



HAL
open science

Data-based comparison of frequency analysis methods: A general framework

Benjamin Renard, K. Kochanek, M. Lang, F. Garavaglia, E. Paquet, L.
Neppel, K. Najib, Julie Carreau, P. Arnaud, Y. Aubert, et al.

► To cite this version:

Benjamin Renard, K. Kochanek, M. Lang, F. Garavaglia, E. Paquet, et al.. Data-based comparison of frequency analysis methods: A general framework. *Water Resources Research*, 2013, 49, p. 1 - p. 19. 10.1002/wrcr.20087 . hal-00811184

HAL Id: hal-00811184

<https://hal.science/hal-00811184v1>

Submitted on 10 Apr 2013

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.

Data-based comparison of frequency analysis methods: a general framework

B. Renard⁽¹⁾, K. Kochanek^(1,7), M. Lang⁽¹⁾, F. Garavaglia⁽²⁾, E. Paquet⁽²⁾, L. Neppel⁽³⁾, K. Najib⁽³⁾, J. Carreau⁽³⁾, P. Arnaud⁽⁴⁾, Y. Aubert⁽⁴⁾, F. Borchì⁽⁵⁾, J.-M. Soubeyrou⁽⁵⁾, S. Jourdain⁽⁵⁾, J.-M. Veysseire⁽⁵⁾, E. Sauquet⁽¹⁾, T. Cipriani⁽¹⁾ and A. Auffray⁽⁶⁾

(1) Irstea, UR HHLY Hydrology-Hydraulics, Lyon, France

(2) EDF-DTG, Grenoble, France

(3) University Montpellier II, UMR HydroSciences, Montpellier, France

(4) Irstea, UR OHAX, Aix-en-Provence, France

(5) Meteo-France, Direction de la Climatologie, Toulouse, France

(6) Meteo-France, Direction Interrégionale Centre-Est, Lyon, France

(7) Institute of Geophysics, Polish Academy of Sciences, Warsaw, Poland

Submitted for publication in Water Resources Research

October 2012

22

23 **Abstract**

24 An abundance of methods have been developed over the years to perform the frequency
25 analysis (FA) of extreme environmental variables. Although numerous comparisons between
26 these methods have been implemented, no general comparison framework has been agreed
27 upon so far. The objective of this paper is to build the foundation of a data-based comparison
28 framework, which aims at complementing more standard comparison schemes based on
29 Monte Carlo simulations or statistical testing. This framework is based on the following
30 general principles: (i) emphasis is put on the predictive ability of competing FA
31 implementations, rather than their sole descriptive ability measured by some goodness-of-fit
32 criterion; (ii) predictive ability is quantified by means of reliability indices, describing the
33 consistency between validation data (not used for calibration) and FA predictions; (iii)
34 stability is also quantified, i.e. the ability of a FA implementation to yield similar estimates
35 when calibration data change; (iv) the necessity to subject uncertainty estimates to the same
36 scrutiny as point-estimates is recognized, and a practical approach based on the use of the
37 predictive distribution is proposed for this purpose. This framework is then applied to a case
38 study involving 364 gauging stations in France, where 10 FA implementations are compared.
39 These implementations correspond to the local, regional and local-regional estimation of
40 Gumbel and Generalized Extreme Value distributions. Results show that reliability and
41 stability indices are able to reveal marked difference between FA implementations. Moreover,
42 the case study also confirms that using the predictive distribution to indirectly scrutinize
43 uncertainty estimates is a viable approach, with distinct FA implementations showing marked
44 differences in the reliability of their uncertainty estimates. The proposed comparison
45 framework therefore constitutes a valuable tool to compare the predictive reliability of
46 competing FA implementations, along with the reliability of their uncertainty estimates.

47

47 **1. Introduction**

48 Frequency analysis (FA) of extremes is one of the cornerstones of hazard quantification and
49 risk assessment. Its basic objective is to estimate the distribution of some environmental
50 variable X . Such a distribution can be used to estimate the exceedance probability of a given
51 value of X , or alternatively, to estimate the p -quantile of X (where p denotes the non-
52 exceedance probability). The estimation of quantiles is of great importance since they are
53 used to design civil engineering structures (e.g. dams, reservoirs, bridges) or to map hazard-
54 prone areas where restrictions may be enforced (e.g. building restrictions in flood zones).

55 FA has been the subject of extensive research, yielding an abundance of approaches that can
56 roughly be classified as follows:

- 57 • At-Site FA is a standard statistical analysis: parameters of a pre-specified distribution are
58 estimated based on at-site observations of the variable X .
- 59 • Climate/Weather-informed at-site FA uses additional meteorological [e.g., weather type,
60 *Garavaglia et al.*, 2010] or climatic [e.g. Interdecadal Pacific Oscillation IPO, *Micevski et*
61 *al.*, 2006a] information. This family of methods stems from the observation that the
62 distribution of X depends on some climate or weather state variable.
- 63 • Historical and paleoflood analyses are based on documentary sources or proxy data from
64 e.g. sediment deposits. Such information is used to extend the record period from the last
65 decades to several centuries (historical data) or millennia (paleoflood data). Specific statistical
66 frameworks have been developed to treat such additional information [e.g. *Stedinger and*
67 *Cohn*, 1986; *O'Connel et al.*, 2002; *Parent and Bernier*, 2003; *Naulet et al.*, 2005; *Reis and*
68 *Stedinger*, 2005; *Neppel et al.*, 2010; *Payrastre et al.*, 2011].
- 69 • Regional Frequency Analysis (RFA) jointly uses data from several sites to perform the
70 inference, which may improve the precision of estimates [see e.g. *Durrans and Kirby*, 2004;
71 *Yu et al.*, 2004; *Overeem et al.*, 2008; *Kysely et al.*, 2011 for recent examples]. Moreover,
72 RFA allows estimating quantiles and related uncertainties at an ungauged site.
- 73 • Model-based FA (sometimes referred to as “continuous simulation methods”) uses a
74 simulation model reproducing the main characteristics of the environmental variable [*Arnaud*
75 *and Lavabre*, 1999, for rainfall; *Boughton and Droop*, 2003, for floods]. Quantiles are then
76 directly derived from long series generated from the model.

77 Within each of these families, a large number of variants exist, differing in e.g. the assumed
78 parent distribution (e.g. Generalized Extreme Value (GEV), Log-Pearson), the parameter
79 estimation approach (e.g. maximum likelihood (ML), moment), the definition of homogenous
80 regions or the choice of the simulation model. To avoid ambiguity, the following terminology
81 is systematically used in this paper: a “FA family” refers to any of the previously described
82 families, while a specific variant within a family is referred to as a “FA implementation”. For
83 instance, the local estimation of a GEV distribution with (i) the ML approach, and (ii) the
84 moments approach will be considered as two distinct FA implementations, belonging to the
85 same FA family.

86 In practice, users may feel lost facing so many FA implementations. Consequently, national
87 guideline documents for flood FA help practitioners in realizing their analyses with best-
88 practice methods. Such documents were released e.g. in the UK [*Reed et al.*, 1999], in the US
89 [*Interagency Advisory Committee on Water Data*, 1982], in Switzerland [*Spreafico et al.*,
90 2003] or in Australia [*Institution of Engineers Australia*, 1987].

91 In addition to these end-user-oriented guideline documents, a large number of comparative
92 studies between competing FA implementations have been reported in the research literature
93 (a non-exhaustive review will be proposed in section 2). However, as noted by *Bobee et al.*
94 [1993], the comparison framework varies from one study to another. *Bobee et al.* therefore
95 advocated “*a systematic approach to comparing distributions used in flood frequency*
96 *analysis*”, which is still not agreed upon to our knowledge.

97 Moreover, in recent years there has been a growing emphasis on the importance of
98 quantifying and communicating uncertainties in FA implementations [e.g., *Hall et al.*, 2004;
99 *Naulet et al.*, 2005; *Renard et al.*, 2006a; *Renard et al.*, 2006b; *Kysely*, 2008; *Lee and Kim*,
100 2008; *Hine and Hall*, 2010; *Lima and Lall*, 2010; *Neppel et al.*, 2010]. However, while most
101 FA implementations include an evaluation of uncertainties, the question of the reliability of
102 estimated uncertainties has received less attention in FA [but see e.g. *Kysely*, 2008;
103 *Garavaglia et al.*, 2011, for recent exceptions]. Other fields of environmental sciences (e.g.
104 weather forecasting [*Dawid*, 1984; *Atger*, 1999; *Gneiting et al.*, 2007] or hydrological
105 modeling [*Hall et al.*, 2007; *Laio and Tamea*, 2007; *Thyer et al.*, 2009; *Renard et al.*, 2010])
106 have recognized the need to scrutinize uncertainty estimates.

107 The general objective of this paper is to build the foundation of a methodological framework
108 devoted to the data-based comparison of FA implementations. This framework aims to

109 complement (but not replace) other comparison frameworks based for instance on Monte-
110 Carlo simulations or statistical testing. Importantly, the framework we are proposing is built
111 in order to meet the following requirements:

112 [R1] It should enable the inclusion of any FA implementation, whatever its family (at-
113 site, regional, model-based etc.).

114 [R2] It should enable the comparison of estimated uncertainties.

115 This paper is organized as follows. Section 2 proposes a short review of commonly used
116 comparison frameworks, and emphasizes the differences between them in terms of underlying
117 objectives, advantages and limitations. Section 3 then describes the data-based comparison
118 framework. In particular, section 3.2 proposes several indices to quantify the performance of
119 competing FA implementations, and section 3.3 introduces the predictive distribution as an
120 indirect way to compare uncertainty estimates. A case study based on 364 gauging stations in
121 France illustrates the application of the comparison framework (section 4). Limitations are
122 discussed in section 5, before summarizing the main conclusions in section 6.

123 **2. A short review of standard comparison frameworks**

124 **2.1. Simulation-based comparisons**

125 Simulation-based approaches use Monte-Carlo-generated data. Knowing the true distribution,
126 the performance of a FA implementation can be quantified by means of formal and objective
127 statistical criteria such as bias, root mean squared error (RMSE), etc. This approach has been
128 widely used for the comparison between various distributions and/or estimation approaches
129 [e.g., *Hosking et al.*, 1985; *Kroll and Stedinger*, 1996; *Madsen et al.*, 1997a; *Madsen et al.*,
130 1997b; *Sankarasubramanian and Srinivasan*, 1999; *Durrans and Tomic*, 2001; *Ribatet et al.*,
131 2007; *He and Valeo*, 2009; *Meshgi and Khalili*, 2009], and for robustness studies [i.e., the
132 performance of a method outside its conditions of application, see e.g., *Stedinger and Cohn*,
133 1986; *England et al.*, 2003b; *Markiewicz and Strupczewski*, 2009]. Moreover, an important
134 advantage of simulation-based approaches is that they enable a formal evaluation of estimated
135 uncertainties [see e.g. *Stedinger*, 1983b; *Stedinger and Tasker*, 1985; *Chowdhury and*
136 *Stedinger*, 1991; *Cohn et al.*, 2001; *Kysely*, 2008; *Stedinger et al.*, 2008].

137 Simulation-based studies are hence useful, even necessary, to verify the internal consistency
138 of a given FA implementation and to provide information about its main strengths and
139 weaknesses. Indeed, a FA implementation performing poorly with synthetic data is unlikely to
140 become highly capable with real data. Similarly, a FA implementation showing little

141 robustness with slight departures from its underlying assumptions should be considered with
142 caution, since real data are unlikely to perfectly fulfill these assumptions.

143 However, good/better performance of a FA implementation with synthetic data is indicative
144 only, but not conclusive, about its performance in practice. Indeed, determining whether the
145 simulation setup is realistic enough to ensure that the good/better performances of a given FA
146 implementation will also hold in real life is difficult. This is especially the case when FA
147 implementations from distinct families are to be compared (requirement [R1]): for instance,
148 deriving a simulation setup where local, regional and model-based FA implementations could
149 be compared in a fair way is far from obvious.

150 **2.2. Data-based comparisons**

151 Data-based comparisons can complement simulation studies. Indeed, by using real data, they
152 circumvent the difficulty of building realistic simulation setups. However, the main difficulty
153 is that the truth is unknown, thus precluding the use of formal statistical criteria like bias or
154 RMSE. Specific comparison schemes are therefore required. Data-based comparisons are
155 mainly implemented using statistical tests and split-sample validation.

156 **2.2.1. Statistical tests**

157 A statistical test is used to evaluate whether observations can be considered as realizations
158 from the assumed distribution family [e.g., *Chowdhury et al.*, 1991; *Laio*, 2004]. We stress
159 that while this is an important question, choosing a distribution family is not the final
160 objective of frequency analysis: indeed, even if the parent distribution family were known, the
161 *estimated* distribution used for decision and design would still be affected by estimation
162 errors, thus requiring further evaluation of its performance.

163 Statistical tests are hence useful to reject FA implementations that cannot be statistically
164 reconciled with observations. Unfortunately, as noted by *Bobee et al.* [1993], such tests are
165 not powerful with the typical sample size available for environmental data (usually hardly
166 exceeding 50 elements). Consequently, it is often observed that several competing FA
167 implementations cannot be rejected [*Laio*, 2004]. Again, this calls for alternative comparison
168 approaches to attempt further distinguishing between such FA implementations.

169 Another difficulty is that statistical tests are simply not available for many FA
170 implementations. This is problematic when FA implementations from distinct families are to
171 be compared (requirement [R1]), since in general tests will be available for only a few of
172 them. General-purpose testing procedures do exist [e.g. Cramer–von Mises or Anderson-

173 Darling tests, see *Stephens*, 1974], but in their standard form they compare observations with
174 a fully specified distribution: this is not a realistic setting in frequency analysis where
175 parameters are unknown and need to be inferred. Applying these tests in their standard form
176 would systematically favor over-parameterized implementations. Consequently, specific
177 corrections need to be implemented to account for estimation uncertainty, which is not an
178 obvious task. As an illustration, *Laio* [2004] derived tests customized to extreme value
179 distributions, but these tests are only applicable with particular estimators.

180 **2.2.2. Split-sample evaluation**

181 Another data-based comparison approach is based on the splitting of observations into a
182 calibration (or estimation/training) set and a validation (or testing) set [*Gunasekara and*
183 *Cunnane*, 1992]. This approach distinguishes between the descriptive and predictive abilities
184 of FA implementations, which are two fundamentally distinct properties. The former refers to
185 the ability of a FA implementation to *describe* past events used for parameter estimation (i.e.,
186 calibration events), whereas the latter refers to the ability to *predict* new events (i.e. validation
187 events). While a FA implementation of poor descriptive ability has slim chance to become
188 highly capable in predictive mode, the contrary is not true: a FA implementation that can
189 provide a good description of calibration data may become inefficient in predictive mode.

