A comparison of regional flood frequency analysis approaches in a simulation framework

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2016WR018604

A comparison of regional flood frequency analysis approaches in a simulation framework

D. Ganora[1] and F. Laio[1]

[1]Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Torino, Italy

Key Points:
- Different regional approaches are compared in a simulation-based framework
- The regional models are applied to virtual scenarios with different degree of heterogeneity
- Spatially smooth estimation technique performs better in a wide range of conditions

Supporting Information:
- Supporting Information S1

Correspondence to:
D. Ganora,
daniele.ganora@polito.it

Citation:
Ganora, D., and F. Laio (2016), A comparison of regional flood frequency analysis approaches in a simulation framework, *Water Resour. Res.*, *52*, 5644–5661, doi:10.1002/2016WR018604.

Received 11 JAN 2016
Accepted 28 JUN 2016
Accepted article online 1 JUL 2016
Published online 31 JUL 2016

Abstract Regional frequency analysis (RFA) is a well-established methodology to provide an estimate of the flood frequency curve at ungauged (or scarcely gauged) sites. Different RFA approaches exist, depending on the way the information is transferred to the site of interest, but it is not clear in the literature if a specific method systematically outperforms the others. The aim of this study is to provide a framework wherein carrying out the intercomparison by building up a virtual environment based on synthetically generated data. The considered regional approaches include: (i) a unique regional curve for the whole region; (ii) a multiple-region model where homogeneous subregions are determined through cluster analysis; (iii) a Region-of-Influence model which defines a homogeneous subregion for each site; (iv) a spatially smooth estimation procedure where the parameters of the regional model vary continuously along the space. Virtual environments are generated considering different patterns of heterogeneity, including step change and smooth variations. If the region is heterogeneous, with the parent distribution changing continuously within the region, the spatially smooth regional approach outperforms the others, with overall errors 10–50% lower than the other methods. In the case of a step-change, the spatially smooth and clustering procedures perform similarly if the heterogeneity is moderate, while clustering procedures work better when the step-change is severe. To extend our findings, an extensive sensitivity analysis has been performed to investigate the effect of sample length, number of virtual stations, return period of the predicted quantile, variability of the scale parameter of the parent distribution, number of predictor variables and different parent distribution. Overall, the spatially smooth approach appears as the most robust approach as its performances are more stable across different patterns of heterogeneity, especially when short records are considered.

1. Introduction

The probabilistic analysis of discharge-related variables is a fundamental step to support, among other tasks, risk analysis, mitigation measures, and the design of hydraulic infrastructures. However, in basins where few or no discharge data are available, it is not possible to directly provide a robust statistical analysis; these cases belong to the "ungauged basins" family, which has been extensively studied in the last decade, also thanks to the Prediction in Ungauged Basins (PUB) initiative held by the International Association of Hydrological Sciences (IAHS). Different methodologies can be adopted for the purpose of predicting the variable of interest in ungauged basins, usually referred to as "regional models" in the literature.

Blöschl et al. [2013] have recently reviewed the most recent procedures to evaluate annual runoff, seasonal runoff, flow duration curves, low flows, floods and hydrographs in ungauged (or poorly gauged) basins. The book describes a variety of case studies; however, no clear consensus emerges about the methodology to be preferred. In fact, although all regional models are based on the fundamental concept of substituting the temporal information at a site by exploiting observations at other sites, different approaches exist to transfer information to the site of interest. The great variety of methods and implementations used in regional analysis is not surprising when considering the great variety of conditions (different variables, climatic regimes, soil variability, etc.) the methods are applied to.

In this context, *Wagener et al*. [2013] noticed that the information transfer can generally follow two paths. In the first case, catchments are grouped together according to their mutual distances computed on a set of selected basin characteristics, also known as descriptors (e.g., geographical location, area, mean elevation, precipitation regime, soil type, etc.); neighbor catchments are expected to have a similar hydrological behavior. Each group, called a "region," should be statistically homogeneous to guarantee that the basins

effectively behave in a similar way. For instance, when dealing with flood frequency analysis, a region is interpreted as homogeneous if the records observed in the catchments belonging to the region can be considered as drawn from the same underlying probability distribution. Also note that the regions may or may not be contiguous in the geographical space [e.g., *Ouarda et al*., 2001]. Prediction is performed by associating the ungauged basin to the closest region in the descriptor space, and by using the hydrological information of the region (records at the gauged sites) to estimate the variable of interest in the ungauged site.

