Poncelet, C., Merz, R., Merz, B., Parajka, J., Oudin, L., Andréassian, V., Perrin, C. (2017): Process-based interpretation of conceptual hydrological model performance using a multinational catchment set. - *Water Resources Research*, *53*, 8, pp. 7247—7268.

DOI: http://doi.org/10.1002/2016WR019991

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR019991

Key Points:
- The article uses large sample hydrology to identify catchment controls on daily runoff simulations from the GR6J lumped model
- Nonflashy, nonseasonal, large, and nonarid catchments show the best performance for four uncorrelated efficiency criteria
- The study underline the value of multinational dataset to increase results robustness

Supporting Information:
- Supporting Information S1
- Data Set S1

Correspondence to:
C. Poncelet,
carine.poncelet@irstea.fr

Citation:
Poncelet, C., R. Merz, B. Merz, J. Parajka, L. Oudin, V. Andréassian, and C. Perrin (2017), Process-based interpretation of conceptual hydrological model performance using a multinational catchment set, *Water Resour. Res.*, *53*, 7247–7268, doi:10.1002/2016WR019991.

Received 21 OCT 2016
Accepted 16 JUL 2017
Accepted article online 20 JUL 2017
Published online 22 AUG 2017

Process-based interpretation of conceptual hydrological model performance using a multinational catchment set

Carine Poncelet[1], Ralf Merz[2], Bruno Merz[3], Juraj Parajka[4], Ludovic Oudin[5], Vazken Andréassian[1], and Charles Perrin[1]

[1]Irstea, UR Hydrosystèmes et Bioprocédés (HBAN), Antony, France, [2]UFZ German Research Centre for Environment, Catchment Hydrology Team, Halle (Saale), Germany, [3]GFZ German Research Centre for Geosciences, Section Hydrology, Potsdam, Germany, [4]TUW, Institute of Hydrology and Water Resource Management, Vienna, Austria, [5]Sorbonne Universitrés, UPMC, Paris 6, UMR Metis, Paris, France

Abstract Most of previous assessments of hydrologic model performance are fragmented, based on small number of catchments, different methods or time periods and do not link the results to landscape or climate characteristics. This study uses large-sample hydrology to identify major catchment controls on daily runoff simulations. It is based on a conceptual lumped hydrological model (GR6J), a collection of 29 catchment characteristics, a multinational set of 1103 catchments located in Austria, France, and Germany and four runoff model efficiency criteria. Two analyses are conducted to assess how features and criteria are linked: (i) a one-dimensional analysis based on the Kruskal-Wallis test and (ii) a multidimensional analysis based on regression trees and investigating the interplay between features. The catchment features most affecting model performance are the flashiness of precipitation and streamflow (computed as the ratio of absolute day-to-day fluctuations by the total amount in a year), the seasonality of evaporation, the catchment area, and the catchment aridity. Nonflashy, nonseasonal, large, and nonarid catchments show the best performance for all the tested criteria. We argue that this higher performance is due to fewer nonlinear responses (higher correlation between precipitation and streamflow) and lower input and output variability for such catchments. Finally, we show that, compared to national sets, multinational sets increase results transferability because they explore a wider range of hydroclimatic conditions.

1. Introduction

Achieving accurate streamflow simulations is a common objective to most hydrological modelers. To this end, modelers typically focus on: (i) the quality of model inputs [*Gupta and Sorooshian*, 1985; *Oudin et al.*, 2006; *Arheimer et al.*, 2012], (ii) the improvement of model structures [*Perrin et al.*, 2003; *Das et al.*, 2008; *Fenicia et al.*, 2011] or (iii) model calibration [*Duan et al.*, 2006; *Kuzmin et al.*, 2008; *Efstratiadis and Koutsoyiannis*, 2010] or regionalization [*Hrachowitz et al.*, 2011; *Parajka et al.*, 2013]. Advances in these areas resulted in a wide variety of models and modeling setups, none of them systematically outperforming the others [*Pechlivanidis et al.*, 2011; *Clark et al.*, 2016].

To improve streamflow simulations, the evaluation of model performance is of primary importance. Yet, there is no general agreement on a standard procedure for evaluating model performance [*Ritter et al.*, 2013] or on what is actually a "good" simulation [*Crochemore et al.*, 2015]. Almost every modeling study evaluates the model performance, but the results are often fragmented, based on different methods or time periods and do not link the results to landscape or climate characteristics. Moreover, the results are typically analyzed for small number of catchments or only in individual countries. For example, *Merz et al.* [2009] evaluated performance of a conceptual hydrologic model in 269 Austrian catchments and reported an increase in runoff model efficiency with increasing size of the catchments. Similar increase was found in *van Esse et al.* [2013] for 237 catchments in France. This analysis reported also better model performance in wetter than in drier catchments. Aridity, precipitation intermittency, and runoff seasonality were found as the main factors influencing variations in model performance also in 671 US catchments *Newman et al.* [2015]. These results correspond well with previous spatial performance patterns of *Clark et al.* [2008] who applied many conceptual models to a subset of the MOPEX basin set and found poor performance in arid regions.

The previous studies typically analyzed the links between model performance and catchment characteristics by using small number of catchment attributes. The most common characteristics are size of catchment, aridity, mean catchment elevation or precipitation [*Parajka et al*., 2013]. Moreover, the links between model efficiency and catchment characteristics are usually considered one at a time and the interplay between landscape and climate characteristics is ignored. The main objective of this study is to investigate the link between daily runoff simulations and climate and landscape characteristics using a large multinational data set. The analysis is performed for 1103 catchments in Europe, by using 29 catchments characteristics and four model efficiency criteria obtained by the GR6J rainfall-runoff model, a lumped conceptual model that already proved to be a particularly competitive model on a variety of French catchments [*Pushpalatha et al*., 2011]. The specific research questions are: (i) What are the relationships between model performance and catchment characteristics? (ii) Can these relationships be interpreted based on what we know of hydrological processes?, and (iii) Does the multinational set improve the transferability of results compared to national analyses? The paper is organized as follows: sections 2 and 3 describe the data and methods designed for this study, section 4 presents and discusses the results, section 5 summarizes the findings.

2. Data

2.1. Databases Presentation

The features used for catchments description are derived from databases with contrasted spatial resolution: global/European data sets for physical features and national data sets for climate and streamflow features. The Shuttle Radar Topography Mission (SRTM) uses radar imaging to provide high quality, global maps of elevation at a 100 m resolution [*Rodriguez et al*., 2006]. The Corinne Land Cover (CLC) data set is constructed by coupling satellite images with photointerpretation. Vector maps at a 1/100,000 resolution are produced at the European scale and provide good quality estimates of land cover [*EEA*, 2007]. SRTM and CLC databases have both been successfully used for hydrological applications [*Lehner and Grill*, 2013; *Duan et al*., 2006]. The European Soil Database (ESDB) classify European soils according to the FAO85 recommendations [*Nachtergaele*, 2008]. Based on this classification, pedotransfert rules [*King et al*., 1994] are applied to derive advanced soil characteristics such as texture or depth at a 1000 m resolution. To our knowledge, the ESDB has not been used in previous hydrological modeling papers, probably because more detailed data exist at the national scale. The confidence level maps included in the ESDB typically show a moderate data quality for the three countries [*Finke et al*., 2001].

Precipitation and temperature for Austria are gauge-based [*Merz et al*., 2011]. The daily values of precipitation and air temperature were spatially interpolated by methods using elevation as auxiliary information. External drift kriging was used for precipitation and the least squares trend prediction method was used for air temperature. Precipitations and temperature for France come from the Safran analysis [*Vidal et al*., 2010]. Safran is a gauge-based analysis system using the optimal interpolation (OI) method. The OI technique computes the analyzed value by modifying a first-guess field (e.g., prediction model Arpege or ECMWF operational archives) with the weighted mean of the differences between observed and first-guess values at station locations within a search distance. Precipitations for Germany are derived from the REGNIE gauge-based analysis [*Rauthe et al*., 2013; *Görgen et al.*, 2010 (in the supporting information)]. Interpolation of station data in REGNIE combines background fields and residuals regionalization. Background fields are produced using multiple linear regression of five explanatory variables (geographical longitude and latitude, height above sea level, exposition and slope at the stations). Residuals between observation and background field value at the station are regionalized using the inverse distance weights scheme. Temperature for Germany is obtained by interpolation of station data. Interpolation was performed using external drift kriging with elevation as explanatory variable.

The national rainfall-runoff data have been extensively used and the data quality is high for the three countries. Differences in the measurement density and the interpolation procedures between the countries impact features computation. However, because rainfall-runoff data quality impact model performance and because our goal is to assess what catchment features affect model performance, it was necessary to model runoff using the most accurate available rainfall-runoff data. This is why the national data sets were used and not global climate data sets.

2.2. Catchment Features

The catchments features describe each catchment in terms of their physical, climate, and streamflow characteristics. Climate and streamflow features are computed over the 1978–2002 period, for which all catchments have less than 3 years of missing streamflow data (see Table 1).

The soil available water content (*AWC*) is computed as follow:

$$AWC = (AWCs + AWCd) * DR \tag{1}$$

with *AWCs* [mm/m] and *AWCd* [mm/m] the soil available water content of the superficial and deep soil layers, respectively, and *DR* [m] the soil depth. *AWCs*, *AWCd*, and *DR* are directly obtained from the ESDB database.

The actual evapotranspiration is computed using the Turc formula [*Turc*, 1954]:

$$AE_i = \frac{P_i}{\left[1 + \left(\frac{P_i}{E0_i}\right)^n\right]^{\frac{1}{n}}} \tag{2}$$

with AE^i the actual evapotranspiration of year *i*, P^i the precipitation during year *i*, $E0^i$ the potential evapotranspiration in year *I*, and *n* the exponent (chosen at $n = 2$).

The flashiness is quantified by the Richards-Baker flashiness index, which is the ratio of absolute day-to-day fluctuations of the variable of interest by the total amount in a year [*Holko et al.*, 2011]:

$$FI_i = \frac{\sum |X(t_i) - X(t_{i-1})|}{\sum X(t_i)} \tag{3}$$

with *FI* the flashiness index, *X* the flow of interest (*E*0, *P* or *Q*), *i* the year, and t^j the day within year *i*. *FI* is a dimensionless measure which ranges between 0 and 2. Zero represents an absolutely constant flow; increased *FI* values indicate increased flashiness (fluctuations) of flow. When streamflow is the features of interest, the flashiness is comparable to streamflow autocorrelation.

The fraction of solid precipitation is computed based on air temperature [*L'hôte et al.*, 2005]:

Table 1. List of Features Used in This Study[a]

Name	Abbreviation and Units	Computed From	Reference	Aggregation
Area	*A* (km^2)	Topographic maps		
Elevation	*Z* [m]	DEM from the Shuttle Radar Topography Mission (SRTM)	*Rodriguez et al.* [2006]	*m*, *cv*
Soil available water content	*AWC* (mm)	Computed from the European Soil Database (ESDB)	*Finke et al.* [2001]	*m*, *cv*
Soil depth	*DR* (cm)	ESDB	*Finke et al.* [2001]	*m*, *cv*
Percentage of forest	*pF* ()	Corinne Land Cover 2006	*EEA* [2007]	
Streamflow	*Q* (mm/y)	HYDRO database for France, The State Offices for Germany, Hydrographic service of Austria (HZB)	HYDRO [*Leleu et al.*, 2014], The State Offices for Germany (detailed in Acknowledgements section), HZB (ehyd.gv.at)	*m*, *cv*, *ir*
Precipitation	*P* (mm/y)	SAFRAN for France, Deutsche Wetter Dienst (REGNIE) for Germany, Interpolation of station data for Austria	SAFRAN [*Vidal et al.*, 2010], REGNIE [*Rauthe et al.*, 2013], [*Merz et al.*, 2011 for Austria]	*m*, *cv*, *ir*
Actual evapotranspiration	AE (mm/y)	Computed from precipitation and potential evaporation	*Turc* [1954]	*m*, *cv*, *ir*
Potential evapotranspiration	*E*0 (mm/y)	Computed from temperature (same sources as precipitation)	*Oudin et al.* [2005] for France and Germany, *Parajka et al.* [2003] for Austria	*m*, *cv*, *ir*
Flashiness of *E*0	*FIE* (–)	Computed from potential evaporation	Adapted from *Holko et al.* [2011]	*m*, *cv*
Fashiness of *P*	*FIP* (–)	Computed from precipitation	Adapted from *Holko et al.* [2011]	*m*, *cv*
Flashiness of *Q*	*FIQ* (–)	Computed from streamflow	*Holko et al.* [2011]	*m*, *cv*
Fraction of solid precipitation	*Fs* (–)	Computed from temperature	*L'hôte et al.* [2005]	
Aridity index	*AI* (–)	Computed from precipitation and potential evaporation	*Budyko* [1974]	
Water yield	*WY* (–)	Computed from precipitation and streamflow		

[a]If the feature displays variability, the aggregation methods are gathered in the "aggregation" column (*m* is the arithmetic mean, *cv* is the coefficient of variation and *ir* is seasonality) and specified in the text.

$$\begin{cases} F_T(t)=0 & si\ T > 3^{\circ}\mathrm{C} \\ F_T(t)=1-\dfrac{T(t)-(-1)}{3-(-1)} & si\ -1 < T < 3^{\circ}\mathrm{C} \\ F_T(t)=1 & si\ T < -1^{\circ}\mathrm{C} \end{cases} \quad (4)$$

$$Fs=\frac{\sum F_T(t) * P(t)}{\sum P(t)} \quad (5)$$

with T the air temperature on day t, $FT(t)$ the fraction of solid precipitation of day t, and Fs the fraction of solid precipitation used for this study.