190 Split-sample procedures have been mainly implemented for comparing regional FA
191 implementations [e.g., *GREHYS*, 1996; *Grover et al.*, 2002; *Ouarda et al.*, 2006; *Neppel et*
192 *al.*, 2007; *Szolgay et al.*, 2009]. The evaluation is usually achieved by comparing quantiles
193 computed from validation sites (generally using an at-site estimate based on a long series) and
194 quantiles given by the regional FA implementation (ignoring data at the validation site).
195 Standard measures like bias or RMSE can then be used by considering locally-estimated
196 quantiles as surrogate for the unknown true quantiles. While this may be acceptable for
197 moderate quantiles when the record length at the validation site is large, it might become
198 unrealistic for larger quantiles, which are affected by significant sampling errors. Further
199 refinements of this general approach have been proposed [in particular, see the Bootstrap-
200 based scheme implemented by *GREHYS*, 1996].

201 Split-sample comparisons have also been attempted and discussed for local FA
202 implementations [see in particular *Beard*, 1974; *Interagency Advisory Committee on Water*
203 *Data*, 1982; *Gunasekara and Cunnane*, 1992; *Garavaglia et al.*, 2011], but far less frequently

204 than for regional FA. This is because each series has to be decomposed into calibration and
205 validation periods for local FA implementations, which requires using very long series.

206 Split-sample procedures are of interest because they compare FA implementations in the
207 context they are designed for, where the objective is to predict upcoming events (“How
208 should a dam be designed to ensure that it will withstand *upcoming* floods?”), as opposed to
209 describe past event (“How should a dam be designed to ensure that it would have withstood
210 *observed* floods?”). Moreover, they use FA implementations in operational-like conditions,
211 where both model errors (i.e. misspecified distribution) and estimation errors coexist. We
212 stress the difference between this objective and the objective behind statistical tests
213 (identifying the parent distribution, or at least rejecting inappropriate ones).

214 Unfortunately, split-sample procedures are challenging to apply for two main reasons: (i) as in
215 any data-based procedure, the truth is unknown; (ii) they require a large amount of data to be
216 of any practical interest.

217 **3. A data-based comparison framework**

218 **3.1. Notation and basic hypotheses**

219 The data-based framework described in this section follows the path of split-sampling
220 evaluation as described in previous section 2.2.2. Let X be the variable whose distribution is
221 sought. It is assumed that a (large) dataset of observations from X is available, denoted by

222 $\mathbf{x} = \left(x_k^{(i)} \right)_{i=1:N_{site}, k=1:n^{(i)}}$. The superscript $^{(i)}$ denotes the site, the subscript k denotes the time step.

223 Note that the number of observations at each site does not need to be identical. Using a
224 similar notation, we denote by \mathbf{c} the subset of \mathbf{x} used for calibration, and \mathbf{v} the complementary
225 subset used for validation. In cases where no distinction is needed, we use the generic
226 notation \mathbf{d} to denote any one of \mathbf{c} or \mathbf{v} .

227 The cumulative distribution function (cdf) of the unknown parent distribution of X at site i is
228 denoted by $F^{(i)}$. A given FA implementation M makes an assumption on the distribution of X ,
229 yielding a cdf $F_M^{(i)}(y | \boldsymbol{\theta})$. In this notation, y is the value at which the cdf is evaluated and $\boldsymbol{\theta}$
230 represents a vector of unknown parameters. In most cases the distributional assumption is
231 explicit, but it may also be implicit in the case of model-based implementations (e.g. for
232 floods $F_M^{(i)}(y | \boldsymbol{\theta})$ would result from the rainfall-runoff transformation encapsulated in the
233 hydrologic model). Parameter estimation is then performed by the FA implementation,
234 yielding a particular parameter value $\hat{\boldsymbol{\theta}}$. The estimated distribution is then defined by

235 $\hat{F}_M^{(i)}(y) = F_M^{(i)}(y | \hat{\theta})$. We stress the naming and notational distinction that will be consistently
236 used throughout this paper between the *parent* distribution ($F^{(i)}$, the unknown distribution
237 that generated observations), the *assumed* distribution ($F_M^{(i)}$, with unknown parameters) and
238 the *estimated* distribution ($\hat{F}_M^{(i)}$ corresponding to the assumed distribution with one particular
239 parameter value).

240 The performance indices defined in the next sections can be used for comparison under the
241 following minimal hypotheses:

242 [H1] “Extremes” correspond to large values.

243 [H2] At-site data are temporally independent.

244 Assumption [H1] states that large return periods are associated with large values which is the
245 case for most environmental variables (e.g. flood, wind, precipitation, etc.). If this assumption
246 does not hold (e.g. low flow analysis with annual minimum values), all indices can be readily
247 modified to account for extremes in the left tail of the distribution.

248 Assumption [H2] is more stringent: while the assumption of serial independence can be
249 deemed acceptable for variables related to extreme localized events (e.g. storm winds, heavy
250 rainfalls, floods, [Pujol *et al.*, 2007]), other variables sometimes exhibit significant serial
251 dependence [e.g. Hamed and Rao, 1998; Cohn and Lins, 2005; Koutsoyiannis, 2010]. A
252 detailed analysis of the effect of serial dependence on the comparison framework lies well
253 beyond the scope of this paper. It is therefore assumed that data can be considered as serially
254 independent, either because physical or empirical evidence suggests so or thanks to some data
255 pre-processing (e.g. data sub-sampling).

256 **3.2. Performance indices**

257 **3.2.1. Reliability and stability**

258 The performance of competing FA implementations is judged according to two criteria:
259 reliability and stability [Garavaglia *et al.*, 2011]. A reliable FA implementation yields an
260 estimated distribution close to the (unknown) parent distribution, or in other words, it is able
261 to assign correct exceedance probabilities. In practice, since the parent distribution is
262 unknown, reliability has to be evaluated using observed data.

263 The stability of a FA implementation describes its ability to yield similar estimates when
264 different data are used for calibration. In an industrial context, stable estimates are sought

265 when a whole group of structures is to be designed (e.g. power plants or dams fleet). Indeed,
266 quantile estimates strongly varying with new observations would result in a frequent
267 questioning of the design, which is problematic since a built structure cannot be continuously
268 modified to track estimates' variability. Moreover, unstable estimates might cause the actual
269 protection level to differ strongly from e.g. dam to dam, even if all dams are designed with the
270 same target protection level.

271 It is stressed that both criteria do not play the same role in judging the performance of a FA
272 implementation. In particular, stability cannot be used alone, because it does not give any
273 information about the ability to predict observations (a FA implementation can be stable but
274 totally unreliable). Consequently, reliability is assessed first in the comparison framework.
275 When several FA implementations appear equally reliable, the additional insights provided by
276 stability can be used to further discriminate between them.

277 **3.2.2. Reliability: pval**

278 This first reliability index aims to evaluate the overall agreement between the estimated
279 distribution $\hat{F}_M^{(i)}$ and observations $d_k^{(i)}$ (either calibration or validation data can be used). For
280 a given site i and time step k , it is defined as follows:

$$pval_k^{(i)} = \hat{F}_M^{(i)}(d_k^{(i)}) \quad (1)$$

281 Under the assumption that the estimation is reliable ($\hat{F}_M^{(i)} = F^{(i)} \forall i$), $pval_k^{(i)}$ are realizations
282 from a uniform distribution on each site i : $pval_k^{(i)} \sim U[0;1] \forall i$ (see Appendix 1). Graphical
283 diagnostics to assess the agreement between observed $(pval_k^{(i)})_{k=1:n^{(i)}}$ and their theoretical
284 distribution under the reliability hypothesis will be described in subsequent section 3.2.5.

285 The reliability assumption ($\hat{F}_M^{(i)} = F^{(i)} \forall i$) is worth commenting. It is quite clear that it will
286 never be strictly met because of model and estimation errors. However, we use it as a working
287 assumption, and we are looking in the data for evidence conflicting with it (which would
288 materialize in the case of index $pval$ by non-uniformly distributed values). This is the same
289 rationale than that behind the use of a H_0 hypothesis in statistical testing. However, we are
290 only performing graphical diagnostics derived over an ensemble of sites here. While this
291 allows making comparative statements on the relative reliability of FA implementations, it

292 does not provide a formal decision rule to reject the reliability assumption, as a statistical test
 293 would. The reason for this is discussed in subsequent section 5.3.

294 **3.2.3. Reliability: N_T**

295 The second reliability index is based on the number of exceedances of an estimated T -year
 296 quantile [e.g. *Interagency Advisory Committee on Water Data*, 1982, Appendix 14;
 297 *Gunasekara and Cunnane*, 1992; *Garavaglia et al.*, 2010]:

$$N_T^{(i)} = \sum_{k=1}^{n^{(i)}} 1_{\{\hat{q}_T^{(i)}; +\infty\}}(d_k^{(i)}) \quad (2)$$

Where $1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$

298 Under the reliability assumption ($\hat{q}_T^{(i)} = q_T^{(i)}$), $N_T^{(i)}$ is a realization from the binomial
 299 distribution: $N_T^{(i)} \sim \text{Bin}(n^{(i)}, 1/T)$ (see Appendix 1). As previously, dedicated graphical
 300 diagnostics will be described in subsequent section 3.2.5. Contrarily to index *pval* which
 301 quantifies the overall reliability, N_T focuses on reliability for prescribed T -year quantiles.

302 **3.2.4. Reliability: FF**

303 The index FF , used by e.g. *England et al.* [2003a] and *Garavaglia et al.* [2011], corresponds
 304 to the index *pval* computed on the maximum observed value of each site, $d_{\max}^{(i)}$:

$$FF^{(i)} = \hat{F}_M^{(i)}(d_{\max}^{(i)}) \quad (3)$$

305 Under the reliability assumption ($\hat{F}_M^{(i)} = F^{(i)}$), $FF^{(i)}$ is a realization from a Kumaraswamy
 306 distribution with parameters $(n^{(i)}; 1)$: $FF^{(i)} \sim K[n^{(i)}; 1]$, whose cdf can be written as follows
 307 (see Appendix 1):

$$F_K(t) = t^{n^{(i)}}, 0 \leq t \leq 1 \quad (4)$$

308 **3.2.5. Graphical diagnostics based on reliability indices**

309 Graphical diagnostics of reliability are based on the comparison between the reliability
 310 indices and their theoretical distribution under the reliability assumption. For a given site i
 311 with $n^{(i)}$ observations, let $z^{(i)}$ be any one of the indices defined in sections 3.2.2-3.2.4 (e.g.
 312 FF), and $H^{(i)}$ the cdf of its theoretical distribution under the reliability assumption (e.g. cdf of

313 a Kumaraswamy distribution $K(n^{(i)};1)$). A technical difficulty arises because $H^{(i)}$ depends on
314 the number of observations $n^{(i)}$, which in general varies from site to site.

315 This issue can easily be overcome in the case of indices having a continuous cdf (namely,
316 $pval$ and FF) by using a probability-probability plot (pp-plot) representation: probability-
317 transformed indices $H^{(i)}(z^{(i)})$ are plotted against empirical frequencies (see Figure 1a-c for
318 illustrations). Under the reliability hypothesis, the probability-transformed values $H^{(i)}(z^{(i)})$ are
319 indeed uniformly distributed between 0 and 1, irrespective of the number of observations $n^{(i)}$.
320 Departures from the diagonal in the pp-plot have specific interpretations in terms of
321 under/over-estimation or predictive failures (see Figure 1). Moreover, additional axis
322 transformations might be valuable to focus on particular areas of the pp-plot. Typically, the
323 pp-plot can be transformed into a Gumbel quantile-quantile plot (qq-plot) by applying the
324 Gumbel quantile function to each axis (see Figure 1d-f for illustrations). This allows focusing
325 on extreme values of indices. Note that any other continuous quantile function can be used,
326 depending on the area of interest in the pp-plot.

327 The case of the discrete index N_T is more problematic, because the cdf of its theoretical
328 binomial distribution is not continuous. Probability-transformed indices $H^{(i)}(N_T^{(i)})$ are
329 therefore not uniformly distributed. A possibility to overcome this difficulty is to randomize
330 the values $H^{(i)}(N_T^{(i)})$ in order to discard the cdf discontinuity induced by the discrete nature
331 of the index N_T . This randomization is performed as follows. Let $b(-1) = 0$ and
332 $b(j) = H^{(i)}(j) = \Pr(N \leq j)$, $j \geq 0$, where N is a random variable following a $Bin(n^{(i)}, 1/T)$
333 distribution. At a given site i , the value $N_T^{(i)}$ is transformed into probability space by
334 randomly sampling a value $w^{(i)}$ from an uniform distribution between $b(N_T^{(i)} - 1)$ and
335 $b(N_T^{(i)})$. This is to be compared with the non-randomized probability transformation, which
336 corresponds to setting $w^{(i)} = b(N_T^{(i)})$. This randomization ensures that the values $w^{(i)}$ are
337 uniformly distributed between 0 and 1 under the reliability hypothesis (see Appendix 1). It is
338 then possible to use the same pp-plot and qq-plot representations as discussed for continuous
339 indices (see Figure 4d-e for illustrations).

340 **3.2.6. Stability: $SPAN_T$**

341 The stability of quantile estimates can be quantified by contrasting the values obtained with
342 two different calibration datasets c_1 and c_2 . The index $SPAN_T$ proposed by *Garavaglia et al.*

343 [2011] is used in this paper. It is a measure of the relative deviation between the two
344 estimated T -year quantiles. Let $\hat{q}_T^{(i)}$ denotes the T -year quantile at site i , derived from the
345 estimated distribution $\hat{F}_M^{(i)}$. For a given site i , $SPAN_T$ is defined as follows:

$$SPAN_T^{(i)} = \frac{|\hat{q}_T^{(i)}(\mathbf{c}_1) - \hat{q}_T^{(i)}(\mathbf{c}_2)|}{\frac{1}{2}(\hat{q}_T^{(i)}(\mathbf{c}_1) + \hat{q}_T^{(i)}(\mathbf{c}_2))} \quad (5)$$

346 The comparison between competing FA implementations can then be performed by
347 comparing the distribution of $SPAN_T^{(i)}$ over all sites $i = 1:N_{site}$: the FA implementation whose
348 $SPAN_T$ distribution remains the closest to zero is the most stable.

349 **3.3. Comparing uncertainties**

350 **3.3.1. Motivation**

351 One of the requirements for the comparison framework is to enable the comparison of
352 estimated uncertainties. The term “*estimated uncertainties*” aims to emphasize the fact that
353 uncertainty quantification depends on the assumptions underlying the FA implementation
354 (e.g. distribution family, estimation approach, etc.). Consequently, there is a distinct
355 possibility that such estimated uncertainties are unreliable if those assumptions are unrealistic
356 [see e.g. the discussion by *Daly*, 2006 in the context of spatial interpolation methods].

357 Evaluating uncertainty estimates cannot be performed by counting the percentage of points
358 inside a $\alpha\%$ confidence interval in Figure 2, because the values on the x -axis are based on
359 estimates of the exceedance probability (by means of a plotting position formula). As such,
360 those values are affected by considerable uncertainties. The approach taken in this paper to
361 circumvent this difficulty is to transform the uncertainty intervals shown in Figure 2 into a
362 new distribution, named the predictive distribution.

