



HAL
open science

Dependence of model-based extreme flood estimation on the calibration period: the case study of the Kamp River (Austria)

Pierre Brigode, Emmanuel Paquet, Pietro Bernardara, Joël Gailhard, Federico Garavaglia, Pierre Ribstein, François Bourgin, Charles Perrin, Vazken Andréassian

► To cite this version:

Pierre Brigode, Emmanuel Paquet, Pietro Bernardara, Joël Gailhard, Federico Garavaglia, et al.. Dependence of model-based extreme flood estimation on the calibration period: the case study of the Kamp River (Austria). *Hydrological Sciences Journal*, 2015, Special issue: Modelling Temporally-variable Catchments, 60 (7-8), pp.1-14. 10.1080/02626667.2015.1006632 . hal-01164840

HAL Id: hal-01164840

<https://hal.science/hal-01164840>

Submitted on 17 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

4
5 **Dependence of model-based extreme flood estimation on the**
6 **calibration period: the case study of the Kamp River**
7 **(Austria)**
8

9 P. Brigode¹, E. Paquet², P. Bernardara³, J. Gailhard², F. Garavaglia², P. Ribstein⁴, F.
10 Bourgin¹, C. Perrin¹ & V. Andréassian¹.

11 ¹ *UR HBAN, Irstea, Antony, France*

12 ² *DMM, DTG, Électricité de France, Grenoble, France.*

13 ³ *LNHE, R&D, Électricité de France, Chatou, France.*

14 ⁴ *Sorbonne Universités, UPMC Univ. Paris 06, UMR 7619 METIS, Paris, France.*

15 pierre.brigode@irstea.fr

16
17
18 Received 20 January 2014; accepted 25 November 2014

19 Accepted author version posted online: 13 Jan 2015. Published online: 16 Jun 2015.

20 **Editor** Z.W. Kundzewicz; **Associate editor** A. Viglione.

21 **Abstract** The Kamp River is a particularly interesting case study for testing flood frequency estimation
22 methods, since it experienced a major flood in August 2002. Here, this catchment is studied in order to
23 quantify the influence of such a remarkable flood event on the calibration of a rainfall-runoff model, in
24 particular when it is used in a stochastic simulation method for flood estimation, by performing
25 numerous rainfall-runoff model calibrations (based on split-sample and bootstrap tests). The results
26 confirmed the usefulness of the multi-period and bootstrap testing schemes to identify the dependence
27 of model performance and flood estimates on the information contained in the calibration period. The
28 August 2002 event appears to play a dominating role for the Kamp River, since the presence or absence
29 of the event within the calibration sub-periods strongly influences the rainfall-runoff model calibration
30 and the extreme flood estimations that are based on the calibrated model.

31 **Key words:** Non stationarity; IAHS workshop; model calibration and evaluation; SCHADEX, extreme
32 floods, bootstrap.

33
34 **Dépendance des estimations de crues extrêmes (basées sur un modèle pluie-**
35 **débit) à la période de calage: étude de cas de la rivière Kamp (Autriche)**

36
37 **Résumé** La rivière Kamp est un cas d'étude particulièrement intéressant pour le test de méthodes de
38 prédétermination des crues, puisqu'elle a vu une crue exceptionnelle se produire en août 2002. Dans cet
39 article, nous étudions ce bassin versant pour quantifier l'influence de ce type de crue remarquable sur le
40 calage d'un modèle pluie-débit, en particulier lorsqu'il est utilisé dans une méthode de simulation
41 stochastique pour la prédétermination des crues. Pour cela, nous réalisons de nombreux calages du
42 modèle pluie-débit (en nous basant sur des tests de bootstrap et sur des périodes indépendantes). Les
43 résultats obtenus confirment l'utilité des procédures de calages multi-périodes et de « calages
44 bootstrap » pour identifier la dépendance des performances des modèles hydrologiques et des
45 estimations de crues extrêmes aux informations contenues dans les périodes de calage. L'événement de
46 2002 apparaît jouer un rôle dominant pour la rivière Kamp, puisque la présence de l'événement au sein
47 des périodes de calage influence fortement le calage du modèle pluie-débit et l'estimation des crues
48 extrêmes reposant sur le modèle calé. L'ensemble des jeux de paramètres obtenus avec des périodes de
49 calages ne contenant pas l'épisode de 2002 produit des estimations de crues extrêmes
50 systématiquement plus fortes que celles obtenues avec les autres jeux de paramètres.

51
52 **Mots clé :** Non stationnarité ; Atelier AISH ; calage de modèle et évaluation ; SCHADEX ; crues
53 extrêmes ; bootstrap.

54 **1. INTRODUCTION**

55
56 **1.1. The challenge of hydrological variability**

57
58 The calibration of rainfall-runoff models in the context of a changing climate is
59 currently the subject of an intense discussion in the hydrological modelling
60 community (e.g. by Peel & Blöschl, 2011; Muñoz *et al.*, 2013; Montanari *et al.*, 2013;
61 Hrachowitz *et al.*, 2013 and Thirel *et al.*, 2014). Indeed, observed hydro-
62 meteorological series (precipitation or streamflow for example) used for model
63 calibration are subject to significant variability over time (Milly *et al.*, 2008). This
64 variability could be induced by sudden physiographic changes in the catchment (e.g.
65 forest fire, dam building), climatic condition changes (e.g. air temperature rising)
66 and/or long-term fluctuations being barely detectable by statistical tests
67 (Koutsoyiannis, 2006; Montanari, 2012).

68
69 Hydrological variability challenges the usual calibration approach - traditionally
70 assuming stationary or at least representative hydro-climatological conditions - which
71 consists in using the entire record period for identifying one or several optimal
72 parameter sets. Several studies, based on the split-sample test proposed by Klemeš
73 (1986), investigated the sensitivity of rainfall-runoff simulations to the characteristics
74 of the calibration period (e.g. Donnelly-Makowecki & Moore, 1999; Seibert, 2003;
75 Vaze *et al.*, 2010; Merz *et al.*, 2011; Coron *et al.*, 2012; Brigode *et al.*, 2013a).
76 Gharari *et al.* (2013) recently suggested estimating calibration performance over
77 different sub-periods, in order to identify parameter sets with time-consistent
78 performance, thereby reducing the over-calibration problem (Andréassian *et al.*,
79 2012). Time-varying sensitivity analysis such as the DYNamic Identifiability
80 Analysis (DYNIA, Wagener *et al.* (2003)) have also been proposed to identify
81 “informative regions with respect to model parameters” (Wagener & Kollat, 2007)
82 and to link particular hydro-climatic conditions with time-varying dominant rainfall-
83 runoff model parameters (Herman *et al.*, 2013).

84
85 **1.2. The information content of extreme events**

86
87 The observed hydro-meteorological variability affects mean values as well as extreme
88 values. For instance, Ward *et al.* (2014) recently showed that El Niño Southern
89 Oscillations (ENSO) significantly influence the flood intensity of daily annual peak.
90 Interannual variability could also be characterized by the observation of outliers
91 within the record period, i.e. the “outlying observation that appears to deviate
92 markedly from other members of the sample in which it occurs” (Grubbs, 1969). Such
93 outstanding values have to be taken into account, since they provide valuable
94 information about the extreme hydrological behavior of the studied catchments (Laio
95 *et al.*, 2010). In a statistical framework, methods such as resampling techniques (Katz
96 *et al.*, 2002) can be used to quantify the sensitivity of the extreme-quantile estimation
97 to these observed rare events.

98
99 Nevertheless, in the context of rainfall-runoff model calibration, quantifying the
100 sensitivity of the model’s results to such rare events is more challenging. Berthet *et al.*
101 (2010) showed that only a limited number of time steps truly influences the values of
102 the quadratic calibration criteria usually used for rainfall-runoff model calibration
103 (like Nash & Sutcliffe (1970) Efficiency or Root-Mean-Square Deviation scores).

104 Moreover, Perrin *et al.* (2007) and Seibert & Beven (2009) highlighted that a limited
105 number of streamflow values can contain a significant amount of hydrological
106 information, while Beven & Westerberg (2011) suggested that some periods within
107 the observation records could even be *disinformative* for the models. Singh &
108 Bárdossy (2012) and Singh *et al.* (2012) suggested identifying a limited number of
109 events on which the calibration should be performed, using the statistical concept of
110 data depth.

111

112 The challenge of rainfall-runoff model calibration in a changing climate has been
113 recently studied in a workshop during the 2013 International Association of
114 Hydrological Sciences (IAHS) General Assembly in Göteborg, Sweden, where
115 hydrological modellers were asked to calibrate their models over several selected
116 catchments (Thirel *et al.*, 2014, this issue). Participants were provided a calibration
117 and evaluation protocol as well as a selection of 14 “changing catchments” showing
118 different observed changes such as temperature increases, dam building and land-
119 cover modification.

120

121 Among these 14 catchments, the Kamp River at Zwettl (622 km²) located in northern
122 Austria is a particularly interesting case study, since (i) a significant increase of more
123 than 1°C of the catchment’s air temperature has been estimated over the last 30 years
124 (Thirel *et al.*, 2014, this issue) and (ii) it experienced a major flood event in August
125 2002, which has been extensively studied over the last few years (e.g. Komma *et al.*,
126 2007; Viglione *et al.*, 2010; Viglione *et al.*, 2013). The August 2002 event, which
127 caused major flooding in different regions of central Europe (Blöschl *et al.*, 2013),
128 resulted in an estimated peak flow of 460 m³/s, which is three times higher than the
129 second largest flood observed over the 1951-2005 period (Viglione *et al.*, 2013). The
130 influence of this event has already been studied by Viglione *et al.* (2013) in the
131 context of flood frequency analysis, showing that this event strongly influences the
132 extreme flood estimation if no additional information (e.g. historical data) is used. On
133 this catchment, Brigode *et al.* (2014) also illustrated the strong influence of this event
134 on extreme rainfall estimation and on extreme flood estimation performed with a
135 stochastic flood simulation method.