The aridity index is defined as the ratio of the long-term mean potential evaporation to the long-term mean precipitation [*Budyko*, 1974]:

$$AI=\frac{\overline{E0}}{\bar{P}} \quad (6)$$

with AI the aridity index, P and $E0$ the precipitation and potential evaporation derived from the national rainfall-runoff data set (see Table 1).

The water yield is defined as the ratio of long-term mean streamflow over the long-term mean precipitation:

$$WY=\frac{\bar{Q}}{\bar{P}} \quad (7)$$

with the same notations as above and WY the water yield.

Most of the features display variability (spatial or temporal), because they are computed either per unit of space or per unit of time. Since we are only able to assess model performance at the catchment outlet, we need a single value per catchment to link it with model performance. Hence the question of how each feature is aggregated at the catchment scale is important. We used:

1. the arithmetic mean (m) to describe the overall quantity,
2. the coefficient of variation (cv) to describe variability. Because the climate and streamflow features are computed from different temporal resolutions, the coefficient of variation can refer to different types of variability. In particular for P, Q, and $E0$, the coefficient of variation refers to daily variability. On the other hand for AE, FIE, FIP, and FIQ, the coefficient of variation refers to annual variability,
3. the coefficient of irregularity (ir) to describe the seasonality of climate-related and streamflow-related features [*Mouelhi*, 2003]:

$$ir=\frac{\max(X_m)-\min(X_m)}{\overline{X_m}} \quad (8)$$

with X^m the monthly value of precipitation, evaporation, or streamflow averaged over 1978–2002.

The names of the features were abbreviated in capital letters (see Table 1) and in lower case for the aggregation method. For the few features that express no variability (e.g., the Area or the Aridity Index), only the feature's capital abbreviation is used. Hereafter the term "feature" will include both the feature and its aggregation at the catchment scale. The correlation matrix between the aggregated features is gathered in the supporting information (Figure 2).

2.3. A Multinational Catchment Set

Catchments were chosen according to several criteria: (i) availability of streamflow records over the 1978–2002 period, i.e., less than 3 years of missing data and (ii) unimpacted catchments, i.e., less than 20% artificial land cover [*EEA*, 2007] within the catchment.

Figure 1 shows that most of the catchments are located in France (580 catchments) followed by Germany (309 catchments) and Austria (214 catchments). Each country has specific attributes:

1. Austria has a varied climate with low precipitation in the eastern lowland regions to high precipitation in the western alpine regions. The country is flat or undulating in the east and north, and Alpine in the west and south. In the Alpine parts, the hydrological dynamics are strongly controlled by the seasonal variation of glacier and snow accumulation and melt. In the lower parts, the hydrological regime is more driven by the spatiotemporal variability of rainfall.
2. France has a mainly temperate climate, but its climate conditions are varied: Mediterranean conditions in the south of France, oceanic influences in the west, continental features in the eastern parts and mountainous influences in the Pyrenees and the Alps. The database contains mountainous catchments where snowmelt-fed regimes are observed, small Mediterranean catchments and larger temperate catchments where rainfall and evaporation drive the seasonal variations of runoff as well as groundwater-dominated catchments in the north.
3. Germany is in a transition zone between its maritime climate in the west and a continental climate in the east. Precipitation is dominated by westerly circulation patterns, but large rainfall events can also be produced by other circulation patterns. In the northwest lowlands, winter precipitation immediately affects runoff (pluvial runoff regime), and maximum runoff occurs during the winter months. To the east, the influence of snowcover on seasonal runoff increases. In the low mountain ranges, temporary snow deposits delay the maximum runoff into the spring (nivopluvial regime).

As a result, the main catchment features over the catchment set are regionally variable. Table 2 provides a summary of main catchment features over the multinational set. Figure S1 in the supporting information illustrates the catchments water balance. The shapefile uploaded as supporting information data set contains the catchments boundaries and all features used in this study.

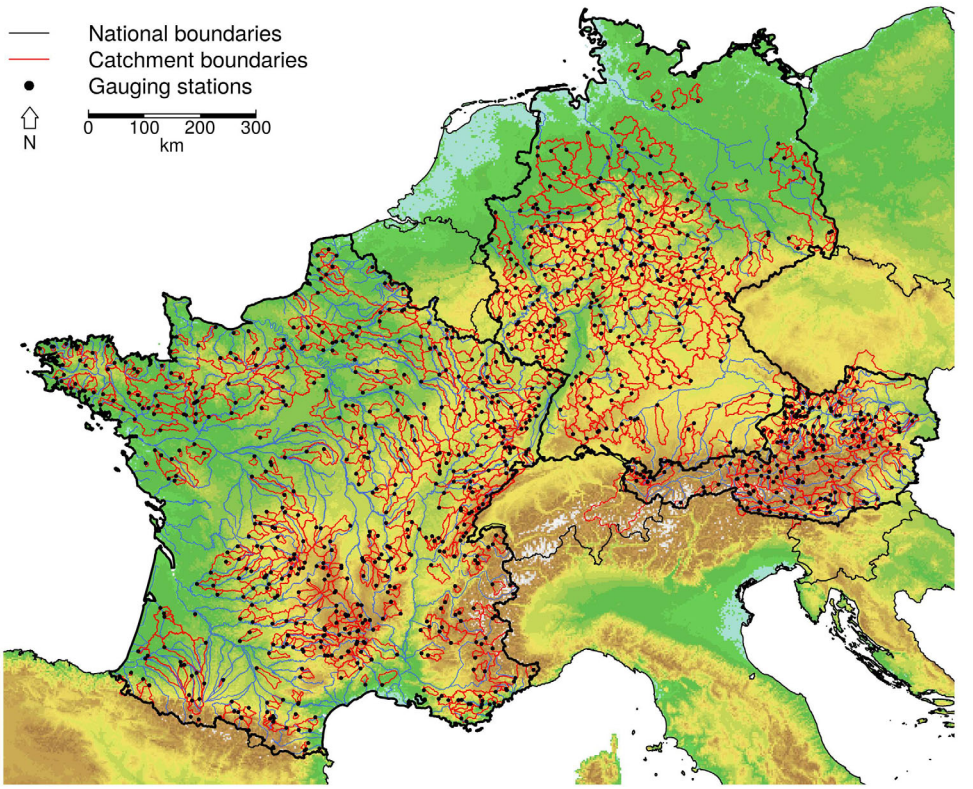

Figure 1. Location of the 1103 catchments in Austria, France and Germany used in the study. Some catchments in the set are nested: the smaller catchments are represented on top of the larger catchments.

Table 2. Quantiles of the Distribution of Main Catchment Features Over the 1103 Catchments Studied

	Minimum	10th	25th	50th	75th	90th	Maximum
Area (km^2)	5	60	120	250	730	2240	27,000
Mean elevation (m a.s.l.)	28	130	270	430	780	1250	2920
Aridity index (–)	0.20	0.39	0.50	0.66	0.77	0.89	1.51
Mean actual evaporation (mm/y)	200	450	480	530	570	600	710
Irregularity of actual evaporation (–)	0.01	0.15	0.17	0.19	0.23	0.28	0.63
Mean precipitation (mm/d)	1.5	2.1	2.3	2.7	3.3	4.1	6.4
Coefficient of variation of precipitation (mm/d)	1.60	1.75	1.80	1.90	2.03	2.35	4.16
Mean flashiness of precipitation (–)	0.99	1.09	1.12	1.18	1.24	1.30	1.52
Coefficient of variation of streamflow flashiness (–)	0.06	0.11	0.13	0.16	0.21	0.28	0.64
Irregularity of streamflow (–)	0.21	0.86	1.19	1.43	1.75	2.04	3.55
Fraction of solid precipitation (–)	0	0.02	0.04	0.09	0.15	0.24	0.68
Water yield (–)	0.06	0.25	0.32	0.40	0.54	0.75	2.74

3. Method

3.1. Hydrological Model

3.1.1. Description

GR6J [*Pushpalatha et al*., 2011] is a lumped model here applied at a daily time step with six free parameters (see Figure 2). Since some of the catchments are located in mountainous areas, the CemaNeige snow accounting routine [*Valéry et al*., 2014] is used in addition to the hydrological model. The model is fed with daily precipitation (*P*) and daily potential evapotranspiration (*E*0). The daily temperature (*T*) is only an input to the snow accounting routine and is used to compute the solid part of precipitation (*Fs*) and the snowpack evolution. GR6J has three conceptual stores:a production store used to compute the actual evapotranspiration (*Es*) and the water amount that reaches the routing store (*Pr*). It is described by its capacity, the *X1* parameter (mm). *Es* and *Ps* are both computed based on *X1* and the level in the store (*S*) and on *Pn* and *En*, respectively.

1. the routing store used to reproduce part of the flow routing (routed flows). It is described by its capacity, the parameter *X3* (mm). In every time step, the routed flows are independent of the soil moisture state and account for 90% of *Pr*.
2. the exponential store used to reproduce long recessions and low flows. It is controlled by the *X6* parameter (mm), a base level in the store.

Two parameters contribute to adjust the catchments water balance through the nonatmospheric exchange function (*L* [mm/d]). *L* computes the quantity of water that is considered lost to/gained from groundwater aquifers or neighboring catchments. It is controlled by two dimensionless parameters *X2* (multiplicative parameter) and *X5* (additive parameter) as a function of the filling rate in the routing store. The reaction time of the catchment is expressed with two unit hydrographs: *UH1* (for routed flows) and *UH2* (for direct flows). Both unit hydrographs share the same base time, i.e., they are controlled by the *X4* parameter (day). Streamflow amounts are regulated mainly by the combination of *X1*, *X2*, and *X5* whereas streamflow time variability is handled by the combination of *X3*, *X4*, and *X6* and, to a lesser extent, by *X1*. Consequently, it is not possible to relate the hydrological response (and hence model performance) directly to individual parameter values.

The CemaNeige snow accounting routine is a snow accumulation/melt module based on the degree-day concept. The snow water equivalent of the snowpack is computed using two parameters: *Ctg* [mm/°C] that describes the thermal inertia of the snowpack and *Kf* [–] a degree-day melting factor. The higher *Ctg* the later the snowmelt and the higher *Kf* the larger the snowmelt.

Because GR6J is built up from simple concepts such as the association of reservoirs and unit hydrograph, it is similar to many classical models such as HBV [*Bergström*, 1995] or VIC [*Liang et al*., 1994]. *Pushpalatha et al*. [2011] compared the performance of GR6J with five other hydrological models on 1000 French catchments and found GR6J's performance competitive. For these two reasons, we considered GR6J to be a good candidate for this experiment. The conclusions drawn are not model-independent but provide general insights into the major catchment controls on daily runoff simulations.