363 This tool is well-known and widely used in Bayesian statistics [e.g. *Gelman et al.*, 1995] and
364 hydrologic modeling [e.g. *Todini and Mantovan*, 2007; *Thyer et al.*, 2009; *Renard et al.*,
365 2010], and has also been proposed for FA applications [*Coles*, 2001, chapter 9; *Cox et al.*,
366 2002; *Meylan et al.*, 2008, chapter 7]. Moreover, the notion of “*expected probability*”
367 discussed by e.g. *Stedinger* [1983a], *Rosbjerg and Madsen* [1998] or *Kuczera* [1999] is
368 conceptually related to the notion of predictive distribution. The main advantage of this
369 approach is that the methodology used to compare estimated distributions (red line in Figure
370 2) can be applied to predictive distributions, hence indirectly comparing estimated

371 uncertainties. This section defines the predictive distribution in both Bayesian and non-
372 Bayesian contexts.

373 **3.3.2. The Bayesian predictive distribution**

374 Following the notation introduced in section 3.1, we use $f_M(y|\theta)$ and $\hat{f}_M(y)$ to denote the
375 probability density function (pdf) of assumed and estimated distributions, respectively.

376 In Bayesian statistics, parameter inference is performed using the posterior distribution
377 $p_M(\theta|\mathbf{c})$, where \mathbf{c} represents the calibration data. The predictive distribution of a future
378 observable y given observed data \mathbf{c} is defined by the following pdf [e.g. *Gelman et al.*, 1995]:

$$\hat{\pi}_M(y) = p_M(y|\mathbf{c}) = \int f_M(y|\theta)p_M(\theta|\mathbf{c})d\theta \quad (6)$$

379 The predictive distribution $\hat{\pi}_M(y)$ hence corresponds to integrating the assumed distribution
380 $f_M(y|\theta)$ over the posterior distribution of θ , $p_M(\theta|\mathbf{c})$, which represents the uncertainty in
381 estimating θ . By contrast, the estimated pdf $\hat{f}_M(y)$ corresponds to using the assumed
382 distribution $f_M(y|\theta)$ for a fixed value $\hat{\theta}$ of its parameters (most commonly the posterior
383 mean, median or mode), hence ignoring estimation uncertainty. Figure 2 illustrates the
384 difference between the predictive distribution $\hat{\pi}_M(y)$ and the estimated distribution $\hat{f}_M(y)$,
385 and compares these distributions with the uncertainty bounds.

386 In practice, the integration in equation (6) cannot be performed analytically in general and has
387 to be approximated numerically. Given that the posterior distribution $p_M(\theta|\mathbf{c})$ is often
388 explored using Markov Chain Monte Carlo (MCMC) samplers, such numerical
389 approximation is usually implemented using a Monte Carlo scheme (see Appendix 2).

390 **3.3.3. Non-Bayesian predictive distributions**

391 The predictive distribution in equation (6) is not defined in a non-Bayesian context, because
392 the posterior distribution $p_M(\theta|\mathbf{c})$ does not exist in frequentist statistics, where θ is
393 considered as a non-random quantity. However, the estimator of θ , noted $\hat{\theta}(X)$, is a random
394 variable and its distribution is defined - it is the sampling distribution of the estimator. Note
395 the distinction between the (random) estimator $\hat{\theta}(X)$ and the (non-random) estimated value
396 $\hat{\theta} = \hat{\theta}(\mathbf{c})$ corresponding to the value taken by the estimator on the calibration sample.

397 The question of deriving a non-Bayesian version of the predictive distribution in equation (6)
398 has attracted a lot of attention amongst statisticians. This has led to the development of
399 innovative (sometimes controversial) inference paradigms, in particular pivotal inference and
400 fiducial probabilities [Fisher, 1930; Dawid and Stone, 1982; Seidenfeld, 1992; Dawid and
401 Wang, 1993; Barnard, 1995; Wang, 2000; Lawless and Fredette, 2005; Hannig et al., 2006],
402 predictive likelihoods [Hinkley, 1979; Butler, 1986; Bjornstad, 1990] and H-likelihoods [Lee
403 and Nelder, 1996; Meng, 2009].

404 Harris [1989] proposed a pragmatic approach: the posterior distribution in equation (6) is
405 simply replaced by the sampling distribution of $\hat{\theta}(X)$. Let $s_M(\boldsymbol{\tau}|\boldsymbol{\theta})$ denote the pdf of this
406 sampling distribution evaluated at $\boldsymbol{\tau}$. Note that in general, the sampling distribution depends
407 on the unknown true parameter value $\boldsymbol{\theta}$. A non-Bayesian version of equation (6) is then:

$$\pi_M^*(y|\boldsymbol{\theta}) = \int f_M(y|\boldsymbol{\tau})s_M(\boldsymbol{\tau}|\boldsymbol{\theta})d\boldsymbol{\tau} \quad (7)$$

408 Compared with equation (6), there is an additional difficulty in equation (7) since the true
409 value $\boldsymbol{\theta}$ is still unknown. Harris' proposal is to replace the unknown $\boldsymbol{\theta}$ by its estimated
410 value $\hat{\boldsymbol{\theta}}$, yielding the following predictive distribution:

$$\hat{\pi}_M(y) = \pi_M^*(y|\hat{\boldsymbol{\theta}}) = \int f_M(y|\boldsymbol{\tau})s_M(\boldsymbol{\tau}|\hat{\boldsymbol{\theta}})d\boldsymbol{\tau} \quad (8)$$

411 Replacing the unknown true value $\boldsymbol{\theta}$ by its estimated value $\hat{\boldsymbol{\theta}}$ is a standard practice when
412 estimating a sampling distribution. This is akin to the Fisher information matrix being
413 replaced by the observed information matrix in ML estimation [e.g. Coles, 2001].

414 The predictive distribution in equation (8) was named the “parametric bootstrap predictive
415 distribution” by Harris [1989], and has been further developed by other authors [e.g., Basu
416 and Harris, 1994; Vidoni, 1995; Fushiki et al., 2005; Fushiki, 2010]. A similar approach
417 termed “bagging predictors” [e.g. Breiman, 1996] is used in the field of machine learning.

418 Similarly to the Bayesian predictive distribution, the integration in equation (8) in general is
419 not performed analytically. Simple algorithms to derive non-Bayesian predictive distributions
420 are described in Appendix 2. It is worth noting that deriving the predictive distribution only
421 requires minimal effort beyond that made to quantify uncertainties.

422 **3.3.4. Indirectly comparing uncertainties via predictive distributions**

423 The comparison of estimated uncertainties is then performed by replacing the cdf of the
424 estimated distribution ($\hat{F}_M^{(i)}$) by the cdf of the predictive distribution $\hat{\Pi}_M(y)$ for all indices in
425 section 3.2. The rationale behind this indirect approach is the following: if implementation A
426 yields a more reliable predictive distribution than implementation B (according to the indices
427 of section 3.2), it suggests that implementation A yields a more reasonable quantification of
428 uncertainties in the sense that after transformation into a predictive distribution (eq. (6)-(8)),
429 these uncertainties are in better agreement with validation data.

430 Throughout the remainder of this paper, we will simply use the naming “predictive
431 distribution” with no further distinction between the Bayesian and the non-Bayesian versions.
432 Indeed, while this distinction is necessary to introduce formal definitions, it is of little
433 relevance in the context of the comparison framework discussed here.

434 **4. Case study**

435 The comparison framework described in previous sections is applied to a large runoff dataset.
436 Ten FA implementations, belonging to three FA families, are compared. These
437 implementations do not constitute an exhaustive representation of existing FA
438 implementations, since the objective of this case study is not to draw definitive conclusions
439 on the merits of existing FA implementations. Instead, it aims at illustrating the application of
440 the performance indices described in section 3, and discussing the insights that can be gained
441 from the application of a data-based comparison exercise.

442 **4.1. Data and FA implementations**

443 Daily runoff series from 364 stations in France are used (Figure 3), corresponding to
444 catchment sizes ranging from 10 to 2,000 km². The time series cover at least 20 years, with
445 more than 200 series spanning over 40 years. The quality of this dataset and its suitability for
446 flood FA has been thoroughly evaluated in previous work [Renard et al., 2008].

447 Annual maxima (AM) are extracted from the daily series. AM values are then treated with 10
448 FA implementations, belonging to three FA families, as summarized in Table 1:

449 1. Local estimation family: six implementations, corresponding to two distributional
450 assumptions (Gumbel (GUM) and Generalized Extreme Value (GEV)) and three
451 parameter estimation methods (Moments (MOM), Maximum Likelihood (ML) and
452 Bayesian (BAY)), are used. The three estimation methods differ in their quantification of

453 uncertainty: (i) a non-parametric bootstrap approach is used for MOM; (ii) a standard
454 Gaussian approximation for the sampling distribution of ML estimators is used; (iii) the
455 posterior distribution of parameters represents the uncertainty in the Bayesian approach.
456 For the latter approach, flat priors are used for location and scale parameters (i.e.,
457 $\pi(\theta) \propto 1$), while an Gaussian prior with mean 0 and standard deviation 0.2 is used for the
458 shape parameter of the GEV distribution.

459 2. Regional estimation family: 2 implementations, corresponding to two distributional
460 assumptions (GUM and GEV), are used. A standard index flood scheme [e.g. *Dalrymple*,
461 1960; *Robson and Reed*, 1999] is used: on the one hand, a regression between the index
462 flood (taken here as the at-site mean) and catchment descriptors is built. On the other
463 hand, a regional distribution is estimated by pooling standardized data (i.e. AM values
464 divided by the index flood) from all sites together. Using the index flood regression
465 together with the regional distribution enables estimating the distribution of AM at any
466 site, including ungauged ones (see Appendix 3 for additional details).

467 3. Local-Regional estimation family: two implementations, corresponding to two
468 distributional assumptions (GUM and GEV), are used. These implementations aim at
469 using both the regional models above and the data observed at the target sites. The
470 Bayesian approach proposed by *Ribatet et al.* [2006] is used: at each target site, the
471 prediction by the regional model is used to define the prior distribution, while at-site data
472 are used to build the likelihood function. The resulting posterior distribution therefore
473 combines local and regional information (see Appendix 3 for details).

474 Note that the ten implementations analyzed in this case study correspond to fairly standard
475 approaches, rather than state-of-the-art methods. Additional implementations could be
476 considered to improve some aspects of the implementations described above. In particular,
477 more advanced regionalization procedures could be investigated [e.g. *Madsen and Rosbjerg*,
478 1997; *Reis et al.*, 2005; *Micevski et al.*, 2006b; *Renard*, 2011]. However, we stress that the
479 objective of this case study is not to provide the best possible estimation of flood quantiles in
480 this particular area, but rather to illustrate the application of the comparison framework to
481 standard FA implementations.

482 **4.2. Reliability and stability decompositions**

483 In order to assess the reliability of the FA implementations, the 364 sites are split into
484 calibration and validation sets as follows:

- 485 • Sites with less than 40 years of data are used for calibration of the regional models (160
486 sites, red and pink dots in Figure 3a).
- 487 • For each site with more than 40 years of data (204 sites, black dots in Figure 3a), 20 years
488 are randomly selected (independently from site to site) for calibration of the local models.
- 489 • For the latter sites, all remaining years (i.e. at least 20 years for each black dot in Figure
490 3a) are used for validation.
- 491 This decomposition allows comparing the reliability of all FA implementations based on
492 exactly the same validation data.
- 493 Additional decompositions are required to assess stability. Since both local and regional
494 implementations are considered, two types of decomposition are proposed:
- 495 • Stability with respect to local information (type I): for each site with more than 40 years
496 of data (black dots in Figure 3a), two 20-year calibration sets are randomly selected.
497 Purely regional implementations will be insensitive to this decomposition, since they do
498 not use local information.
- 499 • Stability with respect to regional information (type II): sites with less than 40 years of data
500 are split into two calibration sets (red and pink dots in Figure 3a). Purely local
501 implementations will be insensitive to this decomposition.

502 **4.3. Results**

503 **4.3.1. Illustration of the reliability diagnostics for one particular** 504 **implementation**

505 In order to illustrate the derivation of the graphical diagnostics of section 3.2.5, reliability is
506 first evaluated for the sole implementation GEV_ML (the estimated distribution is used here).
507 Figure 4a shows the pp-plot of $pval$ for validation data, with each gray line corresponding to a
508 validation site. Overall, the pp-curves are evenly distributed around the diagonal control line,
509 and remain fairly close to it in most cases. However, the $pval$ index only assesses the overall
510 reliability, without particular focus on extremes. More stringent diagnostics are hence
511 required to assess reliability at higher levels.

512 To this aim, Figure 4b shows the pp-plot of FF , for both calibration (blue) and validation
513 (red) data. The S-shaped calibration curve indicates that the observed distribution of FF
514 values is *less* variable than it would be if the parent distribution were used. This is an effect of
515 errors in estimating GEV parameters, whose optimization tend to “over-fit” calibration data.

516 At the opposite, the shape of the validation curve indicates that the distribution of FF values
517 is *more* variable than it would be with the parent distribution. This implies that validation data
518 are too often considered as “extreme” by the model, yielding high/low FF values with an
519 unduly large frequency. In particular, a remarkable feature of this curve is its tendency to be
520 stacked against the right border in the upper right corner: this corresponds to numerous FF
521 values having p -values close to or equal to one, i.e. to observations that are considered as
522 impossible by the model. This is a consequence of estimation errors for the shape parameter,
523 yielding right-bounded GEV distributions whose bound is exceeded by validation data.

524 As suggested in section 3.2.5, an axis transformation can be used to focus on this area of the
525 plot. Figure 4c therefore shows the same curves after transforming both axes into a Gumbel
526 scale. Since this transformation is undefined for FF values equal to one, the corresponding
527 points do not appear in the figure, but their percentage is reported. The large departure from
528 the diagonal appearing in Figure 4c for the validation curve confirms the unduly high
529 frequency of large FF values. In addition, 18% of validation data have a FF value equal to
530 one – in other words, what is considered as impossible by the model actually occurs for 18%
531 of the sites. This corresponds to severe prediction failures.

532 The second row of Figure 4 shows graphical diagnostics related to the N_{10} index. Figure 4d
533 shows the N_{10} pp-plot after the randomization procedure described in section 3.2.5, while
534 Figure 4e shows the qq-plot version of this diagnostic in Gumbel axes. These figures yield
535 similar conclusions to the corresponding FF diagnostics.

536 The opposite behavior of calibration and validation curves is an illustration of the trade-off
537 between descriptive and predictive capability: a too good fit to calibration data may come at
538 the price of a reduced predictive reliability. In turn, this reemphasizes the necessity to assess
539 predictive performances based on validation data. Consequently, all reliability diagnostics
540 will focus on validation data in the remainder of this paper.

541 **4.3.2. Comparison of estimation methods for local FA** 542 **implementations**

543 This section compares the three estimation methods (MOM, ML and BAY) used for local FA
544 implementations. Figure 5 shows the reliability diagnostics for the estimated (first row, the
545 posterior mode is used as parameter estimates) and the predictive (second row) distributions.
546 For brevity, only the qq-plot representations in Gumbel space are reported, since it allows
547 focusing on the most severe prediction failures.