136

137

1.3. Scope of the paper

138

139 This paper aims at (i) applying the calibration protocol proposed by the 2013
140 “hydrology under change” IAHS workshop within the context of extreme flood
141 estimation based on a rainfall-runoff model, (ii) comparing the results obtained using
142 the workshop calibration protocol to the one proposed by Brigode *et al.* (2014) based
143 on bootstrap resampling and (iii) quantifying the influence of the 2002 event on
144 rainfall-runoff model calibration. As in Brigode *et al.* (2014), the SCHADDEX method
145 (*Simulation Climato-Hydrologique pour l'Appréciation des Débits EXtrêmes - Hydro-*
146 *climatic simulation for the estimation of extreme flows*) detailed by Paquet *et al.*
147 (2013) has been applied over the Kamp catchment, considering sub-periods for the
148 calibration of the MORDOR rainfall-runoff model.

149

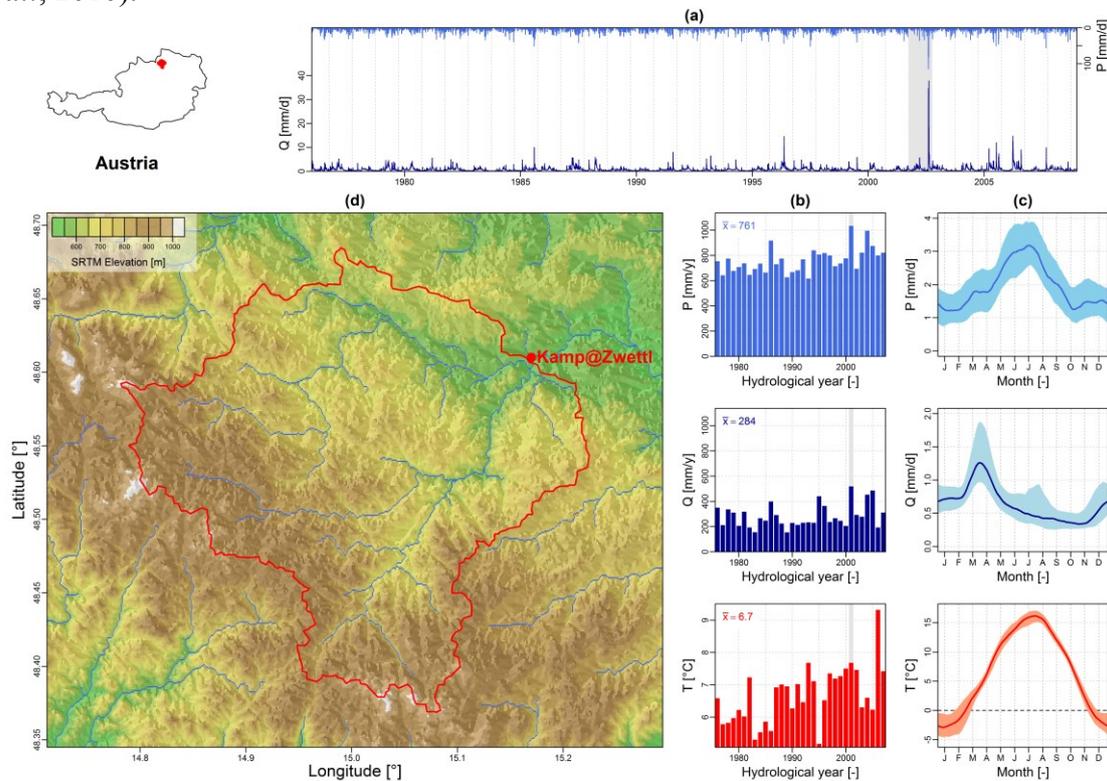
2. CATCHMENT DESCRIPTION

150

151 The Kamp catchment at Zwettl is one of the 14 “changing catchments” selected for
152 the IAHS workshop and described in detail by Thirel *et al.* (2014, this issue). Daily
153 precipitation, temperature and streamflow series have been supplied and are available
154 for the 1976-2008 period. Additionally, elevation data have been extracted from the
155 SRTM 90m data set (Jarvis *et al.*, 2008).

156

157 Figure 1 illustrates the hydroclimatic context of the Kamp catchment at Zwettl, a
158 622 km² catchment located in northern Austria. Catchment elevation ranges from
159 around 500 to 1,000 m a.s.l., with the highest elevation areas located in the southern
160 and western parts of the catchment. The daily streamflow and precipitation series
161 plotted in the upper part of Figure 1 clearly show that the August 2002 flood is not
162 comparable to other observed floods, both in terms of observed precipitation amount
163 and flood magnitude. Due to this extreme event, the 2001-2002 hydrological year has
164 the largest precipitation and runoff annual mean within the 1976-2008 period. On
165 average, the mean annual precipitation and runoff are around 800 mm and 300 mm,
166 respectively, on the Kamp catchment. Precipitation and runoff have clear seasonal
167 behaviours in this region, with the highest precipitation amount observed during
168 summer (typically from June to August) and the highest streamflow amount observed
169 during the March to April month, due to snowmelt. Large floods are usually observed
170 on this catchment during the July to August period, mainly produced by intense
171 rainfall events. Note that snow processes are important on this catchment, since floods
172 are also observed during rain-on-snow or snowmelt events in this region (Viglione *et al.*,
173 2010).



174

175

176

177

178

Fig. 1 Hydroclimatic context of the Kamp River at Zwettl catchment: (a) observed daily streamflow and precipitation time series, (b) streamflow, precipitation and temperature mean annual series, (c) streamflow, precipitation and temperature monthly regimes, and (d) SRTM elevation data (Jarvis *et al.* 2008).

179 **3. METHOD**

180
181 **3.1. The SCHADEX flood simulation method**

182
183 The SCHADEX method (Paquet *et al.*, 2013) is a stochastic flood simulation method
184 developed and applied by Électricité de France (EDF) for the design of dam
185 spillways. It has been applied to more than 80 catchments over France and elsewhere,
186 for example in Austria, Canada and Norway (Lawrence *et al.*, 2014; Brigode *et al.*,
187 2014).

188
189 SCHADEX is a semi-continuous stochastic flood simulation method that generates,
190 for a given catchment, a large number of floods, which result from the combination of
191 two hazards: (i) the rainfall hazard and (ii) the catchment saturation hazard:

- 192 i. Rainfall events are randomly drawn using a rainfall probabilistic model, the
193 Multi-Exponential Weather Pattern distribution (Garavaglia *et al.*, 2010). This
194 rainfall probabilistic model is based on a seasonal and weather pattern sub-
195 sampling of observed rainfall series. The five weather pattern classification
196 proposed for Austria by Brigode *et al.* (2013b) is used for the Kamp
197 catchment.
198 ii. Catchment saturation conditions are not explicitly described as a random
199 variable but are instead generated by a continuous rainfall-runoff simulation
200 over a long record period and thus implicitly represented by the internal
201 variables of the rainfall-runoff model.

202
203 The MORDOR rainfall-runoff model (Garçon, 1999; Andréassian *et al.*, 2006) is used
204 to perform the continuous rainfall-runoff simulation used for the description of the
205 catchment saturation conditions and also to transform a given rainfall event falling
206 over a given catchment into a flood event.

207
208 For each studied catchment, the SCHADEX simulation process generates around two
209 millions of simulated floods, resulting from the combinations of different rainfall
210 events with different catchment saturation conditions. A distribution of simulated
211 flood events is built to provide estimates of extreme flood quantiles, such as the
212 1,000-year return period flood (noted Q_{1000}).

213
214 In this study, only the MORDOR parameter set will change according to the
215 calibration periods. The entire record period will be considered for the estimation of
216 the rainfall probabilistic model parameters and for the computation of a modelled
217 distribution of catchment saturation conditions.

218
219 **3.2. The MORDOR rainfall-runoff model**

220
221 MORDOR is a conceptual rainfall-runoff model developed and intensively used by
222 EDF for operational hydrology in different contexts such as flood forecasting (e.g.
223 Zalachori *et al.*, 2012), low-flow forecasting (e.g. Mathevet *et al.*, 2010; Nicolle *et al.*,
224 2014) and flood frequency estimation (e.g. Paquet *et al.*, 2013). The different
225 components of the hydrological cycle are represented through four reservoirs within
226 MORDOR: (i) a rainfall excess/soil moisture accounting store (noted U) contributing
227 to actual evaporation and to direct runoff, (ii) an evaporating store (noted Z) filled by
228 part of the indirect runoff component and contributing to actual evaporation, (iii) an

229 intermediate store (noted L) determining the partitioning between direct runoff,
 230 indirect runoff and percolation to a deep storage reservoir and (iv) a deep storage
 231 reservoir (noted N) determining baseflow. Last, a unit hydrograph is used for routing
 232 the total simulated runoff.

233
 234 Required inputs of the MORDOR model are air temperature and precipitation series.
 235 With the snow component, MORDOR has 22 free parameters, while it has 11
 236 parameters without the snow component. In this study, the snow component
 237 parameters were fixed after a first MORDOR calibration over the entire record period,
 238 with an objective function combining classical Nash & Sutcliffe (1970) efficiency
 239 (NSE) with a criterion minimizing the difference between observed and simulated
 240 mean annual streamflow. This objective function aims at having snow component
 241 parameters inducing good day-to-day performance of the rainfall-runoff model and
 242 also good model performance in terms of simulated streamflow volume over the
 243 entire record period. The volume difference and NSE score obtained with this
 244 MORDOR parameter set over the 1977-2008 period are -0.1% and 0.85 respectively.
 245 Table 1 summarizes the name, role and unit of the 11 free parameters calibrated in
 246 this study and each parameter's prior ranges. These parameters are estimated using an
 247 automatic optimization scheme developed by EDF and based on a genetic algorithm,
 248 a strategy commonly used in hydrological modelling since the 1990s (e.g. Wang,
 249 1991; Franchini, 1996; Wang, 1997). It has been shown to perform as well as other
 250 algorithms such as the SCE-UA (Duan *et al.*, 1992) over numerous catchments by
 251 Mathevet (2005).