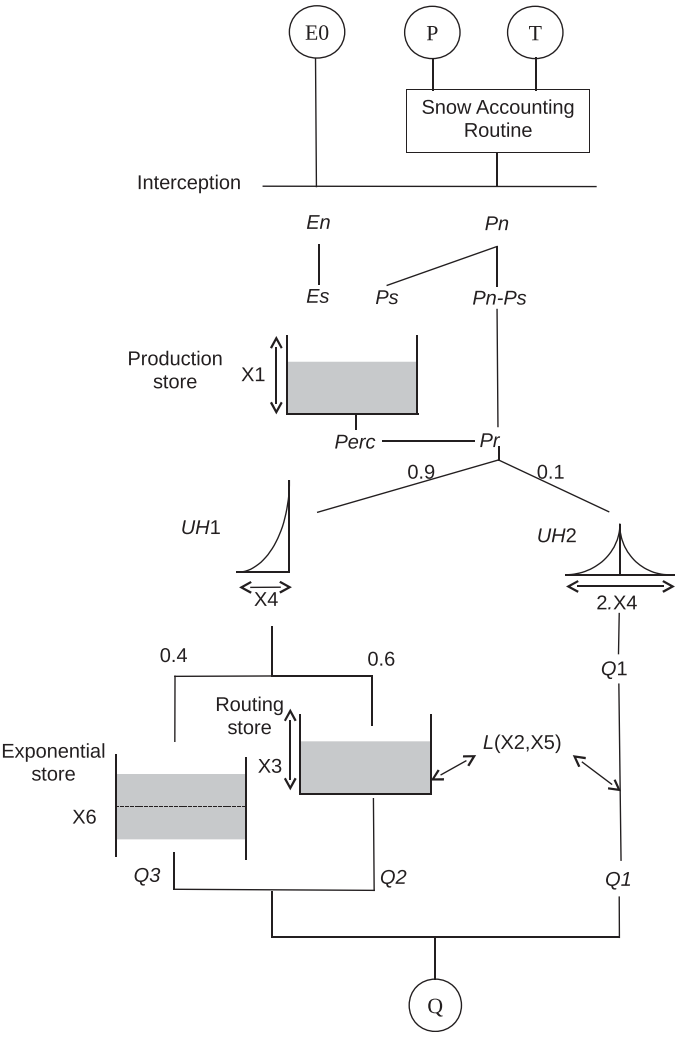

Figure 2. Schematic representation of the GR6J hydrological model with *E*0 potential evapotranspiration, *P* precipitation, *T* temperature, *Q* streamflow. The letter *X* (*X1*, . . ., *X6*) refers to the model parameters. The other letters (*Pn*, *En*, . . .) refer to internal variables, i.e., the water quantities exchanged between the reservoirs. A complete description of the model equations can be found in *Pushpalatha et al.* [2011].

3.1.2. Calibration Strategy

The six parameters of the hydrological model (*X1*, . . ., *X6*) were calibrated for the period between October 1982 and September 1992. The two parameters of CemaNeige were not calibrated but set at the default values of Ctg = 0.25 mm/°C and *Kf* = 3.74. The noncalibration of the CemaNeige parameters do not impact the performance during the validation period. The validation period spans between October 1992 and September 2002. Both periods were preceded by 4 years of warm-up to initialize the content of the stores. We calibrated the model using a single objective function, the Kling-Gupta Efficiency [*Gupta et al.*, 2009] on square-rooted streamflow. Tests, not presented here for the sake of clarity, showed that the results are mostly unimpacted by the choice of the objective function. Hence, we chose a single and simple objective function for calibration to provide more easily transferable results. The optimization algorithm used to calibrate the parameters is a dual global-local strategy. The global search on a coarse grid identifies the best starting point for a local algorithm as presented by *Edijatno et al.* [1999]. It uses a steepest descent method to move step by step in the parameter space, toward the optimum parameter set. This method was tested in several studies and is suitable for models having up to eight parameters to calibrate [*Edijatno et al.*, 1999].

3.2. Performance Assessment

The quality of streamflow simulations is assessed using four efficiency criteria (see Table 3).

We considered the *N** as a high-flow efficiency criteria because it measures how well the model can reproduce the variability of the observations. Since the errors are larger for high flows, *N** puts more weight on these parts of the hydrographs. *Ki** was considered a low-flow efficiency criteria because of the inverse transformation: the low-flow values become preponderant in the computation of *Ki**. To ease the interpretation, we transformed the criteria so that: (i) they will be bounded, (ii) the optimal value of 1 is also the maximum value possible, and (iii) the transformation does not impact the ranking of the performance between the catchments. The *C2M* transformation [*Mathevet et al.*, 2006] is used for the quadratic criteria (*N* and *Ki*). The transformation used on the biases takes the absolute value. In doing so, we lose the information on whether the model overestimates or underestimates the variability or the mean. Given the objective of the paper, which is to identify what affects model performance, we consider it equally bad for a model to underestimate or overestimate streamflow quantity or variability. Hereafter, all results presented will be on

Table 3. List of the Efficiency Criteria Used for Model Performance Evaluation[a]

Name and Reference	Formula	Hydrological Focus	Criteria Transformation	Notation
Nash-Sutcliffe efficiency [*Nash and Sutcliffe*, 1970]	$N=1-\frac{\sum(Qs-Qo)^2}{\sum(Qs-\overline{Qo})^2}$	High flow	*C2M*: $N*=\frac{N}{2-N}$	*N**
Kling-Gupta efficiency on inverse streamflow [*Gupta et al.*, 2009]	$Ki=1-[(1-R)^2+(1-Bm)^2+(1-Bd)^2)]^{0.5}$	Low flow	*C2M*: $Ki*=\frac{Ki}{2-Ki}$	*Ki**
Mean bias	$Bm=\frac{\overline{Qs}}{\overline{Qo}}$	Water balance	$Bm*=1-\lvert 1-Bm\rvert$	*Bm**
Deviation bias	$Bd=\frac{\sigma_{Qs}}{\sigma_{Qo}}$	Variability	$Bd*=1-\lvert 1-Bd\rvert$	*Bd**

[a]The observed streamflow is abbreviated *Qo* and the simulated streamflow *Qs*.

transformed values (noted *N**, *Ki**, *Bm**, and *Bd**). Since the criteria relate to different parts of the hydrograph, they are complementary and cross correlation is low (see Figure S3 of the supporting information). The highest correlation is 0.52 between *N** and *Bd**, because these two criteria are influenced by the highest flows.

3.3. Catchment Features Impacts on Model Performance

The impact of one or several feature(s) on model's performance is assessed by analyzing the model performance during the validation period (1992–2002).

3.3.1. One-Dimensional Analysis

The motivation for this one-dimensional analysis is to assess the impact of each feature taken independently, and better understand their relation to model performance. Given that one feature was considered at a time, correlations between features do not impact the results: correlated features will only have a similar impact on performance. Feature's impact on model performance is assessed by a three-step procedure. The catchment set is first ranked by increasing feature values and divided into five classes composed of an equal number of catchments. Then, the Kruskal-Wallis nonparametric test [*Kruskal and Wallis*, 1952] is used to evaluate whether at least one class has a performance significantly different from the others. The impact of feature *x* on criteria *y* was considered significant if the *p* value returned by the test is lower that 10^{-3}. A justification of the choice of a 10^{-3} threshold over the 0.05 threshold commonly used in hydrological studies is provided in section 4.2 of the supporting information file. The third step of the analysis is to assess whether or not the performance varies monotonously with the feature. To assess this, we simply checked that the mean performance per class increased or decreased with the mean feature value per class. We refrained from using correlation tests (typically the Spearman test) for two reasons: (i) the test proved insufficient for large samples [*Prairie*, 1996] and (ii) we also wished to capture nonmonotonous behaviors.

3.3.2. Multidimensional Analysis

Regression trees were used to take into account features correlations and rank the relative impact of the features on each efficiency criterion. The aim of the analyses via tree-building algorithms is to predict dependent variables from a set of causal effects. Regression tree approaches perform successive binary splittings of a given data set (each efficiency criterion) according to decision variables (the features). The algorithm identifies the best possible predictors, starting from the most discriminating and proceeding to the least important. The optimal choices are determined recursively by increasing the homogeneity within the two resulting clusters. The decision variables are selected automatically by the algorithm among the 29 catchment features. The only constraint we imposed consists in having at least 100 catchments in each final cluster (leaf), to capture general trends. In this study, the regression trees are primarily used to understand what combination of features leads to high or low model performance, rather than to predict the level of efficiency one could expect for a type of catchment.

3.4. Added Value of Multinational Sets

The added value of multinational data sets is defined in terms of results transferability. To assess results transferability, we propose a calibration-validation experiment based on regression trees and different catchment sets. First, four catchment sets are defined according to the location of the catchments: (i) the multinational set, (ii) the catchments located in Austria, (iii) the catchments located in France, and (iv) the catchments located in Germany. Each of these sets can serve as a calibration or a validation set. Therefore 16 combinations are possible for each efficiency criterion. Secondly, *MSE* values are computed for each of

the trees calibrated on set *a* and validated on set *b*: the lower the *MSE*, the more transferable the tree. Lastly a Student *t* test is performed to assess if the trees calibrated on the national sets have significantly different *MSE* values than the trees calibrated on the multinational set when validated on a given set.

4. Results and Discussion

4.1. Model Performance

Figure 3 compares the model performance for the model calibration and validation periods. The performance is assessed by the objective function used for calibration (Kling-Gupta efficiency on square-rooted streamflow).

Median Kling-Gupta efficiency on square-rooted streamflow (*Ks*) over the catchment set is 0.92 during calibration (1982–1992) and 0.88 during validation (1992–2002). In terms of nontransformed Nash-Sutcliffe efficiency (*N*), these values correspond to 0.81 and 0.76, respectively and to 0.69 and 0.65, respectively, in terms of *N**. During calibration, 92% of the catchments have a *Ks* higher than 0.85% and 0% of the catchments have a negative *Ks* value. During validation, 76% of the catchments have a *Ks* higher than 0.85 and only one catchment a negative *Ks* value ($Ks = -0.03$).

The mean performance obtained for this study can be compared to other large-sample studies. For example, *Pushpalatha et al*. [2011] found a mean *N** 0.63 using GR6J on 1000 French catchments, *Parajka et al*. [2007] found a median *N* of 0.71 using HBV on 320 catchments located in Austria. *Arheimer et al*. [2012] found a median value of 0.74 using the HYPE model over 318 Swedish catchments and *Newman et al*. [2015] found that 90% of the catchments had a $N \geq 0.55$ using the Sacramento Soil Moisture Accounting Model over 671 American catchments.

The structure of GR6J appears versatile enough to represent the variety of hydrological behaviors present in the catchment set and provides robust simulations during validation. However, the contrasted performance over the data set means that the model does lack robustness on some catchments, which is an expected outcome of such a large-scale study. The performance differences between the countries is discussed in the supporting information (section 5.2) based on the findings of the analysis.

4.2. One-Dimensional Analysis

All corresponding figures, as well as an overview of the one-dimensional analysis results, can be found in the supporting information (section 4). For the sake of clarity, we choose to discuss only the impact of the features that appeared important in both the one-dimensional and the multidimensional analysis.

4.2.1. Performance of High-Flow Simulation

According to the Kruskal and the monotonous link tests, the most important features to high-flow simulation are: catchment area (*A*), mean flashiness of precipitation (*FIPm*), variation of the flashiness of streamflow (*FIQcv*), mean fraction of solid precipitation (*Fs*), variation of precipitation (*Pcv*), and irregularity of streamflow (*Qir*). For high-flow modeling, information on hydrological data (precipitation and streamflow) seems to have more predictive power than physiographic catchment attributes (morphological, pedological features). This result is in agreement with other studies at the regional and national scale [*Uhlenbrook et al*., 2002]. We will focus here on the impact of *A*, *FIPm*, and *Qir*.

Figure 4 shows that area has a positive impact on high-flow simulations: the larger the catchment, the better the model performance. On a set of 459 Austrian catchments, *Merz et al*. [2009] showed that larger catchments generally have lesser specific flood magnitude. This result is widely found over different data sets, as illustrated by, for example, *Guse et al*. [2010] on 83 German catchments. In addition, when an intense localized precipitation event is missed by the rain gage network, the consequences are more severe in small catchments (e.g., alpine catchments) than in large catchments, in which the rest of the catchment can have a buffering effect on the total streamflow. In other words, larger catchments have a smoother behavior that is easier to reproduce by the model.