548 The *FF* diagnostic in Figure 5a indicates that the estimation method has little impact on
549 reliability, compared to the choice of the distribution (GUM or GEV). Moreover, departures
550 from the diagonal are smaller for the three GUM curves than for the three GEV curves. In
551 addition, for MOM and ML estimation, validation data are considered as impossible by the
552 GEV prediction for more than 15% of the sites. This percentage drops to 7% for BAY, which
553 is a consequence of using an informative prior to constrain the shape parameter. The N_{10} and
554 N_{100} diagnostics in Figure 5b-c yield similar insights. These results indicate that at-site
555 estimation of a GEV distribution with 20 years of data may lead to substantial predictive
556 failures, whatever the estimation method. This is a consequence of the well-documented
557 difficulty in precisely identifying the shape parameter [e.g. *Coles*, 2001; *Garavaglia et al.*,
558 2011]. However, we stress that this does not imply that the GEV distribution should be
559 rejected, but rather that local estimation with moderate sample size is not precise enough for
560 this distribution to yield reliable predictions. In turn, this indicates that ignoring estimation
561 uncertainty is not a viable option for locally-estimated GEV distributions.

562 The second row in Figure 5 shows the same diagnostics applied to the predictive distribution
563 rather than the estimated distribution. The *FF* diagnostic in Figure 5d indicates that the
564 estimation method (MOM, ML or BAY) has little impact for the Gumbel distribution: all
565 three curves are similar and show marked departures below the diagonal. This suggests that
566 even after accounting for uncertainty, a tendency to under-estimation remains with a Gumbel
567 distribution. On the other hand, the estimation method has a stronger impact for the GEV
568 distribution: the GEV_MOM curve shows the largest departure below the diagonal. Departure
569 for the GEV_ML curve is similar although less pronounced. Lastly, the GEV_BAY curve
570 appears much closer to the diagonal, suggesting that once uncertainties are accounted for, the
571 predictions become fairly reliable. The N_{10} and N_{100} diagnostics in Figure 5e-f yield similar
572 conclusions, although they reveal a more pronounced impact of the estimation method for the
573 Gumbel distribution.

574 Overall, the results of this section suggest that while the estimation method does not strongly
575 impact reliability based on the estimated distributions, the method used to quantify
576 uncertainty exerts a stronger leverage according to the predictive distribution.

577 **4.3.3. Comparison of local, regional and local-regional** 578 **implementations**

579 This section compares the three FA families (local, regional and local-regional) for both the
580 Gumbel and the GEV distributions. For simplicity, only implementations GUM_BAY and
581 GEV_BAY are used within the local family. Figure 6 shows the *pval* diagnostic for the
582 estimated distribution, and for the three implementations involving the GEV distribution
583 (similar plots are obtained with the Gumbel distribution, not shown). While local
584 (GEV_BAY) and local-regional (GEV_LR) implementations yield similar diagnostics, the
585 regional implementation (GEV_REG) shows marked departures from the diagonal for many
586 sites. Such departures are more often below the diagonal (under-estimation) than above. Since
587 the *pval* diagnostic does not focus on extremes, this indicates that the regional predictions
588 may be markedly unreliable, even for small to moderate quantiles.

589 Figure 7 shows *FF*, N_{10} and N_{100} reliability diagnostics for the estimated distribution (first
590 row) and the predictive distribution (second row). In Figure 7a (*FF*), the smallest departure
591 from the diagonal corresponds to the GEV_LR implementation, while larger departures are
592 observed for both regional implementations and the local GEV_BAY implementation. Figure
593 7b (N_{10}) highlights a clear distinction between regional implementations on the one hand, and
594 local and local-regional implementations on the other hand. The former show a poor
595 reliability even for predicting moderate 10-year quantiles, which confirms previous findings
596 based on *pval*. On the other hand, local and local-regional implementations show similar
597 predictive reliability for this index. However, the N_{100} index (Figure 7c) indicates that local-
598 regional implementations become more reliable than local ones to predict larger quantiles.
599 This suggests that while local approaches may be sufficient to estimate moderate quantiles,
600 they become less reliable than local-regional approaches when extrapolated to higher
601 quantiles, especially if a GEV distribution is used.

602 The second row of Figure 7 shows the same diagnostics applied to the predictive distribution.
603 Figure 7d (*FF*) does not reveal marked differences between implementations, apart from
604 GEV_BAY whose curve is closer to the diagonal as already observed in section 4.3.2. The
605 N_{10} and N_{100} diagnostics (Figure 7e-f) yield more insights: in both cases, regional
606 implementations show large departures from the diagonal, which suggests an unreliable
607 quantification of uncertainty. On the other hand, local and local-regional implementations
608 have similar curves for both indices, with smaller departures from the diagonal suggesting a
609 more reliable quantification of uncertainty. The behavior of GEV_BAY for indices *FF* and
610 N_{100} is noteworthy: while its predictions based on estimated distributions are unreliable
611 (Figure 7a and c), it still yields fairly reliable predictions once uncertainty is accounted for

612 through the predictive distribution (Figure 7d and f). At the opposite, regional
613 implementations appear unreliable for both estimated and predictive distributions.

614 Lastly, stability is assessed by means of the $SPAN_{100}$ index (Figure 8). Figure 8a compares the
615 type-I stability of estimated (see section 4.2). Local implementations GEV_BAY and
616 GUM_BAY show the lowest stability. On the other hand, both local-regional
617 implementations GEV_LR and GUM_LR have similar stability, and importantly, are more
618 stable than any of the local implementations. Application of the $SPAN_{100}$ index to predictive
619 distributions yield identical insights (Figure 8c). Figure 8b compares the type-II stability of
620 the estimated distributions (see section 4.2). Both regional implementations GEV_REG and
621 GUM_REG show a very low stability, while both local-regional implementations are far
622 more stable. Figure 8d shows a similar pattern for the predictive distribution.

623 Overall, the results of this section suggest that local-regional implementations generally
624 outperform both the purely local and regional implementations they are built upon. This
625 observation holds for both reliability (see e.g. Figure 7c) and stability (Figure 8).

626 **5. Discussion**

627 ***5.1. Ability of reliability and stability indices to benchmark FA*** 628 ***implementations***

629 The case study shows that the indices defined in section 3.2 are able to reveal marked
630 difference between the reliability and stability of competing FA implementations. Moreover,
631 reliability indices appear quite complementary. Index $pval$ is able to reveal reliability failures
632 at moderate levels (e.g. GEV_REG in Figure 6), but will not detect failures specific to
633 extreme levels. Index N_T allows focusing on specific quantiles, and varying the value of T
634 yields insights on the evolution of reliability at increasing levels (see e.g. the evolution of
635 GEV_BAY between N_{10} and N_{100} in Figure 7b-c). Lastly, index FF focuses on the most
636 extreme value observed at each site and is hence the most stringent reliability diagnostic. In
637 particular, it can reveal severe prediction failures, where observations as considered as
638 virtually impossible by the model (e.g. GEV_ML in Figure 4c). However, the set of indices
639 proposed in this paper is not exhaustive and could be completed in future work with
640 additional and possibly more powerful indices. As an illustration, *Garavaglia et al.* [2011]
641 assessed the stability of uncertainty estimates by quantifying the overlapping of confidence

642 intervals obtained using distinct calibration periods. Alternatively, indices based on the
643 duration between exceedances of large quantiles could be derived as an alternative to N_T .

644 **5.2. Feasibility of benchmarking uncertainty estimates**

645 A major objective of this paper was to open uncertainty estimates to the same scrutiny as
646 estimated distributions. This was achieved by transforming these uncertainties into a
647 predictive distribution, which can be scrutinized in the same way as the estimated distribution.
648 The results of the case study confirm that this is a viable approach, with distinct FA
649 implementations showing marked differences in the reliability of their uncertainty estimates.
650 Moreover, these results confirm two important points that are sometimes overlooked: (i)
651 quantifying uncertainty is not sufficient, one also needs to assess whether this quantification is
652 reliable [Hall *et al.*, 2007; Thyer *et al.*, 2009]; (ii) uncertainty estimates derived from a FA
653 implementation whose assumptions are unrealistic are likely to be unreliable, and hence
654 meaningless [Daly, 2006].

655 **5.3. Limitations of the comparison framework**

656 Despite showing its ability to compare FA implementations in terms of stability and
657 reliability, the comparison framework remains based on a few hypotheses that are recalled
658 and discussed in this section.

659 First, it is noted that the proposed framework only yields graphical comparisons between the
660 stability and reliability of competing FA implementations. A natural extension would be to
661 implement formal testing procedures, e.g. to test whether departures from the diagonal in
662 Figure 5 are significant, or whether two FA implementations yield index distributions that are
663 significantly different. Unfortunately, this is a challenging task because indices values are not
664 independent from site to site, due to the spatial dependence between data. Consequently, the
665 development of statistical tests would require a description of this spatial dependence. This
666 was not attempted in this study because it would require making additional assumptions on
667 the structure of spatial dependence beyond that made by the competing FA implementations.

668 Second, it is assumed that the data used for the comparison are temporally independent.
669 Indeed, deriving the distribution of most performance indices (under the reliability
670 hypothesis) requires making this assumption. It may be restrictive in some regions and/or for
671 some hydrologic variables with significant inertia (e.g. low flows or mean annual runoff for
672 groundwater-driven catchments). Future work could therefore evaluate the sensitivity of the
673 comparison framework with temporally dependent data.

674 Lastly, the general philosophy behind the comparison framework involves specific
675 requirements, that are not limitations of the framework itself but might make its application
676 difficult in some contexts. Indeed, applying the comparison framework requires an extended
677 dataset of good-quality long series. The quality of the dataset needs to be thoroughly
678 evaluated to avoid e.g. non-homogeneous data or heavily regulated catchments (see e.g. *Lang*
679 *et al.* [2010] for examples of misleading results caused by the poor quality of a dataset).
680 Moreover, the number of sites needs to be large enough since tools for assessing stability and
681 reliability are based on the distribution of indices over an ensemble of sites. Lastly, long
682 series are also required, since the evaluation of reliability remains limited by the series length:
683 on the one hand, data left out for validation should be numerous to enable a truly challenging
684 assessment of predictive ability; on the other hand, one needs to preserve enough data to
685 calibrate the FA implementation.

686 A consequence of these requirements is that the comparison framework is geared toward large
687 scale, national-wide comparisons rather than smaller-scale studies involving a couple of sites.
688 In particular, the framework cannot compare predictive performance on one particular site.
689 This is an acknowledged limitation, since a FA implementation having the best predictive
690 performance on an ensemble of sites can still fail on one particular site.

691 **5.4. Tailoring and developing comparison schemes**

692 The case study of section 4 is performed at a rather large scale, which may restrict the ability
693 to benchmark FA implementations. Indeed, it is likely that for daily runoff, the “best” FA
694 implementation depends on various catchment properties like catchment size, elevation,
695 climatic area, etc. Consequently, the comparison performed in this case study should be
696 refined at the smaller scale of homogenous hydro-climatic regions. Moreover, the FA
697 implementations compared in this case study are only a small sample of available FA
698 implementations. More precisely, work is currently in progress to extend this comparison to
699 additional distributions (e.g. log-Normal, Pearson family), estimation approaches (e.g. linear
700 moments), approaches to uncertainty quantification (e.g. parametric bootstrap as advocated by
701 *Kysely* [2008]), and alternative regionalization procedures.

702 Lastly, the decomposition into calibration and validation subsets could also be tailored to
703 focus on more specific issues, for instance non-stationarity or low-frequency variability. As
704 an illustration, the decomposition could be stratified according to the value of some climate

705 index (e.g. SOI, NAO) to evaluate the added value of implementations that use climate
706 information.

707 **5.5. Moving toward a systematic approach to comparing FA** 708 **implementations**

709 The data-based comparison approach presented in this paper might be a part of the systematic
710 comparison approach advocated by *Bobee et al* [1993]. However it would not be reasonable
711 to rely on a single comparison framework (let alone on a single comparison metric) to choose
712 between competing FA implementations. We note that there have been controversies on
713 which type of comparison framework should be used (see e.g. [*Wallis and Wood*, 1985;
714 *Beard*, 1987; *Wallis and Wood*, 1987] for data-based vs. simulation studies). However, we
715 claim that the different types of comparison frameworks should not be opposed, but rather be
716 used together since they may actually yield complementary insights. Moreover, concordant
717 results derived from distinct comparison frameworks constitute pieces of evidence that add up
718 to build confidence in their generality [*Gunasekara and Cunnane*, 1992]. As an illustration,
719 most results obtained in this paper are fully consistent with previous simulation-based studies,
720 in particular the poor performance of the GEV distribution with small samples and no prior
721 information [*Martins and Stedinger*, 2000] or the benefit of combining local and regional
722 information [*Stedinger and Lu*, 1995]. The fact that similar findings are found in simulation-
723 based and data-based contexts indicate that they can be extrapolated outside of the particular
724 simulation setups used in the former comparisons.

725 Consequently, a comprehensive comparison of FA implementations might encompass the
726 following steps:

- 727 1. Simulation studies in an “ideal” setup (no model misspecification) are useful to quantify
728 the performance of FA implementations in formal statistical terms (e.g. bias, RMSE,
729 reliable quantification of uncertainty, etc.). Moreover, “non-ideal” setups can be used to
730 assess robustness. FA implementations that grossly fail this simulation step are probably
731 not worth further investigation, but for other implementations, alternative comparison
732 frameworks can provide a complementary point of view on their relative performance.
- 733 2. When available, statistical tests can be used to reject implementations that are in obvious
734 disagreement with observations. An advantage of this comparison approach is that is can
735 be implemented on a site-by-site basis. However, several implementations should be
736 expected to pass the tests given their quite low power with typical sample sizes.

737 3. The data-framework proposed in this paper evaluate the implementations' performance in
738 terms of predictive ability, which closely corresponds to the context those
739 implementations are designed for. However, implementing such a framework requires
740 setting up extensive datasets, and conclusions can only be drawn for an ensemble of sites
741 and can not be individually tailored for each site.

742 **6. Conclusion**

743 This paper proposes a general framework devoted to the data-based comparison of FA
744 implementations. This framework is based on the following general principles:

- 745 • The performance of FA implementations is judged in terms of reliability and stability. The
746 latter is evaluated in predictive mode, i.e. using data that are not used for calibration.
- 747 • The framework does not use any surrogate for the unknown true quantiles, but uses
748 indices reflecting whether validation data are consistent with FA predictions.
- 749 • The necessity to scrutinize uncertainty estimates is recognized, and a practical solution
750 based on the use of the predictive distribution is proposed.

751 The comparison framework is applied to a case study that uses 364 daily runoff series. The
752 performances of ten FA implementations, belonging to three FA families, are compared. This
753 case study demonstrates the ability of the comparison framework to benchmark FA
754 implementations. Local-regional implementations were found to outperform both the purely
755 local and regional implementations they are built upon, both in terms of reliability and
756 stability. Marked differences were also found regarding the reliability of the predictive
757 distribution, which confirms its relevance to indirectly compare uncertainty estimates.

758 Finally, although the comparison framework proposed in this paper proved its usefulness, it
759 remains open to scrutiny and improvement. In particular, other stability and reliability indices
760 could be defined, and comparison schemes could be tailored to specific regions or hydrologic
761 variables. However, the general principles upon which the framework is built intend to be as
762 general as possible. In particular, the importance of predictive reliability and the need to
763 scrutinize uncertainty estimates are two points that hold to any FA implementation. Moreover,
764 combining this data-based framework with alternative comparison schemes (e.g. based on
765 Monte-Carlo simulations and statistical tests) is likely to yield complementary insights.