252
 253 **Table 1** Description of the 11 free parameters of the MORDOR model to be calibrated over the Kamp
 254 River at Zwettl catchment.

Name	Description (and unit)	Range
fe1	Parameter linked to potential evapotranspiration [-]	$0.0005 \leq X \leq 0.1$
fe3	Parameter linked to potential evapotranspiration [-]	$-7 \leq X \leq 0$
kl1	Percolation coefficient 1 of the L reservoir [-]	$0.1 \leq X \leq 0.9$
kl2	Percolation coefficient 2 of the L reservoir [-]	$0.1 \leq X \leq 0.9$
dn	Percolation coefficient of the N reservoir [-]	$1 \leq X \leq 999$
exn	Exponent of the recession limb of the N reservoir [-]	$1 \leq X \leq 8$
fr1	Parameter linked to the routing function [-]	$0.5 \leq X \leq 10$
fr2	Parameter linked to the routing function [-]	$0.5 \leq X \leq 6$
U_{MAX}	Maximum capacity of the U reservoir [mm]	$30 \leq X \leq 200$
L_{MAX}	Maximum capacity of the L reservoir [mm]	$30 \leq X \leq 200$
Z_{MAX}	Maximum capacity of the Z reservoir [mm]	$30 \leq X \leq 200$

255

256

3.3. MORDOR rainfall-runoff model calibration strategies

257

258 The objective function used for the calibration of the MORDOR rainfall-runoff model
 259 (noted OBJ_{EDF} and given in Equation 1) is a combination of two NSE scores: (i) the
 260 NSE score computed with observed and simulated streamflow time series and (ii) the
 261 NSE score computed with observed and simulated cumulative distribution functions
 262 of streamflow series (noted NSE_{CDF}). This combination allows a good trade-off
 263 between the day-to-day performance of the rainfall-runoff model and the model
 264 performance regarding the highest observed streamflow values (Paquet *et al.*, 2013).
 265 It has been recommended within the context of continuous flood simulation (Lamb,
 266 1999). This objective function is commonly used for the calibration of the MORDOR
 267 rainfall-runoff model within the SCHADEX method applications (e.g. Paquet *et al.*,
 268 2013; Lawrence *et al.*, 2014; Brigode *et al.*, 2014). For each MORDOR calibration,

269 only the optimal parameter set in terms of the OBJ_{EDF} objective is considered further.
270 Note that a perfect streamflow simulation has an OBJ_{EDF} value of 0.
271

$$OBJ_{EDF} = (1 - NSE)^2 + 2 * (1 - NSE_{CDF}) \quad (1)$$

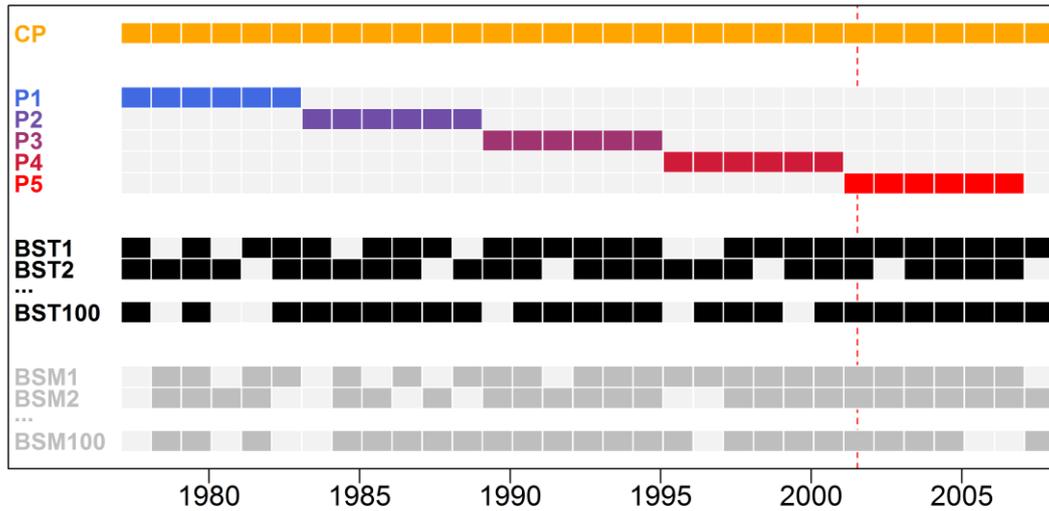
272

273 Three different calibration options were considered, as illustrated in Figure 2:

- 274 i. The MORDOR model was calibrated over the entire record period, leading to
275 a parameter set considered as the reference set (noted CP for complete period
276 in the following).
- 277 ii. MORDOR was calibrated over the five 6-year sub-periods selected in the
278 workshop test protocol (Thirel *et al.*, 2014, this issue). These five parameter
279 sets are noted P1 to P5 parameter sets in the following. The August 2002 flood
280 event is included in the P5 sub-period.
- 281 iii. One hundred independent calibrations (noted BST1, BST2, ..., BST100) were
282 made over 25-year sub-periods, identified through a block-bootstrap method
283 described by Brigode *et al.* (2014). For each of the 100 calibrations, 25
284 hydrological years are identified among the total hydrological years available
285 on the studied catchment (here 31 hydrological years, starting from 1976 and
286 ending in 2008). Note that the number of 25-year combinations from a given
287 set of 31 elements is 736,281 and thus that the 100 combinations tested here
288 are only a sub-set of all the possible combinations. Unlike the P1 to P5 sub-
289 periods, the bootstrapped sub-periods are not independent and are quite
290 similar: they have a majority of the 31 hydrological years observed in common
291 and only differ by the absence/presence of a few years. The rainfall-runoff
292 model is continuously run over the entire record period, but only the selected
293 hydrological years are considered for the computation of the objective
294 function. Note that bootstrap techniques have already been used for the
295 estimation of hydrological parameter uncertainty (Ebtehaj *et al.*, 2010; Selle &
296 Hannah, 2010).
- 297 iv. To highlight the influence of the August 2002 flood on the rainfall-runoff
298 model calibration, the same bootstrap scheme has been repeated but the
299 August 2002 data were systematically excluded from the computation of the
300 objective function. Thus, even if the 2002 hydrological year is selected, the
301 August 2002 month will not be considered for the calibration of the rainfall-
302 runoff model. These parameter sets will be noted BSM1, BSM2, ..., BSM100.

303

304 After each calibration, the MORDOR model was run on the whole period to enable
305 efficiency calculations on all test sub-periods.
306



307
 308
 309
 310
 311
 312
 313

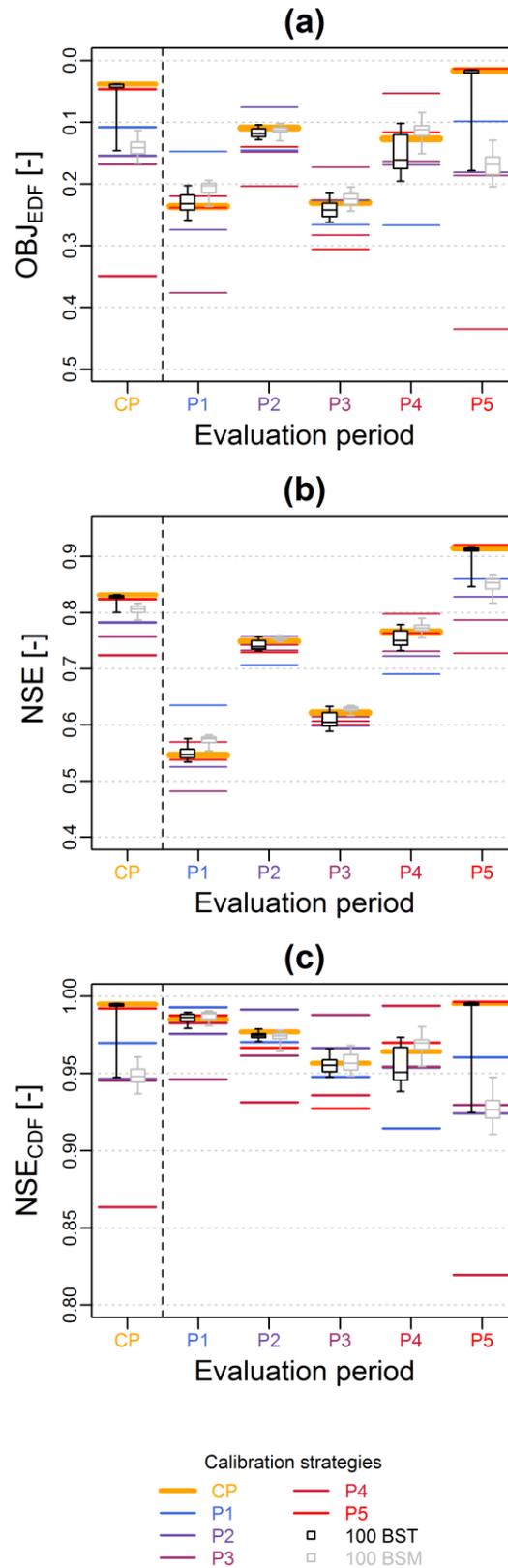
Fig. 2 Illustration of the different calibration strategies used for the calibration of the MORDOR rainfall-runoff model: calibration over the entire record period (CP), calibrations over the five 6-year sub-periods (P1–P5), calibration over 100 25-year sub-periods generated through a block-bootstrap technique (BST1 to BST100) and calibration over 100 25-year sub-periods excluding the August 2002 observations, generated through a block-bootstrap technique (BSM1 to BSM100). The vertical dotted line indicates the 2002 flood event.