The flashiness of precipitation has a negative impact on high-flow simulations: the model performance is lower for catchments with highly variable precipitation inputs. Catchments generally have a low-pass behavior: precipitation is a high-frequency signal when streamflow is a low-frequency signal [*Sivapalan*, 2003]. From the model point of view, it means that the precipitation variability needs to be reduced before reproducing the streamflow variability. *Oudin et al*. [2005] showed on a large multinational set that soil

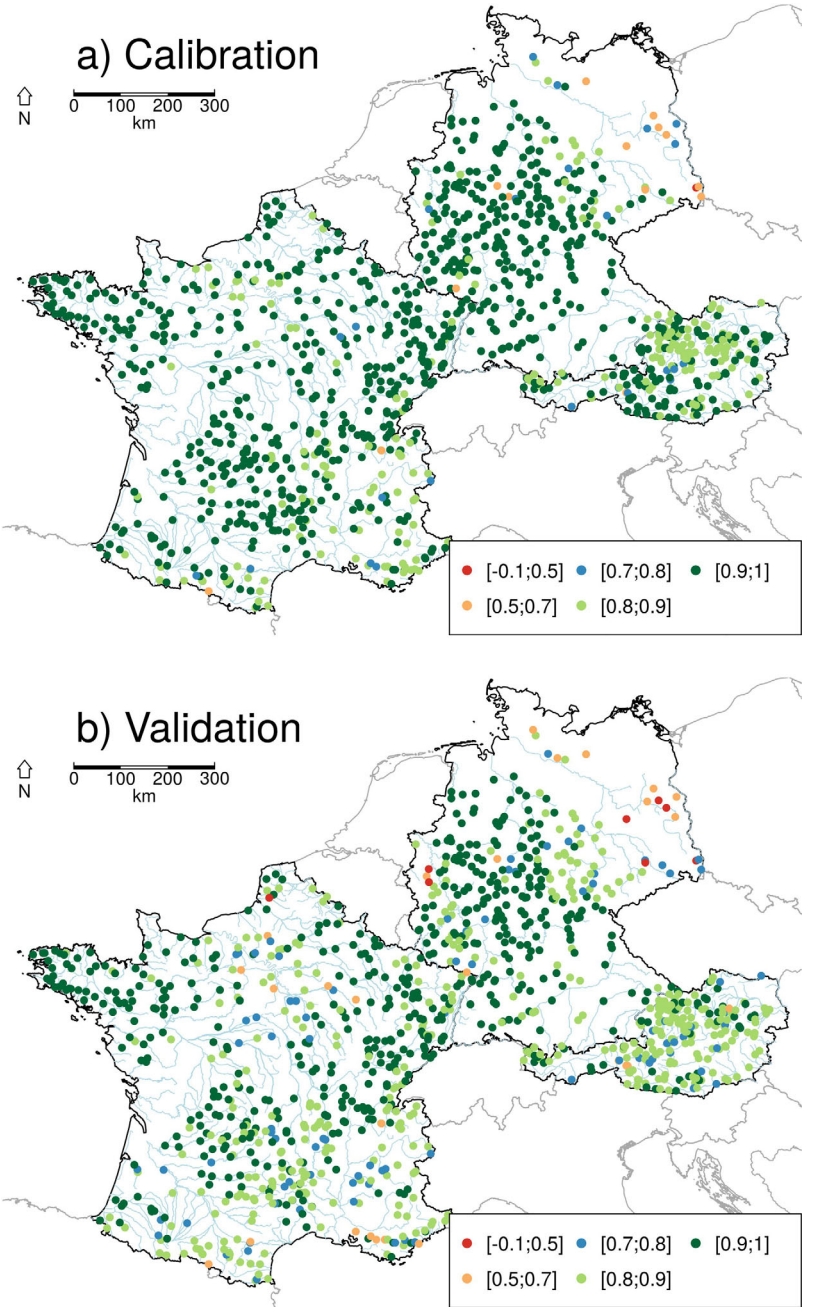

Figure 3. Model performance at the catchments identified by the location of their outlet during: (a) calibration (1982–1992) and (b) validation (1992–2002). The model performance for this figure is measured by the objective function used for calibration (Kling-Gupta efficiency on square-rooted streamflow).

moisture accounting models generally fail to smooth the rainfall input properly. In other words, the model struggles to reproduce the "natural" low-pass behavior of catchments. Therefore, model performance decreases as the precipitation variability increase.

Streamflow seasonality has a positive impact on high-flow simulations: the more seasonal, the higher the performance. A part of this behavior can be explained by the efficiency criteria selected for the high-flow analysis (N^*). Indeed, the Nash-Sutcliffe criterion shows mathematically higher values when streamflow is more seasonal, i.e., when the mean behavior is a poor benchmark [*Garrick et al.*, 1978; *Schaefli and Gupta*, 2007]. In addition, catchments with low streamflow seasonality do not have clearly defined wet and dry periods. Therefore, high flows can occur throughout the year and can be caused by a variety of processes.

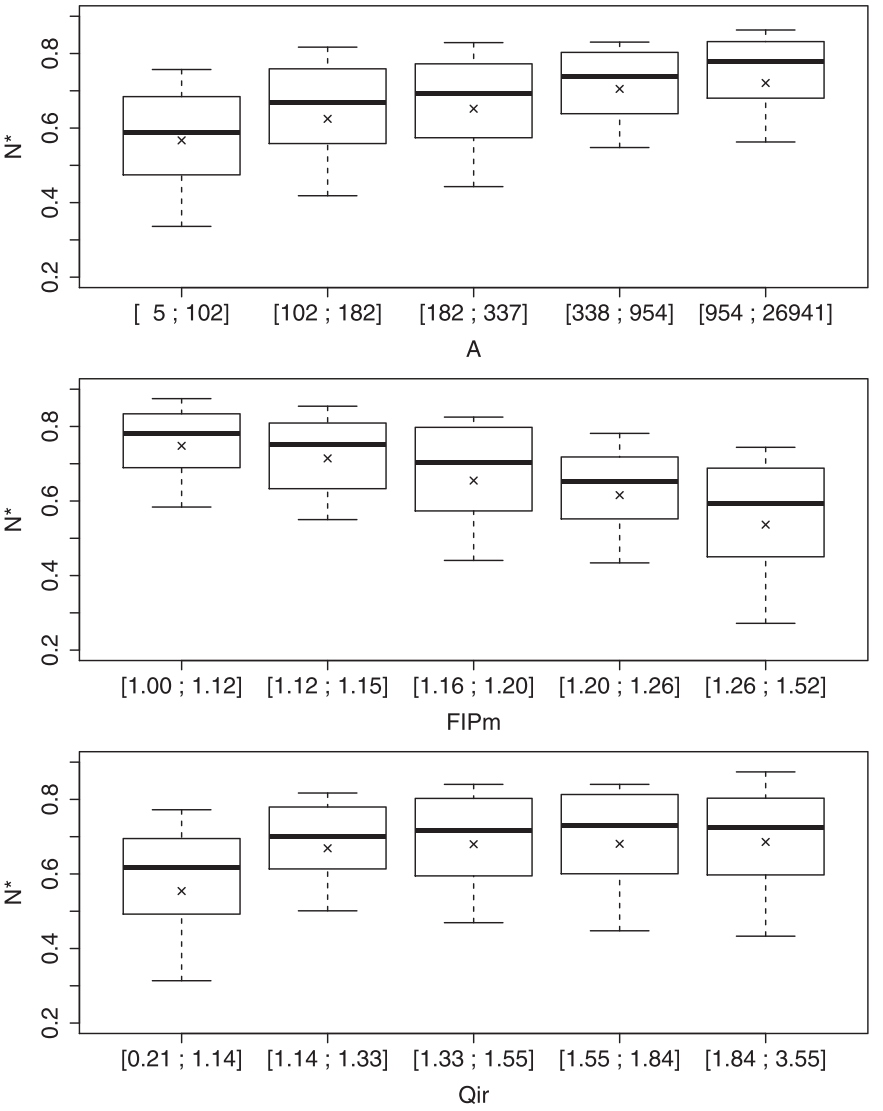

Figure 4. Impact of catchment area (*A*), mean flashiness of precipitation (*FIPm*), and the seasonality of streamflow (*Qir*) on high-flow simulations. According to the design of the one-dimensional analysis, the catchment set is ranked by increasing feature values and then divided into five classes composed of an equal number of catchments.

In particular, rainfall-driven high flows occur more often and are more variable than, for example, snowmelt high flows found in some seasonal catchments [*Merz and Blöschl*, 2003]. Because different processes lead to high-flow generation in nonseasonal catchment they are less predictable and model performance decreases.

4.2.2. Performance of Low-Flow Simulation

According to the Kruskal and monotonous link tests, the features that are most important to low-flow simulation are: catchment area (*A*), mean flashiness of precipitation (*FIPm*) and variation of precipitation (*Pcv*), the latter two being highly correlated. We will focus here on the impact of *A* and *Pcv*.

Figure 5 shows that area has a positive impact on model performance: low-flow simulations are improved on large catchments. As shown, in particular, by *Gupta et al*. [2009] models generally underestimate flow variability. In the case of low-flow simulation, this leads to overestimated flow values. The low-flow overestimation is expected to be emphasized in the case of pronounced low flows and attenuated when the low flows are sustained. Larger catchments are usually located in the lowlands where aquifers are more likely to sustain rivers during the low-flow period. The low flows being sustained, the model overestimation is less important on large catchments and model performance increase.

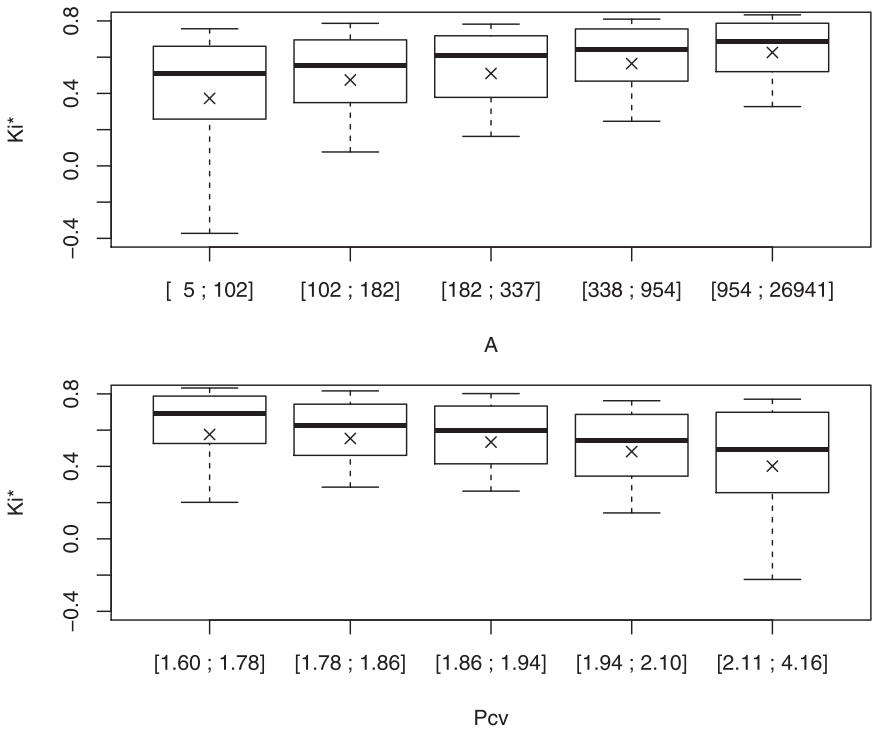

Figure 5. Impact of catchment area (*A*) and variation of precipitation (*Pcv*) on low-flow simulation. According to the design of the one-dimensional analysis, the catchment set is ranked by increasing feature values and then divided into five classes composed of an equal number of catchments.