766 **7. Acknowledgments**

767 This work is funded by the French Research Agency (ANR) through the project EXTRAFL0
768 (<https://extraflo.cemagref.fr/>). The HYDRO database (Ministry of environment) and EDF are
769 gratefully acknowledged for providing the data. The helpful comments by Dan Rosbjerg, Jery
770 Stedinger, three anonymous reviewers and the Associate Editor are gratefully acknowledged.

771 **8. References**

- 772 Arnaud, P., and J. Lavabre (1999), Using a stochastic model for generating hourly
773 hyetographs to study extreme rainfalls, *Hydrological Sciences Journal*, 44(3), 433-446.
- 774 Atger, F. (1999), The skill of ensemble prediction systems, *Monthly Weather Review*, 127(9),
775 1941-1953.
- 776 Barnard, G. A. (1995), Pivotal Models and the Fiducial Argument, *Int. Stat. Rev.*, 63(3), 309-
777 323.
- 778 Basu, A., and I. R. Harris (1994), Robust Predictive-Distributions for Exponential-Families,
779 *Biometrika*, 81(4), 790-794.
- 780 Beard, L. R. (1974), *Flood Flow Frequency Techniques: A Report*, Center for Research in
781 Water Resources.
- 782 Beard, L. R. (1987), Relative Accuracy of Log Pearson-Iii Procedures - Discussion, *Journal*
783 *of Hydraulic Engineering-Asce*, 113(9), 1205-1206.
- 784 Benichou, P., and O. Le Breton (1987), Prise en compte de la topographie pour la
785 cartographie des champs pluviométriques statistiques, *La Météorologie*, 7(19), 23-34.
- 786 Bjornstad, J. F. (1990), Predictive Likelihood: A Review, *Stat. Sci.*, 5(1), 242-265.
- 787 Bobee, B., G. Cavadias, F. Ashkar, J. Bernier, and P. Rasmussen (1993), Towards a
788 Systematic-Approach to Comparing Distributions Used in Flood Frequency-Analysis, *J.*
789 *Hydrol.*, 142(1-4), 121-136.
- 790 Boughton, W., and O. Droop (2003), Continuous simulation for design flood estimation - a
791 review, *Environmental Modelling & Software*, 18(4), 309-318.
- 792 Breiman, L. (1996), Bagging predictors, *Machine Learning*, 24(2), 123-140.
- 793 Butler, R. W. (1986), Predictive Likelihood Inference with Applications, *J. R. Stat. Soc. Ser.*
794 *B-Methodol.*, 48(1), 1-38.
- 795 Chowdhury, J., and J. Stedinger (1991), Confidence Interval for Design Floods with
796 Estimated Skew Coefficient, *Journal of Hydraulic Engineering*, 117(7), 811-831.
- 797 Chowdhury, J. U., J. R. Stedinger, and L. H. Lu (1991), Goodness-of-Fit Tests for Regional
798 Generalized Extreme Value Flood Distributions, *Water Resources Research*, 27(7), 1765-
799 1776.
- 800 Cipriani, T., T. Toilliez, and E. Sauquet (2012), Estimating 10 year return period peak flows
801 and flood durations at ungauged locations in France, *La houille blanche; submitted*.
- 802 Cohn, T. A., and H. F. Lins (2005), Nature's style: Naturally trendy, *Geophys. Res. Lett.*,
803 32(23).
- 804 Cohn, T. A., W. L. Lane, and J. R. Stedinger (2001), Confidence intervals for expected
805 moments algorithm flood quantile estimates, *Water Resour. Res.*, 37(6), 1695-1706.
- 806 Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, 210 pp.,
807 Springer-Verlag, London.
- 808 Cox, D. R., V. S. Isham, and P. J. Northrop (2002), Floods: some probabilistic and statistical
809 approaches, *Philosophical Transactions of the Royal Society a-Mathematical Physical and*
810 *Engineering Sciences*, 360(1796), 1389-1408.

- 811 Dalrymple, T. (1960), Flood frequency analyses, in *Water-supply paper 1543-A*, edited, US
812 Geological Survey.
- 813 Daly, C. (2006), Guidelines for assessing the suitability of spatial climate data sets, *Int. J.*
814 *Climatol.*, 26(6), 707-721.
- 815 Dawid, A. P. (1984), Statistical Theory - the Prequential Approach, *J. R. Stat. Soc. Ser. A-*
816 *Stat. Soc.*, 147, 278-292.
- 817 Dawid, A. P., and M. Stone (1982), The Functional-Model Basis of Fiducial-Inference,
818 *Annals of Statistics*, 10(4), 1054-1067.
- 819 Dawid, A. P., and J. L. Wang (1993), Fiducial Prediction and Semi-Bayesian Inference,
820 *Annals of Statistics*, 21(3), 1119-1138.
- 821 Durrans, S. R., and S. Tomic (2001), Comparison of parametric tail estimators for low-flow
822 frequency analysis, *J. Am. Water Resour. Assoc.*, 37(5), 1203-1214.
- 823 Durrans, S. R., and J. T. Kirby (2004), Regionalization of extreme precipitation estimates for
824 the Alabama rainfall atlas, *J. Hydrol.*, 295(1-4), 101-107.
- 825 England, J. F., R. D. Jarrett, and J. D. Salas (2003a), Data-based comparisons of moments
826 estimators using historical and paleoflood data, *J. Hydrol.*, 278(1-4), 172-196.
- 827 England, J. F., J. D. Salas, and R. D. Jarrett (2003b), Comparisons of two moments-based
828 estimators that utilize historical and paleoflood data for the log Pearson type III distribution,
829 *Water Resources Research*, 39(9).
- 830 Fisher, R. A. (1930), Inverse Probability, *Mathematical Proceedings of the Cambridge*
831 *Philosophical Society*, 26, 528-535.
- 832 Fushiki, T. (2010), Bayesian bootstrap prediction, *J. Stat. Plan. Infer.*, 140(1), 65-74.
- 833 Fushiki, T., F. Komaki, and K. Aihara (2005), Nonparametric bootstrap prediction, *Bernoulli*,
834 11(2), 293-307.
- 835 Garavaglia, F., J. Gailhard, E. Paquet, M. Lang, R. Garcon, and P. Bernardara (2010),
836 Introducing a rainfall compound distribution model based on weather patterns sub-sampling,
837 *Hydrol. Earth Syst. Sci.*, 14(6), 951-964.
- 838 Garavaglia, F., M. Lang, E. Paquet, J. Gailhard, R. Garcon, and B. Renard (2011), Reliability
839 and robustness of a rainfall compound distribution model based on weather pattern sub-
840 sampling, *Hydrology and Earth System Sciences.*, 15(2), 519-532.
- 841 Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian data analysis*, 526
842 pp., Chapman & Hall.
- 843 Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and
844 sharpness, *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 69, 243-268.
- 845 GREHYS (1996), Inter-comparaison of regional flood frequency procedures for Canadian
846 rivers., *J. Hydrol.*, 186, 85-103.
- 847 Grover, P. L., D. H. Burn, and J. M. Cunderlik (2002), A comparison of index flood
848 estimation procedures for ungauged catchments, *Can. J. Civ. Eng.*, 29(5), 734-741.
- 849 Gunasekara, T. A. G., and C. Cunnane (1992), Split Sampling Technique for Selecting a
850 Flood Frequency-Analysis Procedure, *J. Hydrol.*, 130(1-4), 189-200.
- 851 Hall, J., E. O'Connell, and J. Ewen (2007), On not undermining the science: coherence,
852 validation and expertise. Discussion of Invited Commentary by Keith Beven Hydrological
853 Processes, 20, 3141-3146 (2006), *Hydrol. Process.*, 21(7), 985-988.
- 854 Hall, M. J., H. F. P. van den Boogaard, R. C. Fernando, and A. E. Mynett (2004), The
855 construction of confidence intervals for frequency analysis using resampling techniques,
856 *Hydrol. Earth Syst. Sci.*, 8(2), 235-246.
- 857 Hamed, K. H., and A. R. Rao (1998), A modified Mann-Kendall trend test for autocorrelated
858 data, *J. Hydrol.*, 204(1-4), 182-196.
- 859 Hannig, J., H. Iyer, and P. Patterson (2006), Fiducial generalized confidence intervals, *J. Am.*
860 *Stat. Assoc.*, 101(473), 254-269.

- 861 Harris, I. R. (1989), Predictive Fit for Natural Exponential-Families, *Biometrika*, 76(4), 675-
862 684.
- 863 He, J. X., and C. Valeo (2009), Comparative Study of ANNs versus Parametric Methods in
864 Rainfall Frequency Analysis, *J. Hydrol. Eng.*, 14(2), 172-184.
- 865 Hine, D., and J. W. Hall (2010), Information gap analysis of flood model uncertainties and
866 regional frequency analysis, *Water Resources Research*, 46.
- 867 Hinkley, D. (1979), Predictive Likelihood, *Annals of Statistics*, 7(4), 718-728.
- 868 Hosking, J. R. M., R. Wallis James, and F. Wood Eric (1985), An appraisal of the regional
869 flood frequency procedure in the UK flood studies report, *Hydrological Sciences Journal*,
870 30(1), 85-109.
- 871 Institution of Engineers Australia (1987), *Australian Rainfall and Runoff*, Engineers
872 Australia.
- 873 Interagency Advisory Committee on Water Data (1982), *Guidelines for determining flood-
874 flow frequency: Bulletin 17B of the Hydrology Subcommittee*, U.S. Geological Survey,
875 Reston, Va.
- 876 Koutsoyiannis, D. (2010), HESS Opinions 'A random walk on water', *Hydrol. Earth Syst. Sci.*,
877 14(3), 585-601.
- 878 Kroll, C. N., and J. R. Stedinger (1996), Estimation of moments and quantiles using censored
879 data, *Water Resources Research*, 32(4), 1005-1012.
- 880 Kuczera, G. (1999), Comprehensive at-site flood frequency analysis using Monte Carlo
881 Bayesian inference, *Water Resources Research*, 35(5), 1551-1557.
- 882 Kysely, J. (2008), A Cautionary Note on the Use of Nonparametric Bootstrap for Estimating
883 Uncertainties in Extreme-Value Models, *Journal of Applied Meteorology and Climatology*,
884 47(12), 3236-3251.
- 885 Kysely, J., L. Gaál, and J. Pícek (2011), Comparison of regional and at-site approaches to
886 modelling probabilities of heavy precipitation, *Int. J. Climatol.*, 31(10), 1457-1472.
- 887 Laio, F. (2004), Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme
888 value distributions with unknown parameters, *Water Resources Research*, 40(9).
- 889 Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous
890 hydrological variables, *Hydrol. Earth Syst. Sci.*, 11(4), 1267-1277.
- 891 Lang, M., K. Pobanz, B. Renard, E. Renouf, and E. Sauquet (2010), Extrapolation of rating
892 curves by hydraulic modelling, with application to flood frequency analysis, *Hydrological
893 sciences Journal.*, 55(6), 883-898.
- 894 Lawless, J. F., and M. Fredette (2005), Frequentist prediction intervals and predictive
895 distributions, *Biometrika*, 92(3), 529-542.
- 896 Lee, K. S., and S. U. Kim (2008), Identification of uncertainty in low flow frequency analysis
897 using Bayesian MCMC method, *Hydrol. Process.*, 22(12), 1949-1964.
- 898 Lee, Y., and J. A. Nelder (1996), Hierarchical generalized linear models, *J. R. Stat. Soc. Ser.
899 B-Methodol.*, 58(4), 619-656.
- 900 Lima, C. H. R., and U. Lall (2010), Spatial scaling in a changing climate: A hierarchical
901 bayesian model for non-stationary multi-site annual maximum and monthly streamflow, *J.
902 Hydrol.*, 383(3-4), 307-318.
- 903 Madsen, H., and D. Rosbjerg (1997), Generalized least squares and empirical Bayes
904 estimation in regional partial duration series index-flood modeling, *Water Resources
905 Research*, 33(4), 771-781.
- 906 Madsen, H., C. P. Pearson, and D. Rosbjerg (1997a), Comparison of annual maximum series
907 and partial duration series methods for modeling extreme hydrologic events .2. Regional
908 modeling, *Water Resources Research*, 33(4), 759-769.

- 909 Madsen, H., P. F. Rasmussen, and D. Rosbjerg (1997b), Comparison of annual maximum
910 series and partial duration series methods for modeling extreme hydrologic events .1. At-site
911 modeling, *Water Resources Research*, 33(4), 747-757.
- 912 Mardhel, V., P. Frantar, J. Uhan, and A. Mio (2004), Index of development and persistence of
913 the river networks as a component of regional groundwater vulnerability assessment in
914 Slovenia., paper presented at Int. Conf. groundwater vulnerability assessment and mapping,
915 Ustron, Poland, 15-18 June 2004.
- 916 Markiewicz, I., and W. G. Strupczewski (2009), Dispersion measures for flood frequency
917 analysis, *Physics and Chemistry of the Earth*, 34(10-12), 670-678.
- 918 Martins, E. S., and J. R. Stedinger (2000), Generalized maximum-likelihood generalized
919 extreme-value quantile estimators for hydrologic data, *Water Resources Research*, 36(3), 737-
920 744.
- 921 Meng, X. L. (2009), Decoding the H-likelihood, *Stat. Sci.*, 24(3), 280-293.
- 922 Meshgi, A., and D. Khalili (2009), Comprehensive evaluation of regional flood frequency
923 analysis by L- and LH-moments. II. Development of LH-moments parameters for the
924 generalized Pareto and generalized logistic distributions, *Stoch. Environ. Res. Risk Assess.*,
925 23(1), 137-152.
- 926 Meylan, P., A.-C. Favre, and A. Musy (2008), *Hydrologie fréquentielle: Une science*
927 *prédictive*, 173 pp., Presses polytechniques et universitaires romandes, Lausanne.
- 928 Micevski, T., S. W. Franks, and G. Kuczera (2006a), Multidecadal variability in coastal
929 eastern Australian flood data, *J. Hydrol.*, 327(1-2), 219-225.
- 930 Micevski, T., G. Kuczera, and S. W. Franks (2006b), A Bayesian Hierarchical Regional Flood
931 Model, paper presented at 30th Hydrology and Water Resources Symposium, Engineers
932 Australia, Launceston, Tas, Australia, 4-7 Dec.
- 933 Naulet, R., M. Lang, T. B. M. J. Ouarda, D. Coeur, B. Bobee, A. Recking, and D. Moussay
934 (2005), Flood frequency analysis on the Ardeche river using French documentary sources
935 from the last two centuries, *J. Hydrol.*, 313(1-2), 58-78.
- 936 Neppel, L., P. Arnaud, and J. Lavabre (2007), Extreme rainfall mapping: Comparison
937 between two approaches in the Mediterranean area, *C. R. Geosci.*, 339(13), 820-830.
- 938 Neppel, L., B. Renard, M. Lang, P. A. Ayrat, D. Coeur, E. Gaume, N. Jacob, O. Payrastre, K.
939 Pobanz, and F. Vinet (2010), Flood frequency analysis using historical data: accounting for
940 random and systematic errors, *Hydrol. Sci. J.-J. Sci. Hydrol.*, 55(2), 192-208.
- 941 O'Connell, D. R. H., D. A. Ostenaar, D. R. Levish, and R. E. Klinger (2002), Bayesian flood
942 frequency analysis with paleohydrologic bound data, *Water Resources Research*, 38(5).
- 943 Ouarda, T., J. M. Cunderlik, A. St-Hilaire, M. Barbet, P. Bruneau, and B. Bobee (2006),
944 Data-based comparison of seasonality-based regional flood frequency methods, *J. Hydrol.*,
945 330(1-2), 329-339.
- 946 Overeem, A., A. Buishand, and I. Holleman (2008), Rainfall depth-duration-frequency curves
947 and their uncertainties, *J. Hydrol.*, 348(1-2), 124-134.
- 948 Parent, E., and J. Bernier (2003), Bayesian POT modeling for historical data, *J. Hydrol.*, 274,
949 95-108.
- 950 Payrastre, O., E. Gaume, and H. Andrieu (2011), Usefulness of historical information for
951 flood frequency analyses: Developments based on a case study, *Water Resources Research*,
952 47.
- 953 Pujol, N., L. Neppel, and R. Sabatier (2007), Regional tests for trend detection in maximum
954 precipitation series in the French Mediterranean region, *Hydrol. Sci. J.-J. Sci. Hydrol.*, 52(5),
955 956-973.
- 956 Reed, D. W., D. S. Faulkner, A. J. Robson, H. Houghton-Carr, and A. C. Bayliss (1999),
957 *Flood Estimation Handbook*, Institute of Hydrology, Wallingford.