4. RESULTS

4.1. Performance of the MORDOR rainfall-runoff model

Figure 3 presents a summary of the MORDOR performance obtained over the complete record period and over the five 6-year sub-periods, P1 to P5, both in terms of OBJ_{EDF} (the calibration criterion, top panel), NSE (middle panel) and NSE_{CDF} (bottom panel), and considering the different calibration options described above. MORDOR performance is generally good over the complete period, with the NSE score greater than 0.7. However, there is also a substantial variability of NSE scores over the different 6-year sub-periods, with several NSE scores below 0.6.

In terms of OBJ_{EDF} , calibration performance is the poorest for the P3 (1989-1995) sub-period. The performance range obtained with bootstrap calibrations is rather narrow and median performance is generally similar to the CP parameter set performance. This is related to the similarity of the 25-year sub-periods as well as the similarity between the CP period and BS sub-periods, only differing by the presence and/or absence of several hydrological years. BSM calibrations (grey boxplots) generally perform better than the BST for the P1 to P4 sub-periods while they perform less well for the complete period (CP) and the P5 sub-period (which includes 2002), regarding the three different scores. Interestingly, a similar ranking of model performance is obtained for the complete period and the P5 evaluation periods, which are the two periods containing the August 2002 event. For these periods, the CP parameter set is the best parameter set, followed by the P5 parameter set. The P1 to P4 parameter sets perform poorly for P5 compared to the other ones, especially for the NSE_{CDF} score. Finally, the BST calibrations (black boxplots) generally performed better than the BSM calibrations (excluding the August 2002 month).



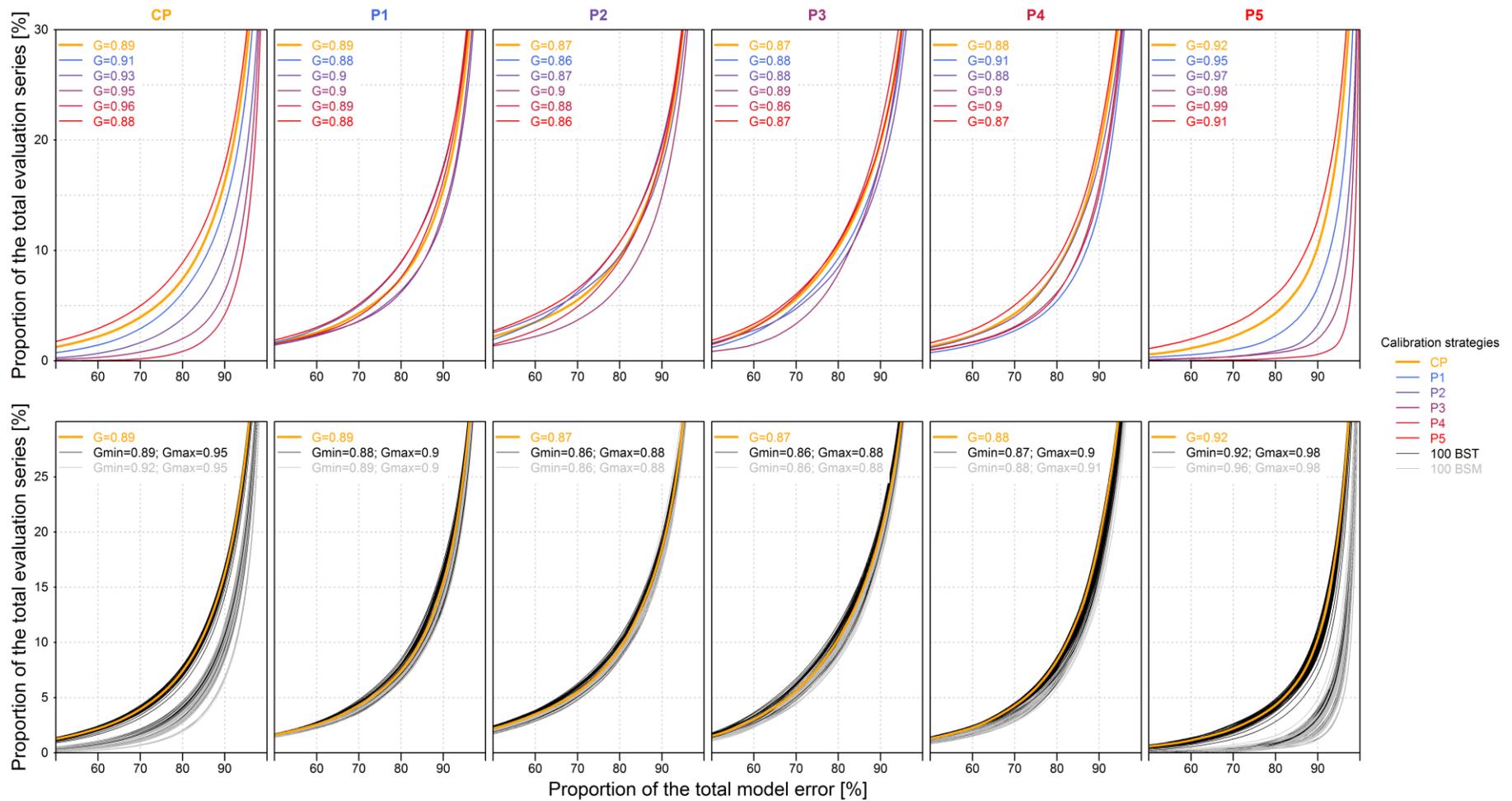
342
 343
 344
 345
 346
 347

Fig. 3 MORDOR performance over six different evaluation periods: the complete period (1976–2008) and five 6-year subperiods (P1–P5), in terms of OBJ_{EDF} (calibration criterion, top panel), NSE (middle panel) and NSE_{CDF} (bottom panel), according to different calibration strategies. Boxplots show the 0.10, 0.25, 0.50, 0.75 and 0.90 percentiles of the performance distributions obtained with the BST and BSM (excluding the August 2002 month) bootstrap calibration strategies.

348 Figure 4 shows Lorenz curves computed for each evaluation period and each
349 MORDOR parameter set. The Lorenz curve is classically used in economics; it was
350 introduced to represent the inequality of the wealth distribution, showing which
351 proportion of the population owns which proportion of the total wealth
352 (Lorenz, 1905). Here, the plotted Lorenz curves show the proportion of the total
353 evaluation series (time steps) as a function of the total model error (here the sum of
354 the squares of the model error), i.e. the cumulative distribution of ranked relative
355 model errors. For example, considering the CP parameter set and evaluating its error
356 over the CP period (orange line on the top left panel), the Lorenz curve reveals that
357 around 80% of the total MORDOR model error is made on less than 15% of the total
358 calibration period time steps. For the P4 parameter set (purple line), 80% of the total
359 MORDOR model error is made on less than 2.5% of the total calibration period time
360 steps. The complete analysis of the Lorenz curves shows first that P1 to P4 sub-
361 periods have similar error distributions considering the different calibration strategies.
362 On average, 80% of the total error is made on 5 to 12% of the calibration period time-
363 steps. For the complete period and the P5 sub-period (both including August 2002
364 flood), different Lorenz curves are obtained. For P1 to P4 parameter sets and
365 bootstrap calibrations not containing the August 2002 flood, a large proportion of the
366 total error is made on a smaller part of the total evaluation time steps, compared to the
367 P5 and CP parameter sets.

368

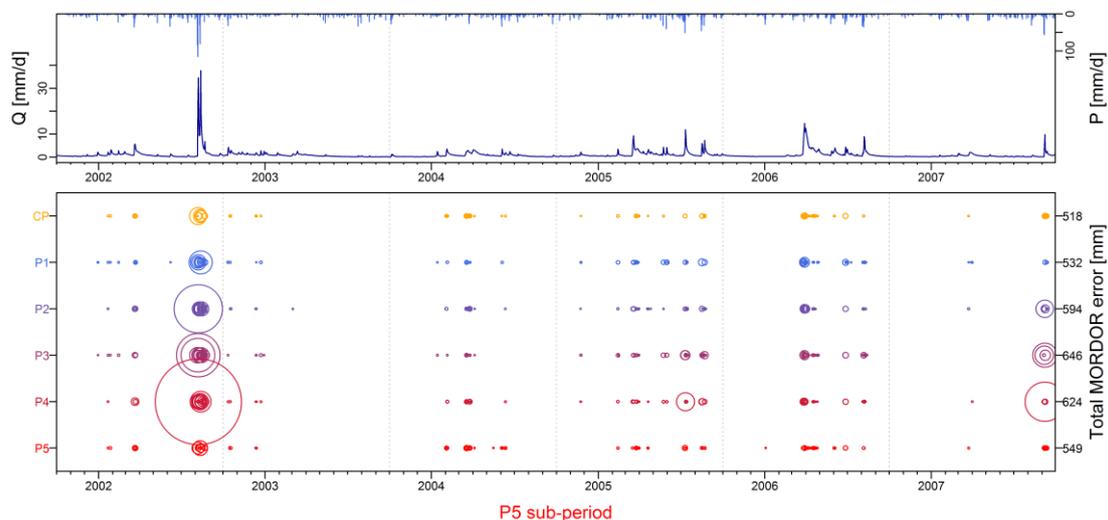
369 The shape of each Lorenz curve can be summarized and quantified with the
370 computation of the Gini index (Gini, 1912), which is the area between the line of
371 perfect equality ($x=y$) and the computed Lorenz curve. The Gini coefficient ranges
372 between 0 and 1: the higher the coefficient, the more uneven the distribution is. Such
373 coefficients have been computed for each Lorenz curve and are indicated on each
374 panel in Figure 4. The highest coefficient values are obtained for the CP and P5
375 periods, considering the P1 to P4 parameter sets and bootstrap calibrations not
376 containing the August 2002 flood. This shows that for these sub-periods and
377 parameter sets, a larger proportion of the total MORDOR model error is made on a
378 smaller part of the evaluation series.