The coefficient of variation of precipitation (*Pcv*) describes the daily variability of precipitation inputs and has a negative impact on model performance. First, it should be noted that the correlation coefficient between *Pcv* and the aridity index (*AI*) is 0.42. In other words, the catchments with variable precipitation are also among the driest. The negative effect of *Pcv* on model performance might, in part, be related to the severity of the low-flow period. In addition, evapotranspiration and soil moisture dynamics are dominant drivers of low flow. These processes are generally difficult to model because of their complexity and insufficient measurements [*Trambauer et al*., 2013] and the GR6J model structure does not explicitly account for these processes. In the case of high *Pcv*, the model first has to cope with precipitation variability before simulating complex processes. As a result, catchment moisture state is poorly estimated, which results in low model performance.

4.2.3. Performance of Water Balance Estimation

According to the Kruskal and monotonous link tests, the most important features to water balance estimations are: variability and irregularity of actual evapotranspiration (*AEcv*, *AEir*), mean flashiness of precipitation (*FIPm*), variation of precipitation (*Pcv*), and the catchment aridity index (*AI*). We will focus here on the impact of *AEir*, *FIPm*, and *AI*.

Figure 6 shows that irregularity of actual evapotranspiration (*AEir*) has a negative impact on model performance: the performance of water balance reproduction decreases when actual evapotranspiration is variable. In this case, the amount of water involved in the components of the water balance is variable, which is harder for the model to reproduce.

FIPm has a negative impact on water balance modeling: the model fails to estimate the water balance accurately for catchments with highly variable precipitation. As shown in section 4.2.1, variable precipitation damage high-flow estimations. Since most water quantities are produced during high flows, the water balance estimation is also degraded. Moreover, precipitation flashiness is correlated with the seasonality of precipitation ($R = 0.53$). Seasonally variable precipitation indicates that the amount of water involved in the components of the water balance is variable over time, a situation that is harder for the model to reproduce.

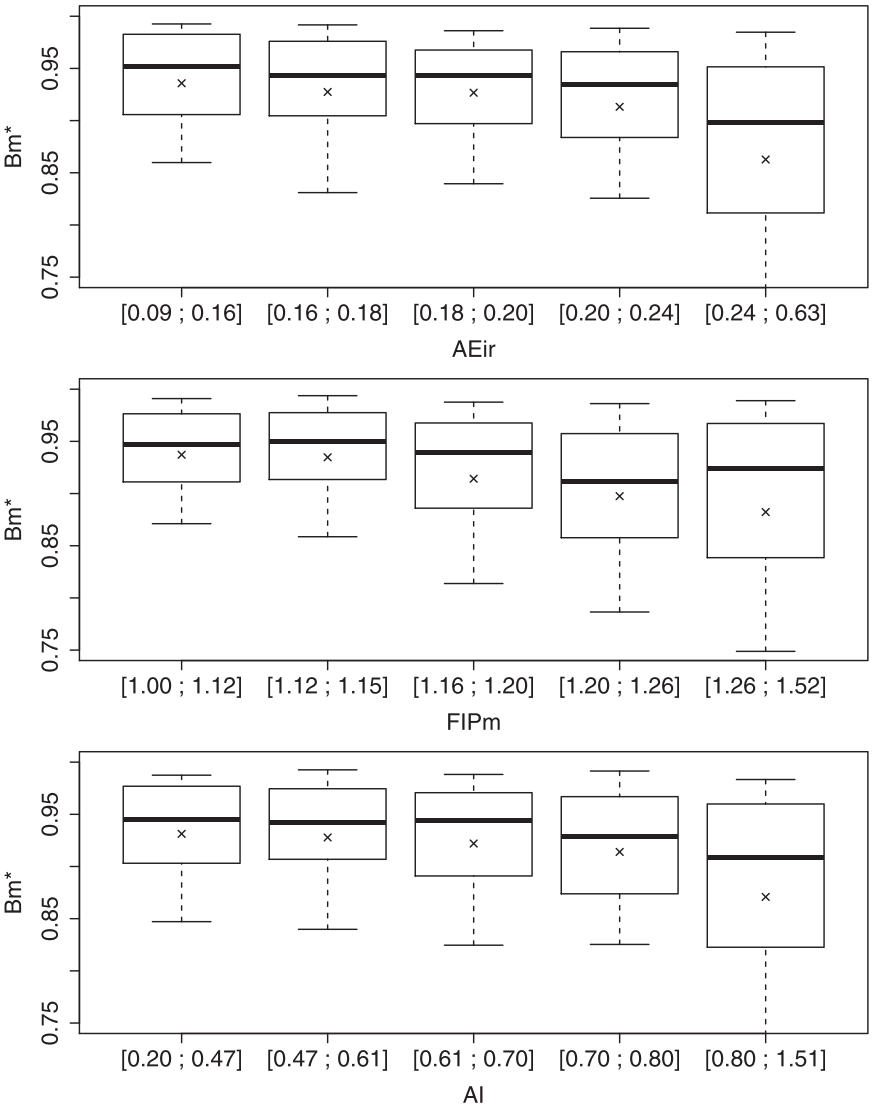

Figure 6. Impact of actual evapotranspiration irregularity (*AEir*), mean flashiness of precipitation (*FIPm*), and catchment aridity (*AI*) on water balance estimation. According to the design of the one-dimensionnal analysis, the catchment set is ranked by increasing feature values and then divided into five classes composed of an equal number of catchments.

Catchment aridity (*AI*) has a negative impact on model performance: water balance estimation decreases as catchment aridity increases. *Wang and Alimohammadi* [2012] studied the relations between water balance components and climate variability on 277 catchments located in the USA. Using the Budyko framework, they related catchment aridity with two states: water-limited catchments and energy-limited catchments. Their results showed that under energy-limited conditions, most of the precipitation anomaly is transferred to the runoff anomaly, but under water-limited conditions, most of the precipitation anomaly is transferred to storage change, and some of precipitation anomaly is transferred to the evapotranspiration anomaly. The catchments in our data set are mostly energy-limited, i.e., a small variation in precipitation leads to high streamflow anomalies. But when aridity increases, they become water-limited and the partitioning of precipitation into runoff, evaporation, and storage becomes more variable. Therefore, the water balance is more difficult for the model to reproduce. These results are complementary to those of *Merz and Blöschl* [2009], who showed that in a wet climate (energy-limited), catchments tend to be wet prior to most high-flow events and hence the runoff coefficients are, generally, high. In wet catchments, the impact of evapotranspiration and groundwater changes is reduced, making the water balance easier for the model to capture.

4.2.4. Performance of Streamflow Variability Estimation

According to the Kruskal and monotonous link tests, the features that are most important to streamflow variability estimations are: catchment area (A), mean flashiness of precipitation (*FIPm*) and variation of the flashiness of streamflow (*FIQcv*). All corresponding figures can be found in the supporting information.

Figure 7 shows that area has a positive impact on model performance (*Bd**): the larger the catchment the better the streamflow variability estimation. Larger catchments generally have lower streamflow variability [*Sivapalan*, 2003]. As a result, the streamflow variability is easier for the model to reproduce.

The mean flashiness of precipitation (*FIPm*) has a negative impact on streamflow variability estimations. As stated before, catchments generally have a low-pass behavior: streamflow variability is smaller compared to precipitation variability. This behavior is more difficult to reproduce when the precipitation variability is high. Therefore, streamflow variability estimation is degraded as *FIPm* values increase.

The variation in streamflow flashiness (*FIQcv*) has a negative impact on streamflow variability estimations. Models generally underestimate flow variability [*Gupta et al*., 2009] and fail to simulate sharp flow peaks

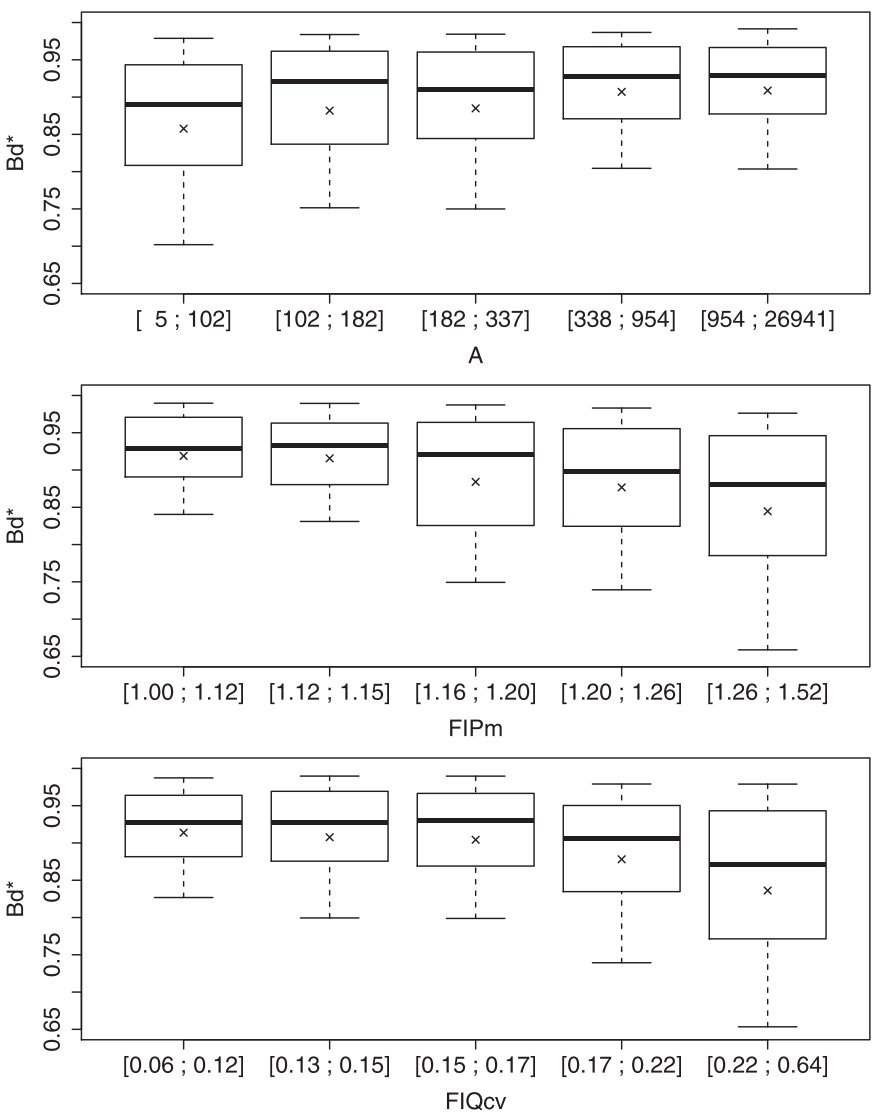

Figure 7. Impact of catchment area (*A*), mean flashiness of precipitation (*FIPm*), and variation of the flashiness of streamflow (*FIQcv*) on streamflow variability estimation (*Bd**). According to the design of the one-dimensional analysis, the catchment set is ranked by increasing feature values and then divided into five classes composed of an equal number of catchments.

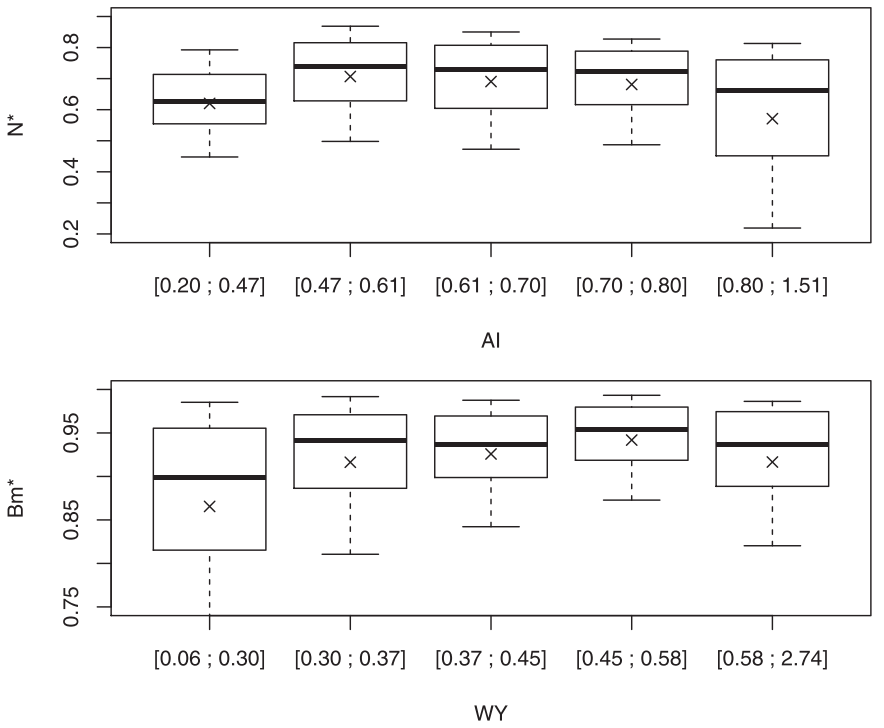

Figure 8. Impact of the aridity index (*AI*) on high-flow efficiency (*N**) and of water yield (*WY*) on water balance reproduction (*Bm**).

reliably [*van Esse et al*., 2013]. These behaviors appear more strongly when the observed variability of streamflow is high.