- 958 Reis, D. S., and J. R. Stedinger (2005), Bayesian MCMC flood frequency analysis with
959 historical information, *J. Hydrol.*, 313(1-2), 97-116.
- 960 Reis, D. S., J. R. Stedinger, and E. S. Martins (2005), Bayesian generalized least squares
961 regression with application to log Pearson type 3 regional skew estimation, *Water Resources*
962 *Research*, 41(10).
- 963 Renard, B. (2011), A Bayesian Hierarchical Approach To Regional Frequency Analysis,
964 *Water Resources Research*, 47.
- 965 Renard, B., V. Garreta, and M. Lang (2006a), An application of Bayesian analysis and
966 MCMC methods to the estimation of a regional trend in annual maxima, *Water Resources*
967 *Research*, 42(12).
- 968 Renard, B., M. Lang, and P. Bois (2006b), Statistical analysis of extreme events in a non-
969 stationary context via a Bayesian framework., *Stoch. Environ. Res. Risk Assess.*, 21, 97-112.
- 970 Renard, B., D. Kavetski, M. Thyer, G. Kuczera, and S. W. Franks (2010), Understanding
971 predictive uncertainty in hydrologic modeling: The challenge of identifying input and
972 structural errors, *Water Resources Research*, 46.
- 973 Renard, B., et al. (2008), Regional methods for trend detection: Assessing field significance
974 and regional consistency, *Water Resources Research*, 44(8).
- 975 Ribatet, M., E. Sauquet, J. M. Gresillon, and T. B. M. J. Ouarda (2006), A regional Bayesian
976 POT model for flood frequency analysis, *Stoch. Environ. Res. Risk Assess.*, 21(4), 327-339.
- 977 Ribatet, M., E. Sauquet, J. M. Gresillon, and T. B. M. J. Ouarda (2007), Usefulness of the
978 reversible jump Markov chain Monte Carlo model in regional flood frequency analysis, *Water*
979 *Resources Research*, 43(8).
- 980 Robson, A. J., and D. W. Reed (1999), *Flood Estimation Handbook. Volume 3: Statistical*
981 *procedures for flood frequency estimation*, 338 pp., Wallingford.
- 982 Rosbjerg, D., and H. Madsen (1998), Design with uncertain design values, in *Hydrology in a*
983 *Changing Environment, Vol III*, edited by H. Wheater and C. Kirby, pp. 155-163, John Wiley
984 & Sons.
- 985 Sankarasubramanian, A., and K. Srinivasan (1999), Investigation and comparison of sampling
986 properties of L-moments and conventional moments, *J. Hydrol.*, 218(1-2), 13-34.
- 987 Seidenfeld, T. (1992), R.A. Fisher's Fiducial Argument and Bayes' Theorem, *Stat. Sci.*, 7(3),
988 358-368.
- 989 Spreafico, M., R. Weingartner, M. Barben, A. Ryser, B. Hingray, A. Musy, and M. Niggli
990 (2003), Evaluation des crues dans les bassins versants de SuisseRep., Département fédéral de
991 l'environnement, des transports, de l'énergie et de la communication, Berne.
- 992 Stedinger, J., and L. Lu (1995), Appraisal of regional and index flood quantile estimators,
993 *Stoch. Hydrol. Hydraul.*, 9(1), 49-75.
- 994 Stedinger, J. R. (1983a), Design-Events with Specified Flood Risk, *Water Resources*
995 *Research*, 19(2), 511-522.
- 996 Stedinger, J. R. (1983b), Confidence-Intervals for Design-Events, *Journal of Hydraulic*
997 *Engineering-Asce*, 109(1), 13-27.
- 998 Stedinger, J. R., and G. D. Tasker (1985), Regional hydrologic analysis: 1. Ordinary,
999 weighted and generalized least squares compared, *Water Resources Research*, 21(9), 1421-
1000 1432 [Correction, *Water Resour. Res.*, 1422(1425), 1844, 1986.].
- 1001 Stedinger, J. R., and T. A. Cohn (1986), Flood Frequency-Analysis with Historical and
1002 Paleoflood Information, *Water Resources Research*, 22(5), 785-793.
- 1003 Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the
1004 generalized likelihood uncertainty estimation (GLUE) method, *Water Resources Research*,
1005 44.
- 1006 Stephens, M. A. (1974), EDF Statistics for Goodness of Fit and Some Comparisons, *J. Am.*
1007 *Stat. Assoc.*, 69(347), 730-737.

- 1008 Szolgay, J., J. Parajka, S. Kohnova, and K. Hlavcova (2009), Comparison of mapping
1009 approaches of design annual maximum daily precipitation, *Atmos. Res.*, 92(3), 289-307.
1010 Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009),
1011 Critical evaluation of parameter consistency and predictive uncertainty in hydrological
1012 modelling: a case study using bayesian total error analysis, *Water Resources Research*, 45.
1013 Todini, E., and P. Mantovan (2007), Comment on: 'On undermining the science?' by Keith
1014 Beven, *Hydrol. Process.*, 21, 1633-1638.
1015 Vidoni, P. (1995), A simple predictive density based on the p*-formula, *Biometrika*, 82(4),
1016 855-863.
1017 Wallis, J. R., and E. F. Wood (1985), Relative Accuracy of Log Pearson-Iii Procedures,
1018 *Journal of Hydraulic Engineering-Asce*, 111(7), 1043-1056.
1019 Wallis, J. R., and E. F. Wood (1987), Relative Accuracy of Log Pearson-Iii Procedures -
1020 Closure, *Journal of Hydraulic Engineering-Asce*, 113(9), 1210-1214.
1021 Wang, Y. H. (2000), Fiducial intervals: What are they?, *American Statistician*, 54(2), 105-
1022 111.
1023 Wasson, J. G., A. Chandesris, H. Pella, and L. Blanc (2004), Les hydro-écorégions: une
1024 approche fonctionnelle de la typologie des rivières pour la directive cadre européenne sur
1025 l'eau, *Ingénieries*, 40, 3-10.
1026 Yu, P. S., T. C. Yang, and C. S. Lin (2004), Regional rainfall intensity formulas based on
1027 scaling property of rainfall, *J. Hydrol.*, 295(1-4), 108-123.
1028
1029

1030 **9. Appendix 1: distribution of performance indices**

1031 **9.1. Pval**

1032 Let $t \in [0;1]$. $\Pr(Pval_k^{(i)} \leq t) = \Pr(\hat{F}_M^{(i)}(D_k^{(i)}) \leq t)$

1033 If the estimation is reliable ($\hat{F}_M^{(i)} = F^{(i)}$):

$$\begin{aligned} \Pr(Pval_k^{(i)} \leq t) &= \Pr(F^{(i)}(D_k^{(i)}) \leq t) \\ 1034 \quad &= \Pr(D_k^{(i)} \leq \{F^{(i)}\}^{-1}(t)) \\ &= F^{(i)}(\{F^{(i)}\}^{-1}(t)) = t \end{aligned}$$

1035 This corresponds to the cdf of a uniform distribution on [0,1].

1036 **9.2. N_T**

1037 For a given time step k , the exceedance of a quantile $\hat{q}_T^{(i)}$ is a Bernoulli trial. If the estimation
1038 is reliable ($\hat{q}_T^{(i)} = q_T^{(i)}$), its success (meaning here the exceedance of the T -quantile) probability
1039 is $1/T$. With the assumption of serial independence, the variable N_T therefore corresponds to
1040 the number of successes in $n^{(i)}$ independent Bernoulli experiments: its distribution is therefore
1041 Binomial, with parameters $(n^{(i)}, 1/T)$.

1042 **9.3. FF**

1043 Let $t \in [0;1]$. $\Pr(FF^{(i)} \leq t) = \Pr(\hat{F}_M^{(i)}(D_{\max}^{(i)}) \leq t)$

1044 If the estimation is reliable ($\hat{F}_M^{(i)} = F^{(i)}$):

$$\begin{aligned} \Pr(FF^{(i)} \leq t) &= \Pr(F^{(i)}(D_{\max}^{(i)}) \leq t) \\ &= \Pr(D_{\max}^{(i)} \leq \{F^{(i)}\}^{-1}(t)) \\ 1045 \quad &= \Pr(D_k^{(i)} \leq \{F^{(i)}\}^{-1}(t) \forall k = 1 \dots n^{(i)}) \\ &= \left[F(\{F^{(i)}\}^{-1}(t)) \right]^{n^{(i)}} = t^{n^{(i)}} \end{aligned}$$

1046 This corresponds to the cdf of the Kumaraswamy distribution with parameters $(n^{(i)}, 1)$. Note
 1047 that the transition between lines 3 and 4 uses the serial independence hypothesis.

1048 **9.4. Randomized probability transformation for N_T**

1049 Let $W_T^{(i)}$ be a random variable whose distribution, conditional on $N_T^{(i)}$, is uniform between
 1050 $b(N_T^{(i)} - 1)$ and $b(N_T^{(i)})$ (see section 3.2.5). Recall that $b(j)$ is defined by $b(j) = \Pr(N \leq j)$,
 1051 with $N \sim \text{Bin}(n^{(i)}, 1/T)$. Let $t \in [0;1]$. The conditional cdf of $W_T^{(i)}$ is:

$$1052 \quad \Pr(W_T^{(i)} \leq t | N_T^{(i)} = j) = \begin{cases} 0 & \text{if } t \leq b(j-1) \\ [t - b(j-1)] / [b(j) - b(j-1)] & \text{if } b(j-1) \leq t \leq b(j) \\ 1 & \text{if } t \geq b(j) \end{cases}$$

1053 The unconditional cdf of $W_T^{(i)}$ can then be derived by using the total probability law:

$$1054 \quad \Pr(W_T^{(i)} \leq t) = \sum_{j=0}^{+\infty} \Pr(W_T^{(i)} \leq t | N_T^{(i)} = j) \Pr(N_T^{(i)} = j)$$

1055 Let k denote the integer verifying $b(k) \leq t < b(k+1)$. The infinite sum above can then be
 1056 decomposed as follows:

$$\begin{aligned} \Pr(W_T^{(i)} \leq t) &= \sum_{j=0}^k \Pr(W_T^{(i)} \leq t | N_T^{(i)} = j) \Pr(N_T^{(i)} = j) \\ 1057 \quad &+ \Pr(W_T^{(i)} \leq t | N_T^{(i)} = k+1) \Pr(N_T^{(i)} = k+1) \\ &+ \sum_{j=k+2}^{+\infty} \Pr(W_T^{(i)} \leq t | N_T^{(i)} = j) \Pr(N_T^{(i)} = j) \end{aligned}$$

1058 When $j \leq k$, $b(j) \leq b(k) \leq t$, and $\Pr(W_T^{(i)} \leq t | N_T^{(i)} = j) = 1$

1059 When $j \geq k+2$, $t < b(k+1) \leq b(j-1)$, and $\Pr(W_T^{(i)} \leq t | N_T^{(i)} = j) = 0$

$$\begin{aligned}
 \text{Consequently, } \Pr(W_T^{(i)} \leq t) &= \sum_{j=0}^k 1 \times \Pr(N_T^{(i)} = j) + \frac{t - b(k)}{b(k+1) - b(k)} \Pr(N_T^{(i)} = k+1) + \sum_{j=k+2}^{+\infty} 0 \times \Pr(N_T^{(i)} = j) \\
 1060 \quad &= \Pr(N_T^{(i)} \leq k) + \frac{t - \Pr(N \leq k)}{\Pr(N \leq k+1) - \Pr(N \leq k)} \Pr(N_T^{(i)} = k+1) \\
 &= \Pr(N_T^{(i)} \leq k) + \frac{t - \Pr(N \leq k)}{\Pr(N = k+1)} \Pr(N_T^{(i)} = k+1)
 \end{aligned}$$

1061 Under the reliability hypothesis, $N_T^{(i)} \sim \text{Bin}(n^{(i)}, 1/T)$, which is the same distribution as that
 1062 of N . Consequently, $\Pr(N \leq k) = \Pr(N_T^{(i)} \leq k)$ and $\Pr(N = k+1) = \Pr(N_T^{(i)} = k+1)$. The
 1063 equation above therefore simplifies as follows:

$$1064 \quad \Pr(W_T^{(i)} \leq t) = \Pr(N_T^{(i)} \leq k) + \frac{t - \Pr(N_T^{(i)} \leq k)}{\Pr(N_T^{(i)} = k+1)} \Pr(N_T^{(i)} = k+1) = t$$

1065 This corresponds to the cdf of a uniform distribution between 0 and 1.

1066 **10. Appendix 2: algorithms for predictive distributions**

1067 **10.1. Bayesian predictive distributions**

1068 It is assumed that the Bayesian inference is performed using a Markov chain Monte Carlo
 1069 (MCMC) sampler, yielding a sample $(\boldsymbol{\theta}^{(i)})_{i=1:N_{sim}}$ from the posterior distribution $p_M(\boldsymbol{\theta} | \mathbf{c})$.