379
380
381
382

Fig. 4 Lorenz curves showing the proportion of the total evaluation series as a function of the total MORDOR model error, computed over six different evaluation periods: the complete period (CP) and five 6-year sub-periods (P1–P5) and for different calibration strategies. Top row: (calibration on CP and P1–P5; and bottom row: calibration on 100 BST (black lines) and 100 BSM (grey lines)).

383 In order to confirm the substantial influence of the August 2002 flood presence within
 384 the MORDOR calibration and evaluation period, a time series of MORDOR model's
 385 error is plotted in Figure 5, for the P5 sub-period (2001-2007) and for the CP, P1 to
 386 P5 parameter sets. In this figure, the time steps representing 80% of the model
 387 cumulated total error are plotted with a circle, whose size is proportional to the model
 388 error made. Note that for the CP and the P5 parameter sets (orange and red circles,
 389 respectively), error is made in calibration while for the other parameter sets, it is an
 390 error in validation. Again, errors made with the CP and the P5 parameter sets are
 391 more evenly distributed than errors made by the P1 to P4 parameter sets, and the
 392 proportion of the 2002 event in the total error appears smaller in the CP and P5 sets.
 393 For the P1 to P4 parameter sets, the main errors are mainly concentrated on the
 394 August 2002 flood event. The P1 parameter error distribution appears to be slightly
 395 different from the P2 to P4 sets, with large errors made on the August 2002 flood
 396 event for the P1 parameter set, while the P2 to P4 parameter sets induce huge errors
 397 made on the August 2002 flood event and also a large error made on the September
 398 2008 flood event.
 399



400
 401 **Fig. 5** (top) Time series of observed precipitation and streamflow Kamp catchment series for the P5
 402 sub-period (2001–2007). (bottom) Time series of MORDOR model error on sub-period P5 for the CP
 403 and P1–P5 parameter sets: time steps where MORDOR error >1 mm/d is plotted with a circle of size
 404 proportional to the error.
 405

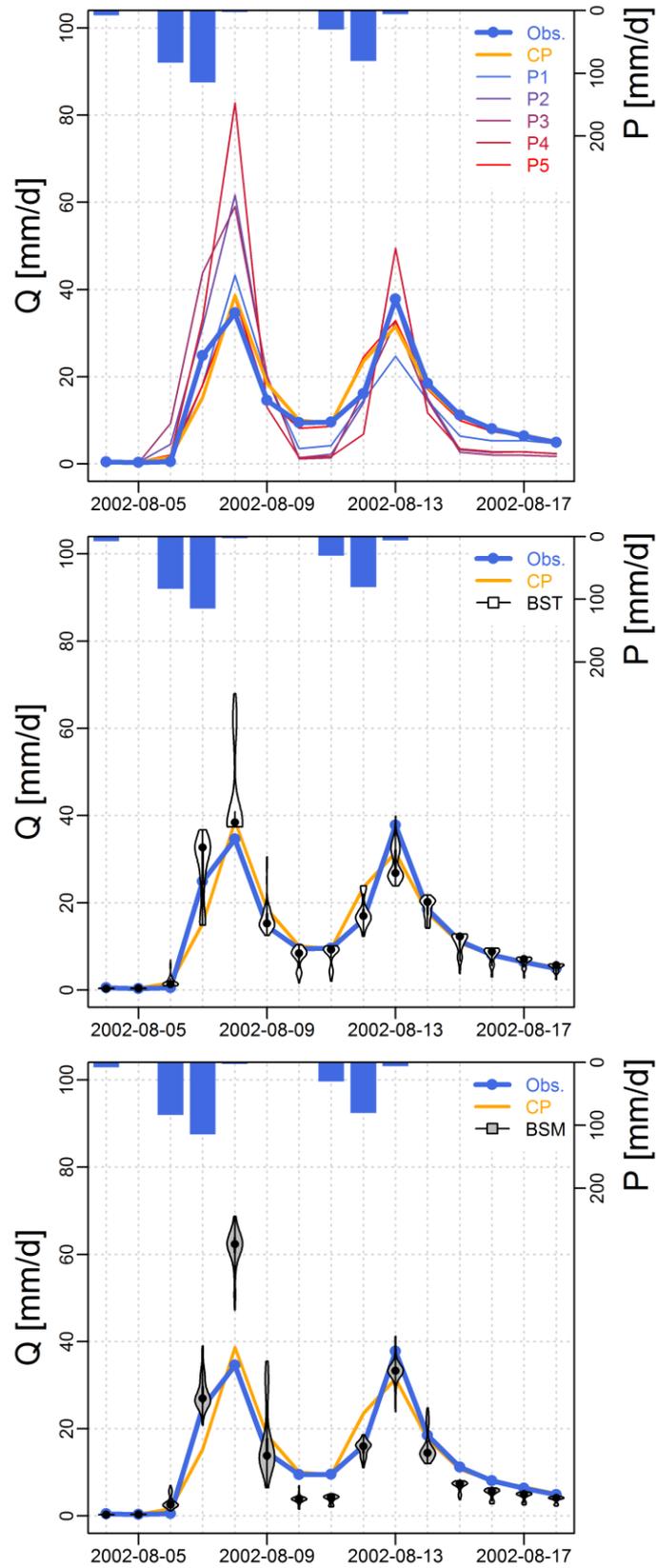
406 Figure 6 zooms in on the MORDOR streamflow simulations of August 2002, using
 407 the different parameter sets obtained with the different calibration options.
 408 Remarkably, this event is well simulated by the CP and P5 parameter sets.
 409 Conversely, it is poorly represented when using the P1 to P4 parameter sets (top
 410 panel), with a particularly strong overestimation of the first flood peak (8th of August)
 411 for the P2 to P4 parameter sets. In general, the P2 to P4 parameter sets induce an
 412 excessively responsive rainfall-runoff relationship by the MORDOR model. When
 413 considering bootstrap calibrations, two different MORDOR model behaviours seem to
 414 be obtained, depending on the presence of the August 2002 month within the
 415 calibration sub-periods. Rather logically, the August 2002 event is well simulated
 416 when it belongs to the calibration sub-period considered (centre panel), while it is
 417 poorly simulated with the parameter sets obtained with calibration sub-periods
 418 systematically excluding this event (bottom panel). Note that the few BST calibrations

419 (centre panel) highly overestimating the first flood peak have all been obtained with
420 calibration sub-periods not containing the 2002 year.

421

422 An investigation of MORDOR internal state dynamics (not shown here) revealed that
423 the responsive rainfall-runoff relationships simulated by several parameter sets are
424 induced, for this catchment, by the value of one particular MORDOR parameter,
425 L_{MAX} . This parameter is the maximum capacity of the L reservoir, which determines
426 the partitioning between a direct runoff, an indirect runoff and a deep percolation.
427 Thus, in the MORDOR model, when a large amount of water reaches this reservoir
428 (which is the case for the August 2002 event) a low L_{MAX} value implies that a large
429 proportion of this water is considered as direct runoff, while a higher L_{MAX} value
430 yields a larger proportion of indirect runoff. Interestingly, all MORDOR calibrations
431 that exclude the August 2002 period are characterized by small L_{MAX} values, while all
432 MORDOR calibration that include this event are characterized by high L_{MAX} values.
433 It clearly shows the weight of the August 2002 event's on the MORDOR calibration
434 and the event uniqueness according to the MORDOR model: the incoming rainfall for
435 this event is so large - relative to the observed streamflow - that the MORDOR model
436 needs to have a high L_{MAX} value for considering a large proportion of this incoming
437 water as indirect runoff and then not have a significant difference between observed
438 and simulated streamflow values. In validation on this event, the parameter sets
439 characterized by low L_{MAX} values (P1 to P4, BSM and 22 BST calibrations) produce
440 overly responsive rainfall-runoff relationships, with a substantial overestimation of
441 the first flood peak (8 August) and an underestimation of the flood recession (e.g. 10
442 and 11 August).

443



444
 445
 446
 447
 448

Fig. 6 August 2002 observed rainfall and streamflow series (solid and dotted lines) and MORDOR streamflow simulations considering different calibration strategies: (top) calibration over the five 6-year subperiods (P1–P5); (centre) calibration over 100 25-year sub-periods (BST); and (bottom) calibration over 100 25-year sub-periods excluding the month August 2002 (BSM).