4.2.5. Nonmonotonous Behaviors: The Case of Aridity Index and Water Yield

Aridity index (*AI*) and water yield (*WY*) returned positive for the Kruskal-Wallis test, but the evolution of model performance between classes was not monotonous. Figure 8 shows how they impact performance for high flows and water balance estimations (respectively, *N** and *Bm**).

Figure 8 shows that model performance decreases for an aridity index between 0.47 and 1.51. The first remark is that the more arid the catchment the lower the model's performance, which is related to the weaker correlation between precipitation and streamflow. For the lowest class of aridity however, performance increases with aridity. This class is mostly composed of mountainous catchments, indicating that the decrease in performance could be linked to the greater uncertainties associated with rainfall inputs in these cases [*Gottardi et al*., 2012].

Performance for water balance estimations increases for a water yield between 0.06 and 0.58 and decreases for higher values. Evapotranspiration and groundwater changes impact on streamflow is reduced in wet catchments, which improves model performance. For the last yield class, where performance decreases, 62% of the catchments have *WY* higher than 0.7 and 15% *WY* higher than 1. These catchments either have a problem with the input data (precipitation might be underestimated) or are receiving underground water from outside the catchment, a situation that is difficult for a catchment model to simulate.

4.3. Multidimensional Analysis

4.3.1. Explanatory Power of the Regression Trees

In this section, we focus on assessing the explanatory power of the features for each criterion. The mean square error (*MSE*) is used to measure the quality of the regression-tree model: the lower the *MSE*, the better the trees explain model performance. Given that the criteria are normalized, MSE values can be directly compared for the four criteria.

Table 4 shows that the trees have similar degrees of complexity: simple (four leaves) for *Ki** to more complex (seven leaves) for *N**. This is due to the constraint of at least 100 catchments per leaf: for some criteria it is not possible to decipher general trends and hence the tree structure is simpler. However, the more

Table 4. Summary of the Regression Trees Complexity and Performance, Measured by the Number of Leaves and the Mean Square Error (*MSE*), Respectively

	1	*N**	*Ki**	*Bm**	*Bd**
Number of leaves	7	4	5	5	
MSE	0.025	0.082	0.006	0.010	

complex trees are not necessarily the ones that perform better; for example, *N** (seven leaves) has a *MSE* of 0.025, whereas *Bm** (five leaves) has a *MSE* of 0.006. Regarding the weak relationships found for *Ki**, we might hypothesize that for low-flow criteria, catchment features might be second-order drivers. Indeed, measurement errors occur more often during low flows [*van Esse et al*., 2013] both due to hydraulic sensitivity of the rating curves and proportionally larger human influences on low flows. Our results suggest that data quality is indeed a first-order factor, decreasing the explanatory power of the catchment features.

4.3.2. Hydrological Interpretation of the Regression Trees

In this section, we focus on ranking the features' importance and understanding how their interaction impacts model performance. The trees for the four criteria are shown in Figure 9.

Figure 9 shows that model performance in high flows (*N**) is mainly influenced by the flashiness of precipitation: the higher the flashiness the lower the performance. This result is in agreement with the one-dimensional analysis: the model fails to reproduce variable behaviors. This effect is attenuated by a larger amount of precipitation ($Pm \geq 2.1$ mm/d, *quantile* 15th), larger catchment area ($A \geq 115$ km^2, 25th) and less forest coverage ($pF \leq 46\%$, 55th). Larger *Pm* and *A* tend to smooth the hydrological response of the catchment, making it easier for the model to reproduce. On the other hand, large forest cover is equivalent to a rough land cover that tends to increase the time of concentration within a catchment [*Samaniego and Bárdossy*, 2007]. Hence, correlation between precipitation and streamflow is reduced, which is harder for the model to deal with. In the case of low precipitation flashiness, model performance is affected by a higher fraction of solid precipitation ($Fs \geq 16\%$, 80th) and more pronounced seasonal streamflow ($Qir \geq 1.2$, 25th). These results appear in contradiction with the results of the one-dimensional analysis where performance increased with seasonality due to an improved predictability of streamflow on these catchments. The correlations between *FIPm*, *Fs*, and *Qir* are weak, so it is unlikely that the change of behavior is due to the inclusion of *FIPm* prior to *Fs* and *Qir*. *Fs* is directly linked with catchment elevation, so that the decrease in model performance is probably related to larger uncertainties on the rainfall inputs in mountainous catchments. Streamflow seasonality decreases model performance for catchments with low variability in precipitation and low fraction of solid precipitation, i.e., for catchments where there is no seasonality or variability in the precipitation inputs. In these cases, the model has to simulate a variable output from nonvariable inputs. It is likely that for such catchments, streamflow variability is caused by groundwater and evapotranspiration dynamics, which are hard to deal with for the model.

Model performance during low flows (*Ki**) is mainly influenced by the water yield: the model performs worse for productive catchments ($WY \geq 0.76$, 90th). Catchment productivity is mostly determined by its high flows since streamflow volumes are much greater for this phase than for low flows. Hence, catchments with high water yield values are characterized by productive high flows and large variability of low-flow features because they are water-limited during the low-flow period [*Wang and Alimohammadi*, 2012]. In other words, if productive catchments are easier to model during high flows, the impact of evapotranspiration and ground-water changes during low flows becomes dominant and is not well reproduced by the model. The effect of water yield is boosted by higher aridity ($AI \geq 0.88$, 90th) and more variable streamflow ($Qcv \geq 1.5$, 80th). The one-dimensional analysis showed that model performance decreases with low-flow severity. Since low flow is less sustained in arid catchments, the performance decreases [*Newman et al*., 2015]. Finally, streamflow variability is more important for the low flows because the model has to reproduce a more variable behavior when there is not much precipitation.

Model performance for water balance reproduction (*Bm**) is mainly influenced by the seasonality of actual evapotranspiration: the model performs worse for seasonal catchments. This result is in line with the one-dimensional analysis and is supported by the results of *Merz and Blöschl* [2009] and *Wang and Alimohammadi* [2012]. Indeed, a seasonal evapotranspiration indicates that the amount of water involved in the components of the balance is variable over time, which is difficult for the model to deal with. This effect is

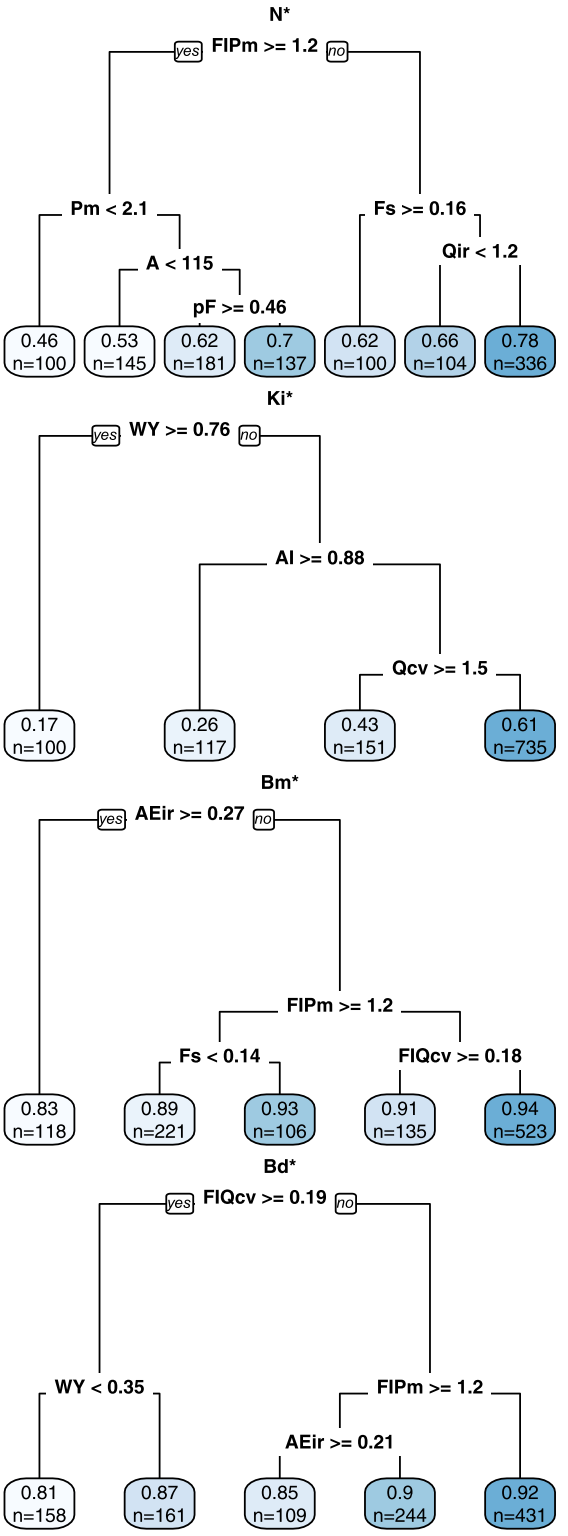

Figure 9. Regression trees calibrated on the whole data set for *K**, *Ki**, *Ns**, and *Bm**. The tree leaves gather information on (i) the mean performance in each leaf and (ii) the number of catchments in the leaf (*n*).

boosted when the seasonality/variability of the other components of the water balance is high: high-precipitation variability (*FIPm* $\geq$ 1.2, 60th, *Fs* $<$ 0.14, 75th) and high streamflow variability (*FIQcv* $\geq$ 0.18, 55th).

Model performance for streamflow variability estimation (*Bd**) is mainly influenced by the variability of streamflow: the model performs worse for catchments with variable streamflow (*FIQcv* $\geq$ 0.19, 70th). Models generally underestimate streamflow variability and this behavior is more pronounced for catchments where the observed variability is high. This effect is attenuated for productive catchments (*WY* $\geq$ 0.35, 35th) and boosted by highly variable precipitation (*FIPm* $\geq$ 1.2, 60th) and evapotranspiration (*AEir* $\geq$ 0.21, 65th), which is in line with the one-dimensional analysis findings. *Bd** is mostly degraded by poor high-flows simulations that occurred preferably in nonproductive catchments. For catchments with variable climatic forcing (high *FIPm* and *AEir*), streamflow variability depends on a variety processes, making it more difficult for the model to predict.

4.4. Are Results Transferability Improved by a Multinational Experiment?

In this section, we are interested in the added value of multinational data sets in terms of results transferability. In other words, we assess if for a given line in Figure 10 the *MSE* in each column is significantly different from the column "*All*." The regression trees calibrated at the national scale as well as an analysis of the model performance per country are gathered in the supporting information (section 5.1).

Figure 10 shows that for all criteria, the MSE values are lower on the diagonal (from bottom left to top right). This is an expected outcome since in this case the trees are calibrated and validated on the same catchment set.

For the high-flow criterion (*N**), the trees calibrated over the Austrian catchments showed lower performance when validated on the other sets. Austrian high flows are mostly snowmelt-fed [*Merz and Blöschl*, 2003] where temperature is the main driver. This is not the case for France and Germany where precipitation is the main driver. The different high-flow driver for Austria causes poor result transferability for high flow (*N**) as well as for the variability bias (*Bd**).