1070 The pdf of the predictive distribution $\hat{\pi}_M(y)$ evaluated at y can then be approximated by:

$$\hat{\pi}_M(y) \approx \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} f_M(y | \boldsymbol{\theta}^{(i)}) \quad (9)$$

1071 Note that it may be more practical to generate a large sample $(y^{(i)})_{i=1:N_{sim}}$ from the predictive
 1072 distribution and use its empirical distribution as an approximation:

1073 Do $i = 1:N_{sim}$

1074 1. Sample $y^{(i)}$ from the distribution with pdf $f_M(y | \boldsymbol{\theta}^{(i)})$

1075 **10.2. Non-Bayesian predictive distributions**

1076 Let $\hat{s}_M(\boldsymbol{\tau})$ denote the pdf of the sampling distribution of the estimator $\hat{\boldsymbol{\theta}}(X)$. The non-
 1077 Bayesian predictive distribution can be approximated using the same algorithms than in
 1078 section 10.1, replacing the sample $(\boldsymbol{\theta}^{(i)})_{i=1:N_{sim}}$ from the posterior distribution by a sample
 1079 $(\boldsymbol{\tau}^{(i)})_{i=1:N_{sim}}$ generated from the sampling distribution $\hat{s}_M(\boldsymbol{\tau})$.

1080 In practice, the algorithm used to generate the sample $(\boldsymbol{\tau}^{(i)})_{i=1:N_{sim}}$ depends on the way $\hat{s}_M(\boldsymbol{\tau})$
1081 is derived. For instance, if bootstrap resampling of observations is used, a sample $(\boldsymbol{\tau}^{(i)})_{i=1:N_{sim}}$
1082 is then available from the bootstrap replications of data. Alternatively, $\hat{s}_M(\boldsymbol{\tau})$ may be derived
1083 using a large-sample Gaussian approximation (as done in many estimation approaches) and
1084 whose generation poses no difficulty. In non-Gaussian approximation of $\hat{s}_M(\boldsymbol{\tau})$ and other
1085 complicated cases, specialized sampling algorithms (e.g. MCMC) may be required.

1086 Finally, some FA implementations provide uncertainties expressed directly on quantiles rather
1087 than on parameters. In such a case, let $\hat{s}_{M,T}(q)$ denote the pdf of the sampling distribution of
1088 the estimated T -year quantile $\hat{Q}_T(X)$. A sample $(y^{(i)})_{i=1:N_{sim}}$ from the predictive distribution
1089 can be generated as follows:

1090 Do $i = 1:N_{sim}$

- 1091 1. Sample u from a uniform distribution on $[0;1]$.
- 1092 2. Compute $T = 1/(1-u)$
- 1093 3. Sample $y^{(i)}$ from the sampling distribution of $\hat{Q}_T(X)$ with pdf $\hat{s}_{M,T}(q)$.

1094 **11. Appendix 3: Regional and local-regional FA** 1095 **implementations**

1096 **11.1. Regional implementations based on an index flood model**

1097 **Index flood regression:** the index flood values at site i , v_i , are linked with catchment
1098 descriptors $w_i^{(1)}, \dots, w_i^{(N_{cov})}$ using the following regression:

$$\log(v_i) = \beta_0 + \sum_{j=1}^{N_{cov}} \beta_j w_i^{(j)} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \quad (10)$$

1099 Building on previous work by *Cipriani et al.* [2012], the following catchment descriptors are
1100 used: (i) catchment area; (ii) mean catchment elevation; (iii) mean of 10-year daily rainfall
1101 within the catchment, as estimated by *Benichou and Le Breton* [1987]; (iv) mean IDPR index.
1102 The latter index (Index of Development and Persistence of the River networks) was proposed
1103 by *Mardhel et al.* [2004] as an indicator of infiltration capacity. Moreover, region-specific
1104 regressions are estimated, with regions shown in Figure 3b and based on the Hydro-
1105 ecoregions defined by *Wasson et al.*[2004].

1106 Estimation of regression parameters $\beta_0, \dots, \beta_{N_{\text{cov}}}$ and residual standard deviation σ is
 1107 performed using a Bayesian approach (with flat priors on $(\beta_0, \dots, \beta_{N_{\text{cov}}}, \log(\sigma))$).

1108 **Regional distribution estimation:** Depending on the FA implementation, regional Gumbel
 1109 or GEV distributions are estimated in each region, based on all standardized data of the region
 1110 pooled together. A Bayesian approach (with flat priors) is used.

1111 **Prediction at target site:** At a target site k , the estimated distribution is a Gumbel or a GEV
 1112 distribution (depending on the FA implementation) with parameters:

$$\begin{aligned} \text{location: } \mu_k &= \hat{v}_k \times \hat{\mu}_{reg}, \text{ with } \hat{v}_k = \exp \left[\hat{\beta}_0 + \sum_{j=1}^{N_{\text{cov}}} \hat{\beta}_j w_k^{(j)} \right] \\ \text{scale: } \lambda_k &= \hat{v}_k \times \hat{\lambda}_{reg} \\ \text{shape (GEV distribution only): } \xi_k &= \hat{\xi}_{reg} \end{aligned} \tag{11}$$

1113 The predictive distribution is derived by propagating forward the MCMC samples of
 1114 $(\beta_0, \dots, \beta_{N_{\text{cov}}}, \sigma, \mu_{reg}, \lambda_{reg}, \xi_{reg})$, as outlined in section 10.1. The MCMC algorithm used in this
 1115 paper is described by *Renard et al.* [2006a].

1116 **11.2. Local-regional implementations**

1117 Propagating forward the MCMC samples of $(\beta_0, \dots, \beta_{N_{\text{cov}}}, \sigma, \mu_{reg}, \lambda_{reg}, \xi_{reg})$ into equation (11)
 1118 yields a large number of replicates for the Gumbel (or GEV) parameters at the target site.
 1119 These replicates can be used to specify a prior for the local-regional implementation. To this
 1120 aim, a Gaussian distribution is estimated based on the replicates, and is used as the prior
 1121 distribution for the local-regional implementation. The rest of the analysis then proceeds as in
 1122 standard local implementations.

1123

1123 **List of captions**

1124 Table 1. Summary of the FA implementations studied in this paper.

1125 Figure 1. Typical shapes for pp-plots (a-c) and qq-plots in Gumbel space (d-f).

1126 Figure 2. Illustration of the difference between the estimated distribution (with pdf $\hat{f}_M(y)$),
1127 and the predictive distribution (with pdf $\hat{\pi}_M(y)$). This illustrative figure results from the
1128 Bayesian estimation of a GEV distribution using 25 observations. Uncertainty intervals are
1129 quantile posterior intervals.

1130 Figure 3. Location of the study sites. (a) Decomposition into “regional sites” used to estimate
1131 the regional models and “local sites” used for local estimation and validation; (b) Regions
1132 derived from the Hydro-ecoregions of *Wasson et al.* [2004].

1133 Figure 4. Reliability diagnostics applied to the implementation GEV-ML (local estimation of
1134 a GEV distribution with maximum likelihood). (a) pval pp-plot. Each gray line refers to a
1135 validation site. (b) FF pp-plot. Red = validation data, blue = calibration data. (c) FF qq-plot in
1136 Gumbel space. The percentages of “impossible observations” (i.e. observations incompatible
1137 with the estimated GEV, yielding FF=1) are provided. (d) Randomized pp-plot of N10
1138 computed on all available observations; (e) Randomized qq-plot of N10 in Gumbel space.

1139 Figure 5. Reliability diagnostics for the six local FA implementations. First row = estimated
1140 distribution, second row = predictive distribution. (a) and (d): FF qq-plot in Gumbel space; (b)
1141 and (e): N10 qq-plot in Gumbel space; (c) and (f): N100 qq-plot in Gumbel space.

1142 Figure 6. pval pp-plot for local, local-regional and regional estimation of the GEV distribution
1143 (estimated distribution).

1144 Figure 7. Reliability diagnostics for six FA implementations (local, regional and local-
1145 regional, with Gumbel and GEV distributions). First row = estimated distribution, second row
1146 = predictive distribution. (a) and (d): FF qq-plot in Gumbel space; (b) and (e): N10 qq-plot in
1147 Gumbel space; (c) and (f): N100 qq-plot in Gumbel space.

1148 Figure 8. Stability diagnostic for six FA implementations (local, regional and mixed local-
1149 regional, with Gumbel and GEV distributions). Left = type I decomposition, right = type II
1150 decomposition. (a) – (b) = estimated distribution, (c) – (d) = predictive distribution.

1151

1151 **Notation list**

1152 $\mathbf{x} = (x_k^{(i)})_{i=1:N_{site}, k=1:n^{(i)}}$ observations

1153 \mathbf{c} subset of \mathbf{x} used for calibration

1154 \mathbf{v} subset of \mathbf{x} used for validation

1155 \mathbf{d} denotes either one of \mathbf{c} or \mathbf{v}

1156 $F^{(i)}(y)$ cdf of the parent distribution (evaluated at some value y)

1157 $F_M^{(i)}(y|\boldsymbol{\theta})$ cdf of the assumed distribution in implementation M , with unknown parameters $\boldsymbol{\theta}$

1158 $f_M(y|\boldsymbol{\theta})$ pdf of the assumed distribution in implementation M , with unknown parameters $\boldsymbol{\theta}$

1159 $\hat{F}_M^{(i)}(y)$ cdf of the estimated distribution in implementation M

1160 $\hat{f}_M(y)$ pdf of the estimated distribution in implementation M

1161 $\hat{\Pi}_M(y)$ cdf of the predictive distribution in implementation M

1162 $\hat{\pi}_M(y)$ pdf of the predictive distribution in implementation M

1163 $\hat{\boldsymbol{\theta}}(X)$ estimator of unknown parameters $\boldsymbol{\theta}$

1164 $\hat{\boldsymbol{\theta}}$ estimated value of $\boldsymbol{\theta}$

1165 $s_M(\boldsymbol{\tau}|\boldsymbol{\theta})$ pdf of the sampling distribution of $\hat{\boldsymbol{\theta}}(X)$ in implementation M (evaluated at some
1166 value $\boldsymbol{\tau}$)

1167 $\hat{s}_M(\boldsymbol{\tau})$ pdf of the estimated sampling distribution of $\hat{\boldsymbol{\theta}}(X)$ in implementation M

1168 $p_M(\boldsymbol{\theta}|\mathbf{c})$ posterior distribution of $\boldsymbol{\theta}$ given observations \mathbf{c} in implementation M

1169

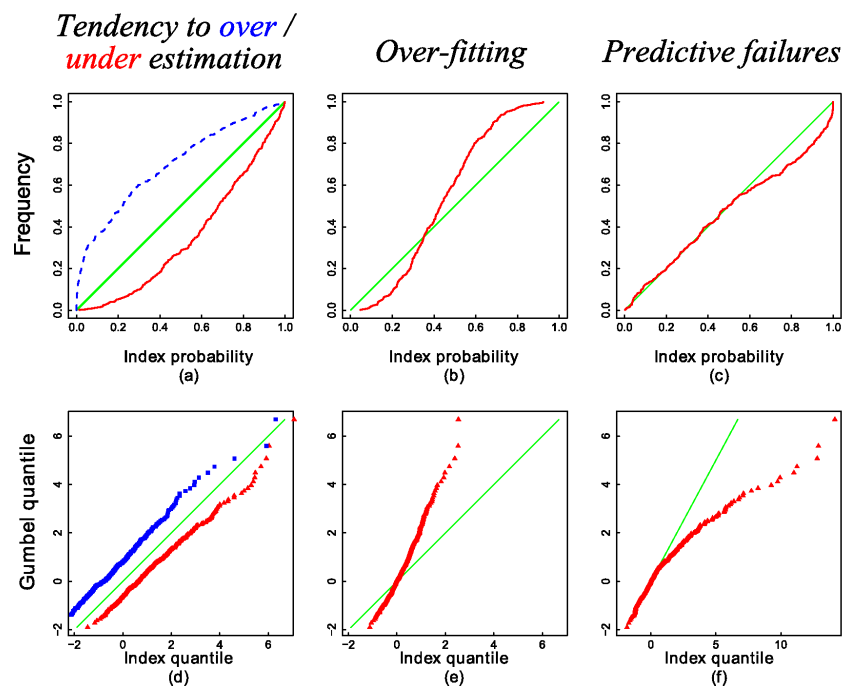
1169
 1170

Table 1. Summary of the FA implementations studied in this paper.

Distribution	Estimation method	Notation	Uncertainty Quantification
<i>FA Family: local estimation</i>			
GEV	Moments	GEV_MOM	Bootstrap
GEV	Maximum Likelihood	GEV_ML	Gaussian approximation ¹
GEV	Bayesian	GEV_BAY	Bayesian
Gumbel	Moments	GUM_MOM	Bootstrap
Gumbel	Maximum Likelihood	GUM_ML	Gaussian approximation ¹
Gumbel	Bayesian	GUM_BAY	Bayesian
<i>FA Family: regional estimation</i>			
GEV	Bayesian	GEV_REG	Bayesian
Gumbel	Bayesian	GUM_REG	Bayesian
<i>FA Family: local-regional estimation</i>			
GEV	Bayesian	GEV_LR	Bayesian
Gumbel	Bayesian	GUM_LR	Bayesian

1171 ¹Asymptotic normality of ML estimator, with covariance matrix equal to the Fisher
 1172 information matrix.

1173

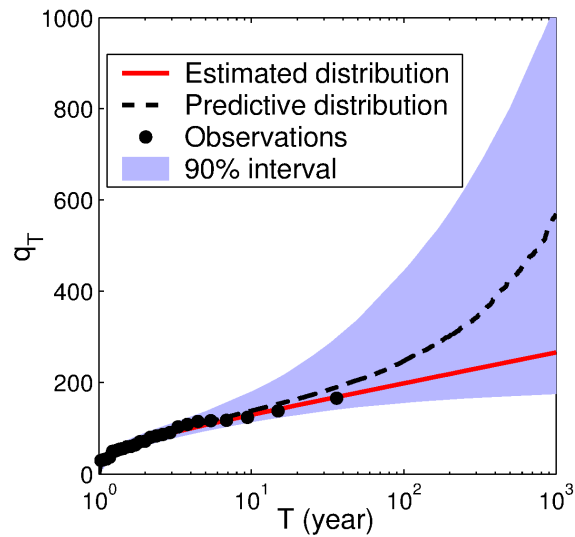


1174

1175

1176

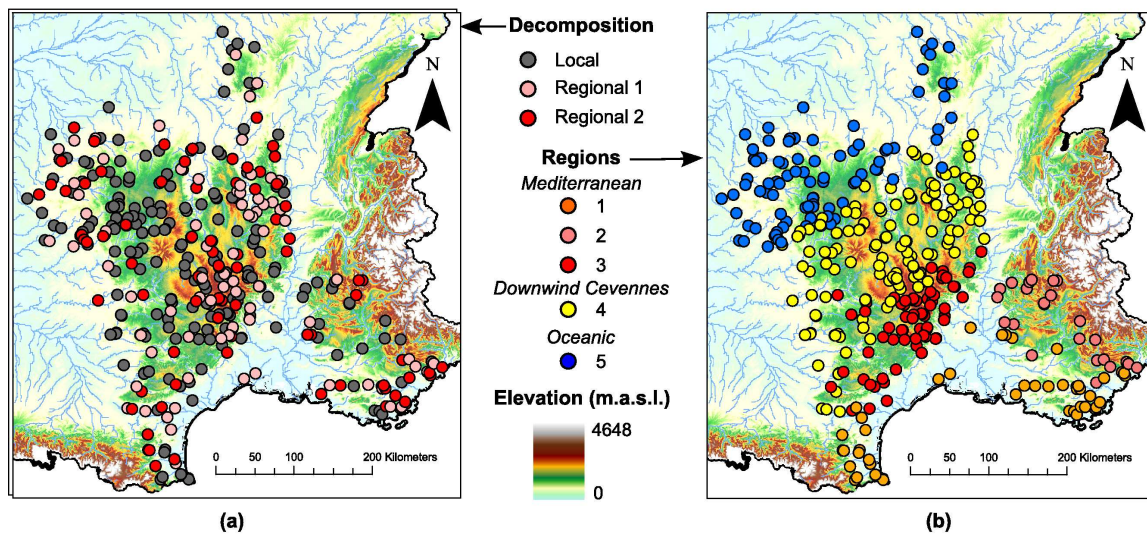
Figure 1. Typical shapes for pp-plots (a-c) and qq-plots in Gumbel space (d-f).



