of the predictive model performance.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Term	Authors	Definition
Conceptualization	Mathevet, Gupta	Ideas; formulation or evolution of overarching research goals and aims
Methodology	Mathevet, Gupta, Andreassian, Perrin, Le Moine	Development or design of methodology; creation of models
Software	Mathevet	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components
Validation	Mathevet, Gupta	Verification, whether as a part of the activity or separate, of the overall replication/ reproducibility of results/experiments and other research outputs
Formal Analysis	Mathevet, Gupta	Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data
Investigation	Mathevet, Gupta	Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection
Resources	Mathevet, Gupta, Andreassian, Perrin, Le Moine	Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools
Data Curation	Mathevet, Andreassian, Perrin, Le Moine	Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse
Writing – Original Draft	Mathevet, Gupta	Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation)

Term	Authors	Definition
Writing – Review & Editing	Mathevet, Gupta, Andreassian, Perrin, Le Moine	Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre-or postpublication stages
Visualization	Mathevet	Preparation, creation and/or presentation of the published work, specifically visualization/ data presentation
Supervision	Mathevet, Gupta	Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team
Revision	Mathevet, Gupta, Perrin, Andreassian	Revision of the paper
Project Administration	NA	Management and coordination responsibility for the research activity planning and execution
Funding Acquisition	NA	Acquisition of the financial support for the project leading to this publication