For the low-flow criterion (*Ki**), the trees calibrated over the Austrian and German catchments showed lower performance when validated on the other sets. Austrian catchments mostly have winter low flows, related

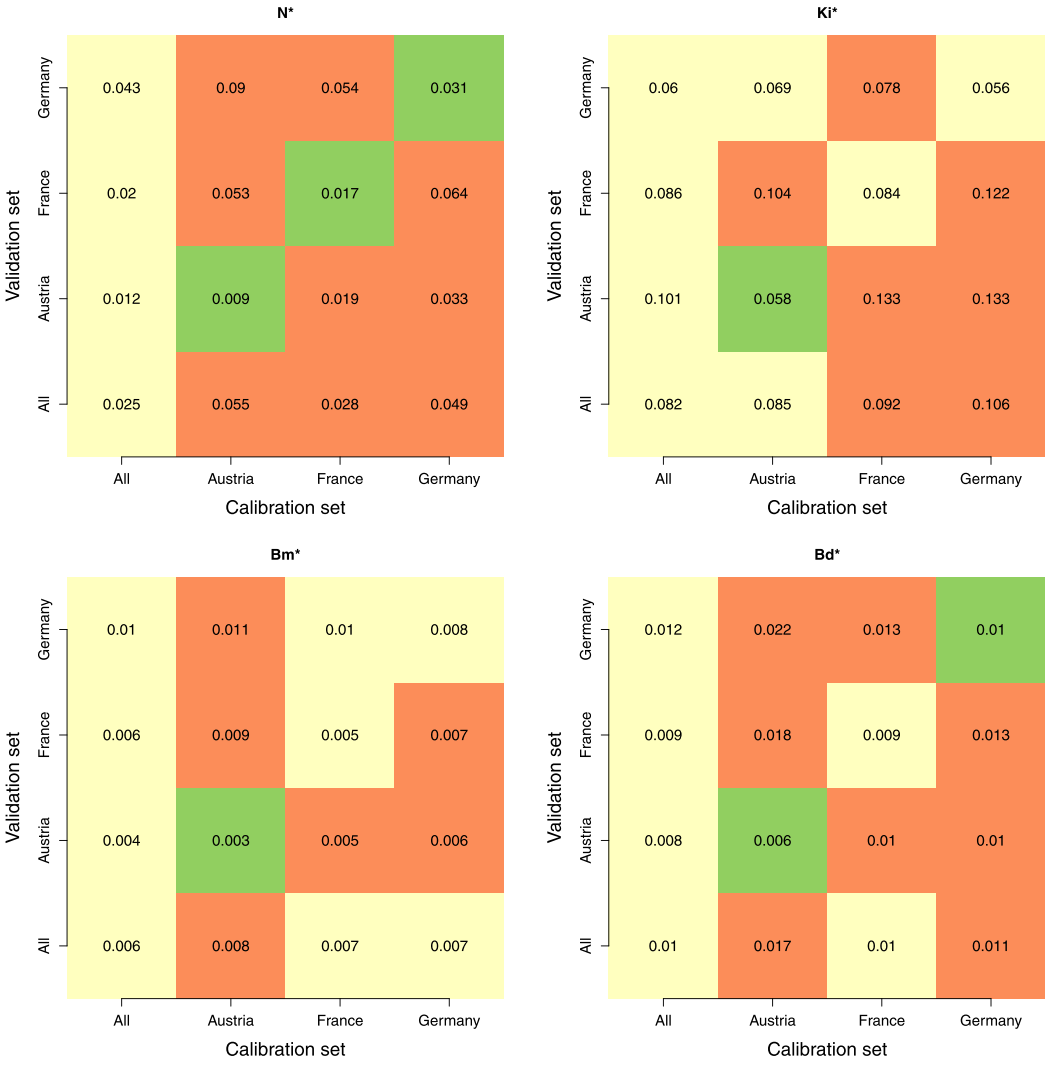

Figure 10. Performance of the regression tree (*MSE*) for the different calibration-validation setups: when the trees are calibrated or validated on catchments located in all countries, Austria, France and Germany. The numbers on the plot are the *MSE* values. The color code presents the result of the *t*-test: red, yellow and green indicates that the performance is significantly worse, equivalent and significantly better, respectively, than the tree calibrated on the whole set.

to the storage of precipitation in the snowpack. On the other hand, French and German catchments mostly have summer low flows, related to the long-term groundwater dynamics and evaporation. German catchments are larger in this data set: the mean catchment area is 1900 km^2 for Germany, 420 km^2 for Austria and 710 km^2 for France. Larger catchments are more likely to have aquifers sustaining rivers during the low-flow period and hence the tree calibrated on the German set is less transferable to Austria and France. The transferability of the trees describing low-flow behavior appears smaller than the transferability for high flows. These results suggest that the low flows drivers are more catchment-specific than the high-flow drivers.

Regarding the mean bias (*Bm**), the trees validated on Germany seems to lack robustness and in addition, even the tree calibrated on Germany does not seem sufficient to explain much of the mean bias. For this study, we did not use geological features but it is likely that for the German catchments, groundwater dynamics are a significant part of the water balance. Since we did not use geological features, the groundwater dynamics are not directly described and hence the explanatory power of the tree is weak for the German catchment set.

The contrasted transferability of trees is due to different streamflow generation drivers between the three countries. As a result, the trees calibrated in a particular country can lack robustness when applied on a

country where main drivers for streamflow generation differ. Large processes variety can exist at the national scale, but multinational sets further increase the explored hydro-climatic conditions and provide stronger results transferability.

5. Conclusions and Perspectives

The aim of this study was to identify major catchment controls on daily runoff simulations. Larger variability and multiple processes involved in the catchment response (especially those related to soil moisture) decrease model performance. In particular, flashiness of both precipitation and streamflow, catchment area, catchment aridity, and seasonality of evaporation are the most significant explanatory features. The performance of the GR6J hydrological model:

Decreases with rainfall variability. The model has more difficulty handling variable precipitation. This is probably due to the larger associated uncertainty and to the increased difficulty of reducing the high frequency of precipitation to the lower frequency of streamflow.

Decreases with streamflow variability: The model has more difficulty reproducing flashier streams. The poorest simulations occur for catchments that are water-limited, i.e., when streamflow variability is due to evapotranspiration and groundwater dynamics.

Increases with catchment size. Larger catchments generally have a smoother behavior that is easier for the model to reproduce. Interestingly, this remains true even for a lumped model. Moreover, the input quantities (precipitation in particular) are known with less uncertainty on large catchments than on the small ones.

Decreases with catchment aridity. The more arid catchments have more nonlinear responses, which are harder for the model to handle because streamflow is less correlated with precipitation inputs and more driven by groundwater and evapotranspiration dynamics, which are poorly known.

Tests, not presented here, show that these results are preserved when the calibration and validation periods are exchanged. Naturally, we do recognize that a more diverse catchment data set would be welcome to confirm our findings, although the data set used in this paper is already quite large. Given the variety of catchments, features and efficiency criteria, we expect the above results to provide general insights into major catchment controls on model performance in Austria, France and Germany. In our opinion, two paths could be pursued for future work.

The first is an improvement of the features describing catchments' geology and pedology. Geological and pedological features are long recognized as significant towards catchments behavior [*Haberlandt et al*., 2001; *Viglione et al*., 2010; *Bouma et al*., 2011]. Quite surprisingly, pedological features were not identified as having a strong impact on model performance. Usually soil moisture plays a crucial role in catchment response because it impacts both the partition of rainfall between runoff and evapotranspiration as well as the transfer to the outlet. Soil moisture is a time and space variable characteristic. The features used in this study relate to a static, potential soil water content and only describe the spatial variability within a catchment. It is likely that the time variability of soil moisture would be more informative towards model performance, as for example soil moisture prior to a rain event or during the low-flow period [*Penna et al*., 2011] and the depth of the water table below the ground surface [*Bronstert et al*., 2012]. Though geological informations are available at the European scale [*Asch*, 2005], they were not used in this analysis because such information is often qualitative and complex to interpret. It is likely that improved geological and pedological features will supplement the results presented here.

The second is the investigation of improved modeling setups. Two fields of investigation could be pursued: (i) enhance model calibration strategies and (ii) improve model structures. The calibration procedure could be improved by using hydrological signatures such as flow duration curves [*Westerberg et al*., 2011], groundwater data [*Madsen*, 2003], soil moisture data [*Grayson et al*., 2002] or actual evapotranspiration data [*Guerschman et al*., 2009]. The same data could also be used in a data assimilation procedure [*Crow and Ryu*, 2009]. Data assimilation and enhanced model calibration might improve the ability of the model to deal with variability and complex processes.

However, the above studies underline that these techniques provide only small improvements over a simple calibration approach, suggesting that processes representation is limited by the model structure itself. Our results suggest that model structure could be improved at least for some catchments presenting specific hydroclimatic settings (arid context and high hydroclimatic variabilities). In that sense, the present study shed more light on the hydrological processes that would need improved representation in the model structure. Consequently, a natural perspective of our study would be to test alternative model structures in order to improve model simulations on these specific catchments. Further works are needed to determine whether adapting model structure to these specific catchments might alter the performance on the rest of the catchments. This poses the question of the genericity of the model structure. In this context, the analysis could be repeated for a model that would allow its structure to vary from one catchment to another. The model used herein (GR6J) belongs to the category of the "one-size-fits-all" models. However, alternative approaches have been developed to adapt the model structure to each catchment [*Clark et al*., 2011; *Fenicia et al*., 2011]. It would be instructive to see how flexible models are able to deal with difficult catchments and see how model performance is affected in the case of a model which structure/parameters explicitly account for catchment features.

Acknowledgments
The authors would like to acknowledge the Knowledge and Innovation Center (Climate KIC) for funding the first author. We also wish to thank the three reviewers and the associate editor for their constructive comments and detailed reviews. We acknowledge SCHAPI for providing streamflow data for France (http://www.hydro.eaufrance.fr/), and Météo France for providing the SAFRAN climate archive over France (https://donneespubliques.meteofrance.fr/) as well as the hydrographic service of Austria (HZB, ehyd.gv.at) for providing climatic and streamflow data for Austria. For providing the discharge data for Germany, we are grateful: Bavarian State Office of Environment (LfU), Baden-Wurttemberg Office of Environment, Measurements and Environmental Protection (LUBW), Brandenburg Office of Environment, Health and Consumer Protection (LUGV), Saxony State Office of Environment, Agriculture and Geology (SMUL), Saxony-Anhalt Office of Flood Protection and Water Management (LHW), Thuringen State Office of Environment and Geology (TLUG), Hessian Agency for the Environment and Geology (HLUG), Rhineland Palatinate Office of Environment, Water Management and the Factory Inspectorate (LUWG), Saarland Ministry for Environment and Consumer Protection (MUV), Office for Nature, Environment and Consumer Protection North Rhine-Westphalia (LANUV NRW), Lower Saxony Office for Water Management, Coast Protection and Nature Protection (NLWKN), Water and Shipping Management of the Fed. Rep. (WSV), prepared by the Federal Institute for Hydrology (BfG). German climatic data can be obtained from the German Weather Service (DWD; ftp://ftp-cdc.dwd.de/pub/CDC/). Soil data are provided by the European Soil DataBase (http://eusoils.jrc.ec.europa.eu/). Land use data are available as CORINE 2006 data set from European Environmental Agency (EEA; http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-2). Digital elevation model can be retrieved from Shuttle Radar Topography Mission (SRTM; http://www.cgiar-csi.org/data/srtm-90m-digital-elevation-database-v4-1).

References

Arheimer, B., J. Dahné, C. Donnelly, G. Lindström, and J. Strömqvist (2012), Water and nutrient simulations using the HYPE model for Sweden *vs*. the Baltic Sea basin—Influence of input-data quality and scale, *Hydrol. Res.*, *43*, 315–329, doi:10.2166/nh.2012.010.

Asch, K. (1993), *IGME 5000: 1:5 Million International Geological Map of Europe and Adjacent Areas*, BGR, Hannover.

Bergström, S. (1995), *Computer Models of Watershed Hydrology—The HBV Model*, Water Resourc. Publ., Highlands Ranch, Colo.

Bouma, J., P. Droogers, M. P. W. Sonneveld, C. J. Ritsema, J. E. Hunink, W. W. Immerzeel, and S. Kauffman (2011), Hydropedological insights when considering catchment classification, *Hydrol. Earth Syst. Sci.*, *15*, 1909–1919.