1177

1178 **Figure 2. Illustration of the difference between the estimated distribution (with pdf $\hat{f}_M(y)$) and the**
1179 **predictive distribution (with pdf $\hat{\pi}_M(y)$). This illustrative figure results from the Bayesian estimation of**
1180 **a GEV distribution using 25 observations. Uncertainty intervals are quantile posterior intervals.**

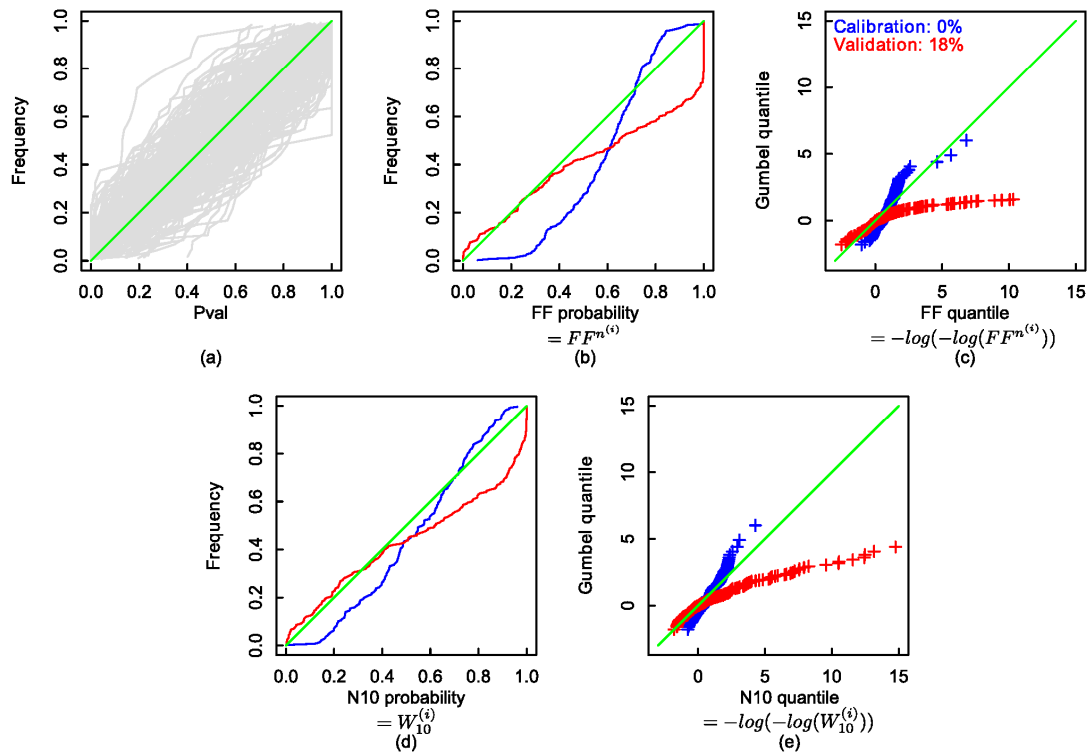
1181



1182

1183 **Figure 3. Location of the study sites. (a) Decomposition into “regional sites” used to estimate the regional**
1184 **models and “local sites” used for local estimation and validation; (b) Regions derived from the Hydro-**
1185 **ecoregions of Wasson *et al.* [2004].**

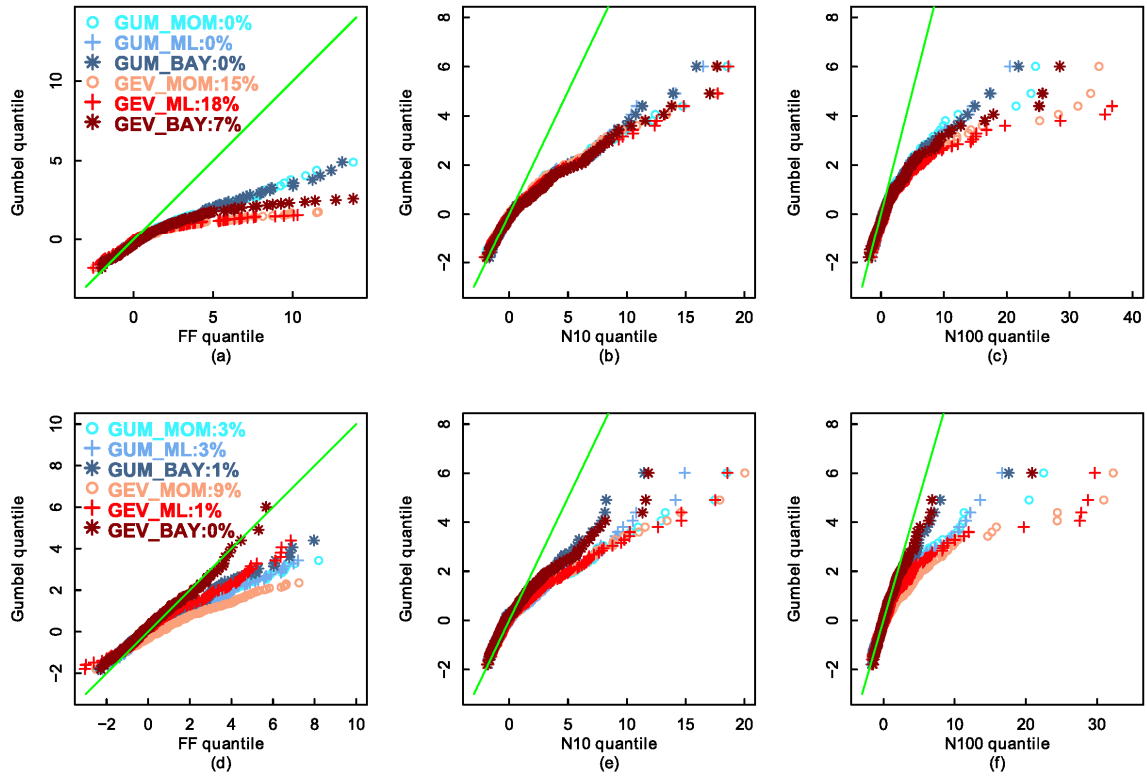
1186



1187

1188 **Figure 4. Reliability diagnostics applied to the implementation GEV-ML (local estimation of a GEV**
 1189 **distribution with maximum likelihood). (a) *pval* pp-plot. Each gray line refers to a validation site. (b) *FF***
 1190 **pp-plot. Red = validation data, blue = calibration data. (c) *FF* qq-plot in Gumbel space. The percentages**
 1191 **of “impossible observations” (i.e. observations incompatible with the estimated GEV, yielding *FF*=1) are**
 1192 **provided. (d) Randomized pp-plot of N_{10} computed on all available observations; (e) Randomized qq-plot**
 1193 **of N_{10} in Gumbel space.**

1194



1195

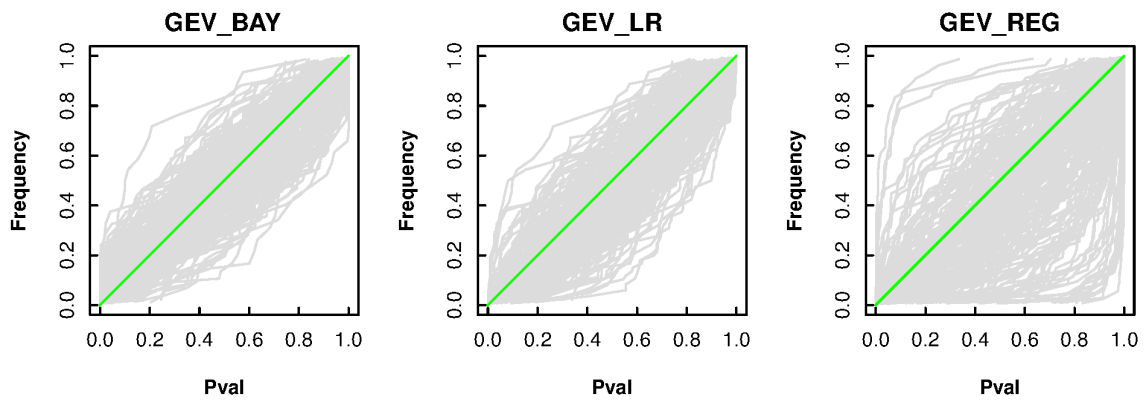
1196

1197

1198

1199

Figure 5. Reliability diagnostics for the six local FA implementations. First row = estimated distribution, second row = predictive distribution. (a) and (d): FF qq-plot in Gumbel space; (b) and (e): N_{10} qq-plot in Gumbel space; (c) and (f): N_{100} qq-plot in Gumbel space.



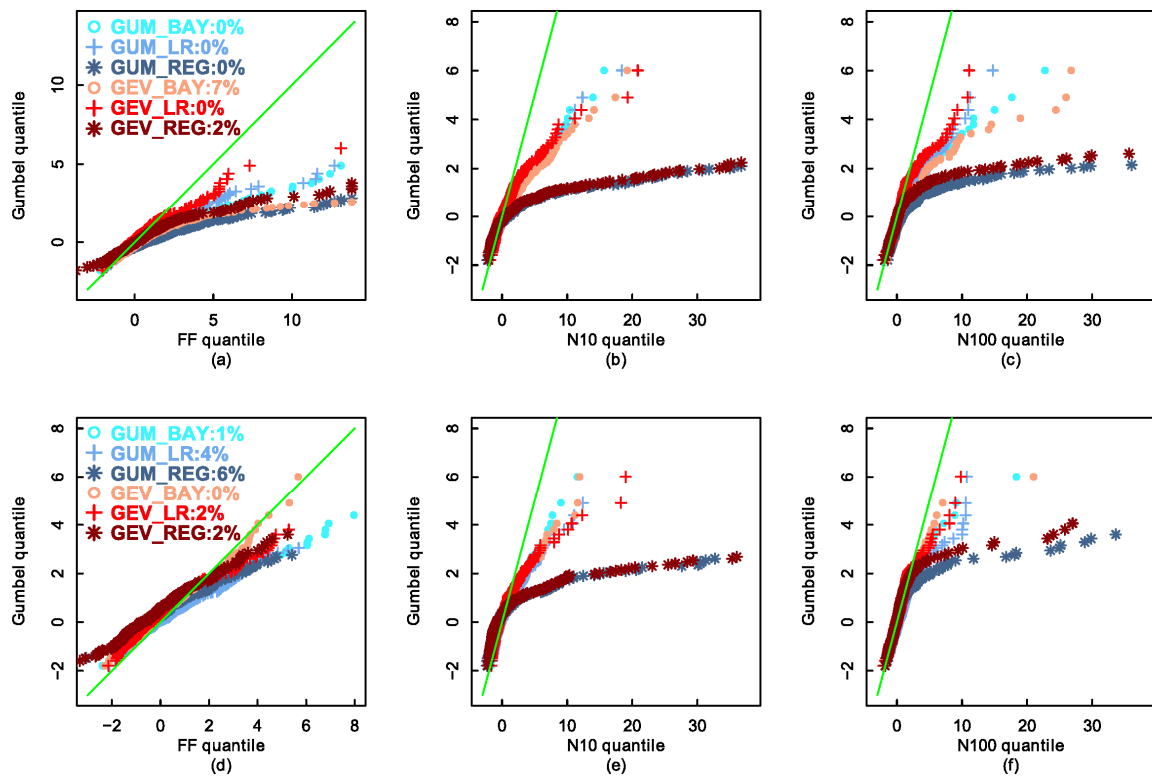
1200

1201

1202

1203

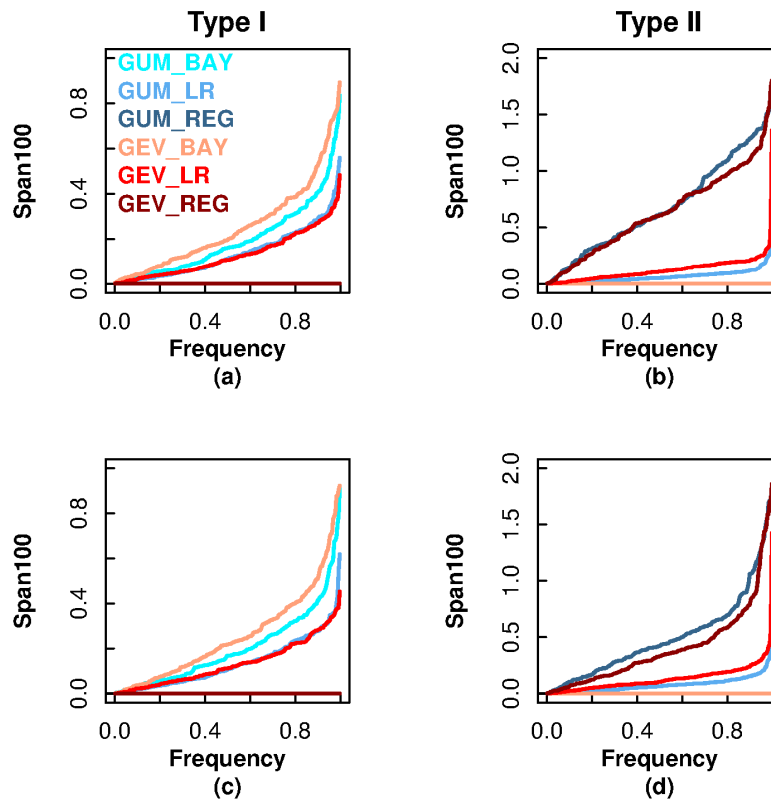
Figure 6. $pval$ pp-plot for local, local-regional and regional estimation of the GEV distribution (estimated distribution).



1204

1205 **Figure 7. Reliability diagnostics for six FA implementations (local, regional and local-regional, with**
1206 **Gumbel and GEV distributions). First row = estimated distribution, second row = predictive distribution.**
1207 **(a) and (d): FF qq-plot in Gumbel space; (b) and (e): N_{10} qq-plot in Gumbel space; (c) and (f): N_{100} qq-plot**
1208 **in Gumbel space.**

1209



1210

1211 **Figure 8. Stability diagnostic for six FA implementations (local, regional and mixed local-regional, with**
1212 **Gumbel and GEV distributions). Left = type I decomposition, right = type II decomposition. (a) – (b) =**
1213 **estimated distribution, (c) – (d) = predictive distribution.**

1214

1215