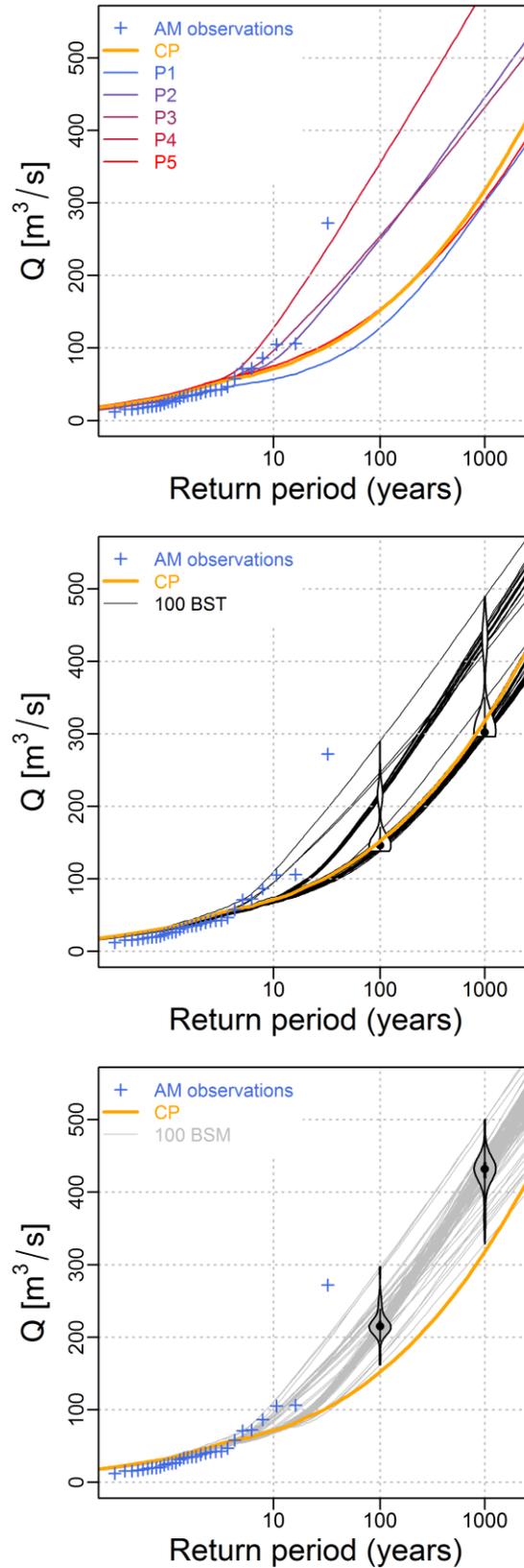
4.2. SCHADEX flood estimations

449
450

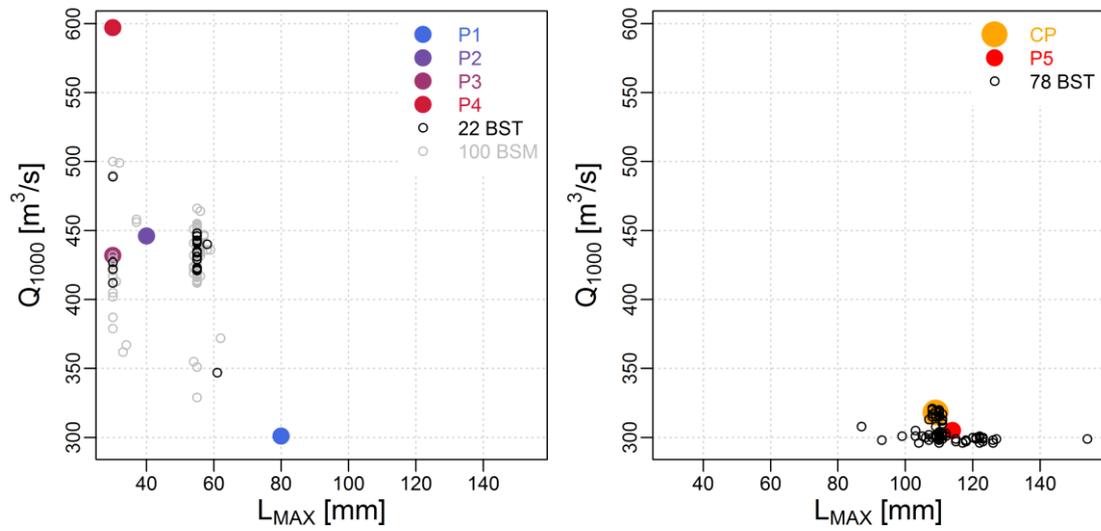
451 Figure 7 presents the SCHADEX flood estimations calculated with the different
452 MORDOR parameter sets, compared to the annual maximum series of daily
453 streamflow of the Kamp catchment. When considering the CP parameter set for the
454 MORDOR rainfall-runoff model, the Q_{1000} value estimated by the SCHADEX method
455 is $318 \text{ m}^3/\text{s}$. The flood estimations computed with the P1 to P5 MORDOR parameter
456 sets are presented on the first panel. The estimations computed using the P5 parameter
457 set are very similar to the reference estimation, unlike estimations made using the P1
458 parameter set (with lower Q_{10} to Q_{1000} values) and estimations made using the P2 to
459 P4 parameter sets (with substantially higher Q_{1000} values). A comparable range of
460 flood estimations is obtained when considering the bootstrap calibrations. Two types
461 of flood distributions are revealed, depending on the presence of the August 2002
462 within the calibration sub-periods: the flood estimations obtained with a calibration
463 sub-period excluding the August 2002 event are higher than the other estimations,
464 with median Q_{1000} values close to $430 \text{ m}^3/\text{s}$ (median value of the 100 estimations
465 presented in the bottom panel, in grey) and $300 \text{ m}^3/\text{s}$ (median value of the 100
466 estimations presented in the centre panel, in black), respectively.

467

468 This counterintuitive result is finally highlighted in Figure 8, where the left panel
469 groups all the Q_{1000} SCHADEX estimations obtained with the MORDOR parameter
470 sets containing the August 2002 month within the calibration period, while the right
471 panel groups all the Q_{1000} SCHADEX estimations obtained with the MORDOR
472 parameter sets excluding the August 2002 month. This figure clearly shows that
473 having the largest observed flood on the Kamp catchment within the MORDOR
474 rainfall-runoff model calibration period produces lower SCHADEX flood estimations.
475 In Figure 8, each Q_{1000} estimation has been plotted against its corresponding
476 MORDOR L_{MAX} parameter value, identified as responsible for the excessively
477 responsive August 2002 simulations (cf. Figure 6). Interestingly, the highest Q_{1000}
478 estimations are obtained with the lowest L_{MAX} values. The presence of the August
479 2002 event within the MORDOR calibration period thus induces a high L_{MAX}
480 parameter value and consequently a low Q_{1000} estimation. In this case, the L reservoir
481 is able to transform large amounts of incoming rainfall into indirect runoff.



482
 483 **Fig. 7** Annual maximum (AM) of daily streamflow observations (+) compared to the SCHADEX
 484 reference flood estimation (large solid lines) and flood estimations performed considering different
 485 MORDOR calibration strategies: (top) calibration over the five 6-year sub-periods (P1–P5); (centre)
 486 calibration over 100 25-year sub-periods (BST); and (bottom) calibration over 100 25-year subperiods
 487 discarding the month August 2002. Violin plots represent the distribution of bootstrap flood
 488 estimations for the 100- and 1000-year return periods.



489
 490
 491
 492

Fig. 8 SCHADEX Q_{1000} flood estimations plotted against the corresponding values of MORDOR parameter, L_{MAX} , both obtained with the MORDOR calibration including August 2002 (left) and excluding August 2002 (right).

5. DISCUSSION AND CONCLUSION

The Kamp at Zwettl is an interesting case study for all hydrologists specializing in the art of flood frequency estimation, since it experienced a remarkable flood in August 2002, reaching a value three times higher than the second largest observed flood in terms of peak value over the 55 years of available observations (Viglione *et al.*, 2013). This case study provides a rare opportunity to investigate the impact of such a remarkable event on flood frequency estimation. Here, numerous extreme flood estimations were made on this catchment with the SCHADEX stochastic method, which is based on a conceptual rainfall-runoff model. Following the calibration protocol of the 2013 “Hydrology under change” IAHS workshop proposed by Thirel *et al.* (2014, this issue), the rainfall-runoff model was calibrated over five 6-year sub-periods. Additionally, bootstrap calibrations were performed, following the methodology proposed by Brigode *et al.* (2014). In total, 206 calibrations of the MORDOR rainfall-runoff model were performed in this study, each of them used for producing different SCHADEX flood estimations.

The results confirmed the usefulness of the multi-period and bootstrap testing schemes to identify the dependence of model performance and flood estimates on the information contained in the calibration period and the presence of large flood events. As already pointed out by Viglione *et al.* (2013), the August 2002 event appears to play a key role in the flood frequency estimation on the Kamp River. Here, the presence of the event within the calibration sub-periods strongly influences the rainfall-runoff model calibration, the validation performance and the extreme flood estimations. All the parameter sets obtained with calibration periods that do not contain the August 2002 month perform poorly on the evaluation periods containing this event. Those parameter sets are characterized by an excessively responsive rainfall-runoff transformation, while the other ones simulate smoother hydrographs. An investigation of the MORDOR model’s internal states reveals that one parameter (L_{MAX}) is responsible for this particular simulation dynamic, and that the L_{MAX} value obtained after calibration depends on the presence or absence of the August 2002 flood within the calibration period. Those “responsive” parameter sets obtained when the August 2002 event is excluded from the calibration period produce higher extreme flood estimations compared to the other parameter sets, confirming the findings of Brigode *et al.* (2014). Thus, Q_{1000} estimates were much higher when model calibration did not include the large 2002 flood event. Interestingly, this sensitivity to the presence of the August 2002 flood is contrary (and thus counterintuitive) to the sensitivity obtained when applying a classical flood frequency analysis method (i.e. statistical estimation of flood quantiles using only streamflow series), highlighted by Viglione *et al.* (2013, Figure 3a and Figure 3b).

The bootstrap calibration methodology is shown to be a useful tool for an objective quantification of the model’s dependence on the calibration period, considering the rainfall-runoff simulations and the extreme flood estimations. Computing a rigorous statistical confidence interval would require more statistical processing, but it nevertheless provides a “first guess” of the uncertainty associated with the calibration period and a range of extreme flood estimations.

Graphical and numerical tools have also been proposed in this study in order to highlight the influence of particular flood events on the calibration of rainfall-runoff

543 models. Lorenz curves and the Gini coefficient provide a simple but efficient way to
544 characterize the distribution of model errors and could be very useful to detect
545 calibration periods where a few time steps cause a large proportion of the model's
546 errors. In this regard, it could be interesting to compare, for different rainfall-runoff
547 models and different catchments, whether the events selected by this kind of analysis
548 as "strongly influencing the model calibration" are the same as the ones selected by
549 other approaches such as DYNIA (Wagener *et al.*, 2003) or ICE (Singh & Bárdossy,
550 2012).