Bronstert, A., et al. (2012), Potentials and constraints of different types of soil moisture observations for flood simulations in headwater catchments, *Nat. Hazards*, *60*, 879–914.

Budyko, M. I. (1974), Climate and Life, Int. Geophys. Ser., vol. 1,8 Academic, New York.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrologic models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.

Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.

Clark, M. P., et al. (2016), Improving the theoretical underpinnings of process-based hydrologic models, *Water Resour. Res.*, *52*, 2350–2365, doi:10.1002/2015WR017910.

Crochemore, L., C. Perrin, V. Andréassian, U. Ehret, S. P. Seibert, S. Grimaldi, H. Gupta, and J.-E. Paturel (2015), Comparing expert judgement and numerical criteria for hydrograph evaluation, *Hydrol. Sci. J.*, *60*, 402–423, doi:10.1080/02626667.2014.903331.

Crow, W. T., and D. Ryu (2009), A new data assimilation approach for improving runoff prediction using remotely-sensed soil moisture retrievals, *Hydrol. Earth Syst. Sci.*, *13*, 1–16.

King, D., J. Daroussin, J. M. Hollis, M. Jamagne, R. J. A. Jones, C. Le Bas, L. Ngongo, A. J. Thomasson, L. Vanmechelen, and E. Van Ranst (1994), A geographical knowledge database on soil properties for environmental studies, Final report of EC Contract No. 3392004 Commission of the European Communities (DGXI), 50 pp.

Das, T., A. Bárdossy, E. Zehe, and Y. He (2008), Comparison of conceptual model performance using different representations of spatial variability, *J. Hydrol.*, *356*, 106–118, doi:10.1016/j.jhydrol.2008.04.008.

Duan, Q., et al. (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*, 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Edijatno, N. 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.

Efstratiadis, A., and D. Koutsoyiannis (2010), One decade of multi-objective calibration approaches in hydrological modelling: A review, *Hydrol. Sci. J.*, *55*(1), 58–78.

EEA (2007), CLC2006 Technical Guidelines, Publ. Off., Luxembourg.

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 Resour. Res.*, *47*, W11510, doi:10.1029/2010WR010174.

Finke, P., R. Hartwich, R. Dudal, J. Ibanez, M. Jamagne, D. King, L. Montanarella, and N. Yassoglou (2001), *Georeferenced Soil Database for Europe*, Eur. Soil Bur. Sci. Comm., Italy.

Görgen, K., et al. (2010), Assessment of Climate Change Impacts on Discharge in the River Rhine Basin, Results of the RheinBlick2050 project, CHR report, 1–23, 229 pp., Lelystad.

Garrick, M., C. Cunnane, and J. E. Nash (1978), A criterion of efficiency for rainfall-runoff models, *J. Hydrol.*, *36*, 375–381.

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

Grayson, R. B., G. Blöschl, A. W. Western, and T. A. McMahon (2002), Advances in the use of observed spatial patterns of catchment hydrological response, *Adv. Water Resour.*, *25*, 1313–1334.

Guerschman, J. P., A. I. J. M. Van Dijk, G. Mattersdorf, J. Beringer, L. B. Hutley, R. Leuning, R. C. Pipunic, and B. S. Sherman (2009), Scaling of potential evapotranspiration with MODIS data reproduces flux observations and catchment water balance observations across Australia, *J. Hydrol.*, *369*, 107–119, 2009.

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

Gupta, V. K., and S. Sorooshian (1985), The relationship between data and the precision of parameter estimates of hydrologic models, *J. Hydrol.*, *81*, 57–77.
Guse, B., T. Hofherr, and B. Merz (2010), Introducing empirical and probabilistic regional envelope curves into a mixed bounded distribution function, *Hydrol. Earth Syst. Sci.*, *14*, 2465–2478, doi:10.5194/hess-14-2465-2010.
Haberlandt, U., B. Klöcking, V. Krysanova, and A. Becker (2001), Regionalisation of the base flow index from dynamically simulated flow components: A case study in the Elbe River Basin, *J. Hydrol.*, *248*, 35–53.
Holko, L., J. Parajka, Z. Kostka, P. Škoda, and G. Blöschl (2011), Flashiness of mountain streams in Slovakia and Austria, *J. Hydrol.*, *405*, 392–401, doi:10.1016/j.jhydrol.2011.05.038.
Hrachowitz, M., et al. (2013), A decade of Predictions in Ungauged Basins (PUB)—A review, *Hydrol. Sci. J.*, *58*(6), 1198–1255, doi:10.1080/02626667.2013.803183.
Kruskal, W. H., and W. A. Wallis (1952), Use of ranks in one-criterion variance analysis, *J. Am. Stat. Assoc.*, *47*, 583–621, doi:10.1080/01621459.1952.10483441.
Kuzmin, V., D. J. Seo, and V. Koren (2008), Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search, *J. Hydrol.*, *353*, 109–128.
Lehner, B., and G. Grill (2013), Global river hydrology and network routing: Baseline data and new approaches to study the world's large river systems, *Hydrol. Processes*, *27*(15), 2171–2186.
Leleu, I., I. Tonnelier, R. Puechberty, P. Gouin, I. Viquendi, L. Cobos, A. Foray, M. Baillon, and P. O. Ndima (2014), Re-founding the national information system designed to manage and give access to hydrometric data, *La Houille Blanche*, *1*, 25–32.
L'hôte, Y., P. Chevallier, A. Coudrain, Y. Lejeune, and P. Etchevers (2005), Relationship between precipitation phase and air temperature: Comparison between the Bolivian Andes and the Swiss Alps, *Hydrol. Sci. J.*, *50*, 988–997, doi:10.1623/hysj.2005.50.6.989.
Liang, X., D. P. Lettenmaier, E. Wood, and S. J. Burges (1994), A simple hydrologically based model of land surface water and energy fluxes for GSMs, *J. Geophy. Res.*, *14*, 415–428.
Madsen, H. (2003), Parameter estimation in distributed hydrological catchment modeling using automatic calibration with multiple objectives, *Adv. Water Resour.*, *26*, 205–216.
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-AISH Publ.*, *307*, 211–219.
Merz, R., and G. Blöschl (2003), A process typology of regional floods, *Water Resour. Res.*, *39*(12), 1340, doi:10.1029/2002WR001952.
Merz, R., and G. Blöschl (2009), Process controls on the statistical flood moments - a data based analysis, *Hydrol. Processes*, *23*, 675–696, doi:10.1002/hyp.7168.
Merz, R., J. Parajka, and G. Blöschl (2009), Scale effects in conceptual hydrological modeling, *Water Resour. Res.*, *45*, W09405, doi:10.1029/2009WR007872.
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.
Mouelhi, S. (2003), Vers une chaîne cohérente de modèles pluie-débit conceptuels globaux aux pas de temps pluriannuel, annuel, mensuel et journalier, PhD thesis, ENGREF, Paris.
Nachtergaele, F., H. Van Velthuizen, L. Verelst, N. Batjes, K. Dijkshoorn, V. Van Engelen, G. Fischer, A. Jones, L. Montanarella, and M. Petri (2008), *Harmonized World Soil Database*, Food Agric. Organ. of the United Nations, Rome. [Available at http://www.fao.org/nr/Water/docs/Harm-World-Soil-DBv7cv.pdf 2008.]
Nash, J., and J. V. 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.
Newman, A. J., et al. (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, doi:10.5194/hess-19-209-2015.
Oudin, L., V. Andréassian, C. Perrin, and F. Anctil (2005), Locating the sources of low-pass behavior within rainfall-runoff models, *Water Resour. Res.*, *40*, W11101, doi:10.1029/2004WR003291.
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.
Oudin, L., C. Perrin, T. Mathevet, V. Andréassian, and C. Michel (2006), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *J. Hydrol.*, *320*, 62–83, doi:10.1016/j.jhydrol.2005.07.016.
Parajka, J., R. Merz, and G. Blöschl (2003), Estimation of daily potential evapotranspiration for regional water balance modeling in Austria, in *11th International Poster Day and Institute of Hydrology Open Day Transport of Water, Chemicals and Energy in the Soil-Crop-Canopy-Atmosphere System*, pp. 299–306, Slovak Acad. of Sci., Bratislava.
Parajka, J. G. Blöschl, and R. Merz (2007), Regional calibration of catchment models: Potential for ungauged catchments, *Water Resour. Res.*, *43*, W06406, doi:10.1029/2006WR005271.
Parajka, J., A. Viglione, M. Rogger, J. L. Salinas, M. Sivapalan, and G. Blöschl (2013), Comparative assessment of predictions in ungauged basins. Part 1: Runoff-hydrograph studies, *Hydrol. Earth Syst. Sci.*, *17*, 1783–1795, doi:10.5194/hess-17-1783-2013.
Pechlivanidis, I. G., B. M. Jackson, N. R. Mcintyre, and H. S. Wheater (2011), Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications, *Global NEST J.*, *13*, 193–214.
Penna, D., H. J. Tromp-van Meerveld, A. Gobbi, M. Borga, G. Dalla Fontana (2011), The influence of soil moisture on threshold runoff generation processes in an alpine headwater catchment, *Hydrol. Earth Syst. Sci.*, *15*, 689–702.
Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*, 275–289.
Prairie, Y. T. (1996), Evaluating the predictive power of regression models, *Can. J. Fish. Aquat. Sci.*, *53*, 490–492.
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*, 66–76, doi:10.1016/j.jhydrol.2011.09.034.
Rauthe, M., H. Steiner, U. Riediger, A. Mazurkiewicz, and A. Gratzki (2013), A Central European precipitation climatology—Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS), *Meteorol. Z.*, *22*, 235–256, doi:10.1127/0941-2948/2013/0436.
Ritter, A., and R. Muñoz-Carpena (2013), Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments, *J. Hydrol.*, *480*, 33–45.
Rodriguez, E., C. S. Morris, and J. E. Belz (2006), A global assessment of the SRTM performance, *Photogramm. Eng. Remote Sens.*, *72*, 249–260.

Samaniego, L., and A. Bárdossy (2007), Relating macroclimatic circulation patterns with characteristics of floods and droughts at the mesoscale, *J. Hydrol.*, *335*, 109–123, doi:10.1016/j.jhydrol.2006.11.004.

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

Sivapalan, M. (2003), Process complexity at hillslope scale, process simplicity at the watershed scale: Is there a connection?, *Hydrol. Processes*, *17*, 1037–1041, doi:10.1002/hyp.5109.

Trambauer, P., S. Maskey, H. Winsemius, M. Werner, and S. Uhlenbrook (2013), A review of continental scale hydrological models and their suitability for drought forecasting in (sub-Saharan) Africa, *Phys. Chem. Earth, Parts A/B/C*, *66*, 16–26, doi:10.1016/j.pce.2013.07.003.

Turc, L. (1954), Le bilan en eau des sols: Relation entre les précipitations, l'évaporation et l'écoulement, *Ann. Agron.*, *5*, 491–595.

Uhlenbrook, S., A. Steinbrich, D. Tetzlaff, and C. Leibundgut (2002), *Regional Analysis of the Generation of Extreme Floods*, vol. 274, pp. 243–250, IAHS Publ., Cape Town, South Africa.

Valéry, A., V. Andréassian, and C. Perrin (2014), "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*, 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.

Vidal, J.-P., E. Martin, L. Franchistéguy, M. Baillon, and J.-M. Soubeyroux (2010), A 50-year high-resolution atmospheric reanalysis over France with the Safran system, *Int. J. Climatol.*, *30*, 1627–1644, doi:10.1002/joc.2003.

Viglione, A., G. B. Giovanni Battista Chirico, J. Komma, R. Woods, M. Borga, and G. Blöschl (2010), Quantifying space-time dynamics of flood event types, *J. Hydrol.*, *394*(1–2), 213–229.

Wang, D., and N. Alimohammadi (2012), Responses of annual runoff, evaporation, and storage change to climate variability at the watershed scale, *Water Resour. Res.*, *48*, W05546, doi:10.1029/2011WR011444.

Westerberg, I. K., J.-L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C.-Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrol. Earth Syst. Sci.*, *15*, 2205–2227.