551

552 Given the dominating impact of the August 2002 data, one could wonder whether in
553 practice a hydrologist engineer working on extreme flood estimation on the Kamp
554 catchment should consider discarding the August 2002 data. On one hand, if he is
555 applying a classical flood frequency analysis method, considering both the observed
556 flood series and additional information such as historical floods or regional
557 information is necessary to significantly reduce the weight of the August 2002 event
558 and thus the flood estimation uncertainty. On the other hand, if he is applying a flood
559 simulation method based on a rainfall-runoff model (e.g. SCHADEX), it would be in
560 principle more appropriate to consider this type of flood events for the rainfall-runoff
561 model calibration, because it enables the model to be trained on exceptional floods
562 and thus to have the opportunity to identify the high flood-prevailing processes, which
563 could differ from current floods (e.g. Rogger *et al.*, 2012). However, the rainfall-
564 runoff model robustness issue addressed in this case study (illustrated in Figures 5 and
565 6 for example) could be used as "process-based arguments" for an expert rainfall-
566 runoff modeller who believes more in his model than in the observed data to discard
567 particular flood event data.

568

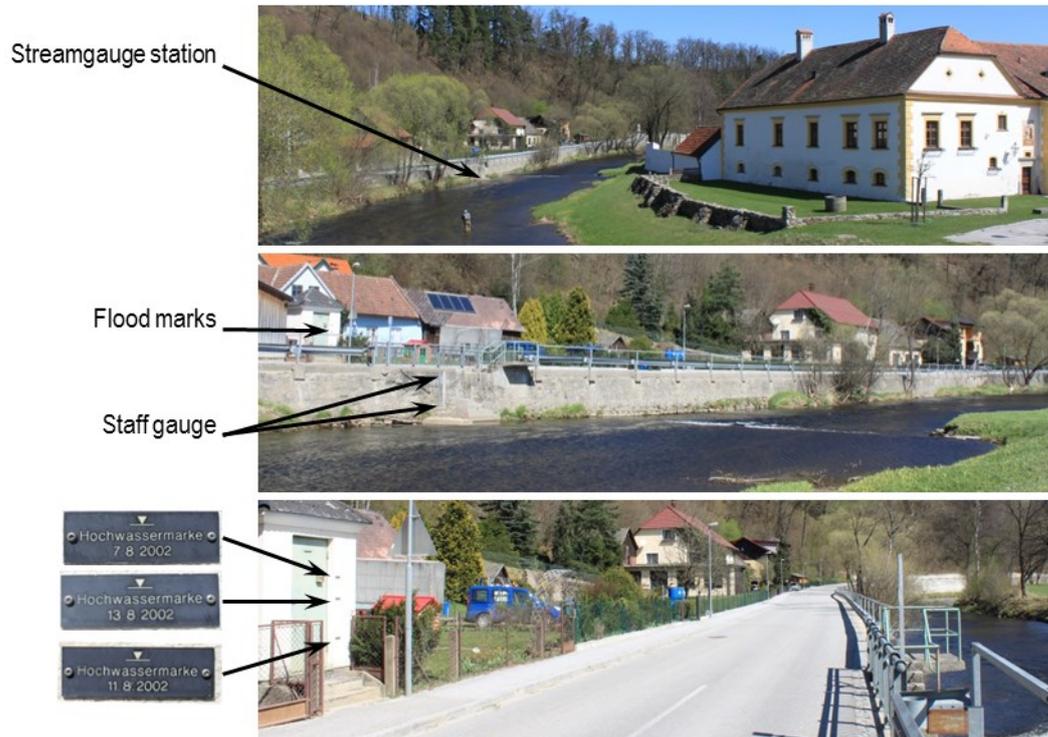
569 In both cases, the question of the uncertainty of the rainfall and streamflow
570 measurement of such events needs to be investigated further. For example, Lang *et al.*
571 (2010) suggested that numerous French gauging streamflow stations are not reliable
572 for floods with a return period higher than 2 years. For the Kamp catchment, Figure 9
573 shows pictures of the gauging station (at Zwettl (Bahnbrücke), station ID 207944) and
574 flood tracks left by the August 2002 flood event, illustrating the potentially
575 considerable uncertainty of streamflow measurement for this event. The Kamp River
576 was clearly out of its banks and thus out of the usual streamflow rating curve. An
577 interesting perspective would be to quantify the measurement uncertainty of this event
578 and then recalibrate the MORDOR model before performing new SCHADEX flood
579 estimations.

580

581 **Acknowledgements**

582 The assistance of Dr. Guillaume Thirel (Irstea), Prof. Ralf Merz (UFZ) and Prof. Juraj
583 Parajka (Vienna University of Technology) with the Kamp River data set is gratefully
584 acknowledged. The authors thank the anonymous reviewers and the Associate Editor,
585 Dr. Alberto Viglione, for their constructive comments which helped improve the
586 quality of the manuscript. Finally, the authors would like to thank Rémy Garçon
587 (EDF-DTG) for his constrictive comments.

588



589

590 **Fig. 9** Zwettl (Bahnbrücke, ID: 207944) streamgauge station and Kamp River. The bottom panel shows
 591 the August 2002 flood marks. Photos: P. Brigode, April 2011.

592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641

6. REFERENCES

- Andréassian, V., Hall, A., Chahinian, N. & Schaake, J. (2006) Large sample basin experiments for hydrological model parameterization: results of the Model Parameter Experiment (MOPEX). In: *Large sample basin experiments for hydrological model parameterization: results of the model parameter experiment (MOPEX)* IAHS Publication 307, 346. Retrieved from <http://www.cabdirect.org/abstracts/20073200660.html>
- Andréassian, V., Moine, N. Le, Perrin, C., Ramos, M.-H., Oudin, L., Mathevet, T., Lerat, J., et al. (2012) All that glitters is not gold: the case of calibrating hydrological models. *Hydrological Processes* **26**(14), 2206–2210. doi:10.1002/hyp.9264
- Berthet, L., Andréassian, V., Perrin, C. & Loumagne, C. (2010) How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion. *Hydrological Sciences Journal* **55**(6), 1063–1073. doi:10.1080/02626667.2010.505891
- Beven, K. & Westerberg, I. (2011) On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrol. Process.* **25**(10), 1676–1680. doi:10.1002/hyp.7963
- Blöschl, G., Nester, T., Komma, J., Parajka, J. & Perdigão, R. A. P. (2013) The June 2013 flood in the Upper Danube Basin, and comparisons with the 2002, 1954 and 1899 floods. *Hydrol. Earth Syst. Sci.* **17**(12), 5197–5212. doi:10.5194/hess-17-5197-2013
- Brigode, P., Bernardara, P., Gailhard, J., Garavaglia, F., Ribstein, P. & Merz, R. (2013b) Optimization of the geopotential heights information used in a rainfall-based weather patterns classification over Austria. *International Journal of Climatology* **33**(6), 1563–1573. doi:10.1002/joc.3535
- Brigode, P., Bernardara, P., Paquet, E., Gailhard, J., Garavaglia, F., Merz, R., Mićović, Z., et al. (2014) Sensitivity analysis of SCHADEX extreme flood estimations to observed hydrometeorological variability. *Water Resources Research*. doi:10.1002/2013WR013687
- Brigode, P., Oudin, L. & Perrin, C. (2013a) Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change? *Journal of Hydrology* **476**(0), 410–425. doi:10.1016/j.jhydrol.2012.11.012
- Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M. & Hendrickx, F. (2012) Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. *Water Resour. Res.* **48**, W05552. doi:10.1029/2011WR011721
- Donnelly-Makowecki, L. M. & Moore, R. D. (1999) Hierarchical testing of three rainfall-runoff models in small forested catchments. *Journal of Hydrology* **219**(3-4), 136–152. doi:10.1016/S0022-1694(99)00056-6
- Duan, Q., Sorooshian, S. & Gupta, V. (1992) Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.* **28**(4), PP. 1015–1031. doi:10.1029/91WR02985
- Ebtehaj, M., Moradkhani, H. & Gupta, H. V. (2010) Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling. *Water Resour. Res.* **46**(7), W07515. doi:10.1029/2009WR007981
- Franchini, M. (1996) Use of a genetic algorithm combined with a local search method for the automatic calibration of conceptual rainfall-runoff models.

- 642 *Hydrological Sciences Journal* **41**(1), 21–39.
643 doi:10.1080/02626669609491476
- 644 Garavaglia, F., Gailhard, J., Paquet, E., Lang, M., Garçon, R. & Bernardara, P. (2010)
645 Introducing a rainfall compound distribution model based on weather patterns
646 sub-sampling. *Hydrol. Earth Syst. Sci.* **14**(6), 951–964. doi:10.5194/hess-14-
647 951-2010
- 648 Garçon, R. (1999) Modèle global pluie-débit pour la prévision et la prédétermination
649 des crues. *La Houille Blanche* (7-8), 88–95. doi:10.1051/lhb/1999088
- 650 Gharari, S., Hrachowitz, M., Fenicia, F. & Savenije, H. H. G. (2013) An approach to
651 identify time consistent model parameters: sub-period calibration. *Hydrol.*
652 *Earth Syst. Sci.* **17**(1), 149–161. doi:10.5194/hess-17-149-2013
- 653 Gini, C. (1912) Variabilità e mutabilità. *Reprinted in Memorie di metodologica*
654 *statistica* (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi **1**.
655 Retrieved from <http://adsabs.harvard.edu/abs/1912vamu.book.....G>
- 656 Grubbs, F. E. (1969) Procedures for Detecting Outlying Observations in Samples.
657 *Technometrics* **11**(1), 1–21. doi:10.1080/00401706.1969.10490657
- 658 Herman, J. D., Reed, P. M. & Wagener, T. (2013) Time-varying sensitivity analysis
659 clarifies the effects of watershed model formulation on model behavior. *Water*
660 *Resources Research*. doi:10.1002/wrcr.20124
- 661 Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M.,
662 Pomeroy, J. W., Arheimer, B., et al. (2013) A decade of Predictions in
663 Ungauged Basins (PUB) - a review. *Hydrological Sciences Journal* **58**(6),
664 1198–1255. doi:10.1080/02626667.2013.803183
- 665 Jarvis, A., Reuter, H. I., Nelson, A. & Guevara, E. (2008) Hole-filled SRTM for the
666 globe Version 4 (available from the CGIAR-CSI SRTM 90m Database).
667 Retrieved from <http://srtm.csi.cgiar.org>
- 668 Katz, R. W., Parlange, M. B. & Naveau, P. (2002) Statistics of extremes in hydrology.
669 *Advances in Water Resources* **25**(8–12), 1287–1304. doi:10.1016/S0309-
670 1708(02)00056-8
- 671 Klemeš, V. (1986) Operational testing of hydrological simulation models.
672 *Hydrological Sciences Journal* **31**(1), 13. doi:10.1080/02626668609491024
- 673 Komma, J., Reszler, C., Blöschl, G. & Haiden, T. (2007) Ensemble prediction of
674 floods – catchment non-linearity and forecast probabilities. *Natural Hazards*
675 *and Earth System Sciences* **7**(4), 431–444. doi:10.5194/nhess-7-431-2007
- 676 Koutsoyiannis, D. (2006) Nonstationarity versus scaling in hydrology. *Journal of*
677 *Hydrology* **324**(1–4), 239–254. doi:10.1016/j.jhydrol.2005.09.022
- 678 Laio, F., Allamano, P. & Claps, P. (2010) Exploiting the information content of
679 hydrological outliers’ for goodness-of-fit testing. *Hydrol. Earth Syst. Sci.*
680 **14**(10), 1909–1917. doi:10.5194/hess-14-1909-2010
- 681 Lamb, R. (1999) Calibration of a conceptual rainfall-runoff model for flood frequency
682 estimation by continuous simulation. *Water Resources Research* **35**(10),
683 3103–3114. doi:10.1029/1999WR900119
- 684 Lang, M., Pobanz, K., Renard, B., Renouf, E. & Sauquet, E. (2010) Extrapolation of
685 rating curves by hydraulic modelling, with application to flood frequency
686 analysis. *Hydrological Sciences Journal* **55**(6), 883–898.
687 doi:10.1080/02626667.2010.504186
- 688 Lawrence, D., Paquet, E., Gailhard, J. & Fleig, A. K. (2014) Stochastic semi-
689 continuous simulation for extreme flood estimation in catchments with
690 combined rainfall–snowmelt flood regimes. *Nat. Hazards Earth Syst. Sci.*
691 **14**(5), 1283–1298. doi:10.5194/nhess-14-1283-2014

- 692 Lorenz, M. O. (1905) Methods of Measuring the Concentration of Wealth.
693 *Publications of the American Statistical Association* **9**(70), 209–219.
694 doi:10.2307/2276207
- 695 Mathevet, T. (2005) *Quels modèles pluie-débit globaux au pas de temps horaire ?*
696 *Développements empiriques et comparaison de modèles sur un large*
697 *échantillon de bassins versants*. ENGREF, Paris.
- 698 Merz, R., Parajka, J. & Blöschl, G. (2011) Time stability of catchment model
699 parameters: Implications for climate impact analyses. *Water Resour. Res.* **47**,
700 17 PP. doi:10.1029/2010WR009505
- 701 Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W.,
702 Lettenmaier, D. P. & Stouffer, R. J. (2008) Stationarity Is Dead: Whither
703 Water Management? *Science* **319**(5863), 573–574.
704 doi:10.1126/science.1151915
- 705 Montanari, A. (2012) Hydrology of the Po River: looking for changing patterns in
706 river discharge. *Hydrol. Earth Syst. Sci.* **16**(10), 3739–3747. doi:10.5194/hess-
707 16-3739-2012
- 708 Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L.,
709 Koutsoyiannis, D., et al. (2013) ‘Panta Rhei-Everything Flows’: Change in
710 hydrology and society - The IAHS Scientific Decade 2013-2022. *Hydrological*
711 *Sciences Journal* **58**(6), 1256–1275. doi:10.1080/02626667.2013.809088
- 712 Muñoz, E., Arumí, J. L. & Rivera, D. (2013) Watersheds are not static: Implications
713 of climate variability and hydrologic dynamics in modeling. *Bosque (Valdivia)*
714 **34**(1), 7–11. doi:10.4067/S0717-92002013000100002
- 715 Nash, J. E. & Sutcliffe, J. V. (1970) River flow forecasting through conceptual
716 models part I – A discussion of principles. *Journal of Hydrology* **10**(3), 282–
717 290. doi:10.1016/0022-1694(70)90255-6
- 718 Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Lay,
719 M. Le, et al. (2014) Benchmarking hydrological models for low-flow
720 simulation and forecasting on French catchments. *Hydrol. Earth Syst. Sci.*
721 **18**(8), 2829–2857. doi:10.5194/hess-18-2829-2014
- 722 Paquet, E., Garavaglia, F., Garçon, R. & Gailhard, J. (2013) The SCHADDEX method:
723 A semi-continuous rainfall–runoff simulation for extreme flood estimation.
724 *Journal of Hydrology* **495**, 23–37. doi:10.1016/j.jhydrol.2013.04.045
- 725 Peel, M. C. & Blöschl, G. (2011) Hydrological modelling in a changing world.
726 *Progress in Physical Geography* **35**(2), 249–261.
- 727 Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C. & Mathevet, T.
728 (2007) Impact of limited streamflow data on the efficiency and the parameters
729 of rainfall-runoff models. *Hydrological Sciences Journal* **52**(1), 131–151.
730 doi:10.1623/hysj.52.1.131
- 731 Rogger, M., Pirkel, H., Viglione, A., Komma, J., Kohl, B., Kirnbauer, R., Merz, R., et
732 al. (2012) Step changes in the flood frequency curve: Process controls. *Water*
733 *Resour. Res.* **48**, 15 PP. doi:201210.1029/2011WR011187
- 734 Seibert, J. (2003) Reliability of model predictions outside calibration conditions.
735 *Nordic Hydrology* **34**(5), 477–492. doi:10.2166/nh.2003.028
- 736 Seibert, J. & Beven, K. J. (2009) Gauging the ungauged basin: how many discharge
737 measurements are needed? *Hydrology and Earth System Sciences* **13**(6), 883–
738 892. doi:10.5194/hess-13-883-2009
- 739 Selle, B. & Hannah, M. (2010) A bootstrap approach to assess parameter uncertainty
740 in simple catchment models. *Environmental Modelling & Software* **25**(8),
741 919–926. doi:10.1016/j.envsoft.2010.03.005

- 742 Singh, S. K. & Bárdossy, A. (2012) Calibration of hydrological models on
 743 hydrologically unusual events. *Advances in Water Resources* **38**, 81–91.
- 744 Singh, S. K., Liang, J. & Bárdossy, A. (2012) Improving the calibration strategy of
 745 the physically-based model WaSiM-ETH using critical events. *Hydrological
 746 Sciences Journal* **57**(8), 1487–1505. doi:10.1080/02626667.2012.727091
- 747 Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P.,
 748 Folton, N., et al. (2014) Hydrology under change. An evaluation protocol to
 749 investigate how hydrological models deal with changing catchments. *Hydrol.
 750 Sci. J. (Special Issue)*.
- 751 Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J.-M., Viney, N. R. & Teng, J. (2010)
 752 Climate non-stationarity - Validity of calibrated rainfall-runoff models for use
 753 in climate change studies. *Journal of Hydrology* **394**(3-4), 447–457.
 754 doi:10.1016/j.jhydrol.2010.09.018
- 755 Viglione, A., Chirico, G. B., Komma, J., Woods, R., Borga, M. & Blöschl, G. (2010)
 756 Quantifying space-time dynamics of flood event types. *Journal of Hydrology*
 757 **394**(1–2), 213–229. doi:10.1016/j.jhydrol.2010.05.041
- 758 Viglione, A., Merz, R., Salinas, J. L. & Blöschl, G. (2013) Flood frequency
 759 hydrology: 3. A Bayesian analysis. *Water Resources Research* **49**(2), 675–
 760 692. doi:10.1029/2011WR010782
- 761 Wagener, T. & Kollat, J. (2007) Numerical and visual evaluation of hydrological and
 762 environmental models using the Monte Carlo analysis toolbox. *Environmental
 763 Modelling & Software* **22**(7), 1021–1033. doi:10.1016/j.envsoft.2006.06.017
- 764 Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S. & Gupta, H. V. (2003)
 765 Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic
 766 identifiability analysis. *Hydrological Processes* **17**(2), 455–476.
 767 doi:10.1002/hyp.1135
- 768 Wang, Q. J. (1991) The Genetic Algorithm and Its Application to Calibrating
 769 Conceptual Rainfall-Runoff Models. *Water Resour. Res.* **27**(9), 2467–2471.
 770 doi:10.1029/91WR01305
- 771 Wang, Q. J. (1997) Using genetic algorithms to optimise model parameters.
 772 *Environmental Modelling & Software* **12**(1), 27–34. doi:10.1016/S1364-
 773 8152(96)00030-8
- 774 Ward, P. J., Eisner, S., Flörke, M., Dettinger, M. D. & Kummu, M. (2014) Annual
 775 flood sensitivities to El Niño–Southern Oscillation at the global scale. *Hydrol.
 776 Earth Syst. Sci.* **18**(1), 47–66. doi:10.5194/hess-18-47-2014
- 777 Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T. & Gailhard, J. (2012)
 778 Statistical processing of forecasts for hydrological ensemble prediction: a
 779 comparative study of different bias correction strategies. *Adv. Sci. Res.* **8**, 135–
 780 141. doi:10.5194/asr-8-135-2012