The second path reported by *Wagener et al*. [2013] to perform a regional analysis considers a relationship, called mapping function, between the catchment descriptors and the hydrological variable under study. This approach does not require grouping similar catchments; however, the mapping function must be valid for the whole region. Prediction is performed applying the mapping function (e.g., a regression model) to the ungauged site, after the relationship has been calibrated using the data available in the gauged basins of the region.

This two-way classification, although able to highlight the main ideas underlying the regionalization procedures, does not resolve the whole complexity of regional models.

In the literature, only few works have systematically applied different regional models on the same data set to identify the most reliable procedure to be adopted for operational applications [e.g., *Kochanek et al*., 2013; *Renard et al*., 2013]. The reliability of each model is usually tested by comparing predicted results against the corresponding observed values [e.g., *Salinas et al*., 2013]. Although this approach is fundamental to define an operational model, it is strictly dependent on the considered case study. As a consequence, general conclusions can hardly be drawn.

The present work reconsiders several different regional modeling approaches for flood frequency analysis from a more general point of view; this is achieved by using a simulation-based framework to create virtual landscapes where all the statistical characteristics of the data are known, and reference variables are thus not affected by limited sample availability. The final aim of the work is not to identify the "best" approach for each scenario, but rather to recognize if a model is able to behave well in a wide range of conditions. In real world applications the actual pattern of heterogeneity cannot be easily identified, and thus it is particularly important to select a robust methodology that have stable performances in many different conditions. To the best of our knowledge, this is the first time in which a systematic application of different regional approaches is performed over a set of different scenarios, each one representing a different possible degree of heterogeneity of the landscape.

2. The Simulation Framework

2.1. Simulations Set-Up

The analysis is based on a two-step procedure, whose first step involves Monte Carlo simulations to generate a number of data sets which are used to feed different regional models in the second step. Each data set is formed by a collection of (virtually) observed time series at different locations which represent sequences of annual maxima. Note that the Monte Carlo framework can be set up in a number of ways by defining different characteristics of the virtual landscape, for instance the parent distribution, the generated sample length, the number of gauging stations, etc. In this paper, we provide a detailed description of some configurations, on the basis of reliable and representative characteristics, and a summary of an extensive sensitivity analysis based on a number of alternative configurations (details are provided in the supporting information).

The location of the gauging stations where the records are generated is defined as follows: the virtual landscape is defined over a one-dimensional unitary space where the coordinate $x \in [0, 1]$ is used to identify the location of each station; eleven equally spaced stations have been considered for this study (the effect of varying the number of stations has been investigated in the sensitivity analysis, see section 3.3). Note that the coordinate x is a generic basin descriptor; it can be interpreted either as a normalized geographical coordinate or as a watershed characteristic. We therefore refer in a first instance to a mono-dimensional descriptor space; extensions to two dimensions are discussed in section 3.3.

To test the regional modeling procedures, different scenarios are defined. The first scenario is a fully homogeneous region. In this case, the time series of annual maxima simulated at all virtual gauging stations are sampled from the same parent distribution; we thus use the same parameters over the whole spatial

Table 1. Symbols and Parameter Values for the Reference Scenarios Used in the Simulations

Variable	Value	Notes
Sample length	$n = 50$	
Number of stations	$n_s = 11$	Equally spaced along x
Return period	$T = 200$ years	
Parent distribution	GEV	Also used for prediction
Location parameter	$\theta_1 = 1$	
Scale parameter	$\theta_2 = 0.5$	
Shape parameter	$\theta_3 = -0.4 + 0.6 \cdot x$	LIN-H scenario
	$\theta_3 = -0.2 + 0.2 \cdot x$	LIN-L scenario
	$\theta_3^L = -0.3$; $\theta_3^R = 0.1$	STEP-H scenario[a]
	$\theta_3^L = -0.2$; $\theta_3^R = 0$	STEP-L scenario[a]

[a] θ^L is valid for $0 \le x \le 0.5$ and θ^R for $0.5 < x \le 1$.

domain to guarantee that the region is truly homogeneous. Alternative scenarios will mimic heterogeneous regions by allowing the parameters of the parent distribution to vary over the spatial domain.

The simulated records are sampled from a generalized extreme value distribution (GEV), whose quantile function reads [e.g., *Grimaldi et al*., 2011]:

$$Q_T = \theta_1 + \frac{\theta_2}{\theta_3}\left\{1 - \left[-\ln\left(1 - \frac{1}{T}\right)\right]^{\theta_3}\right\} \quad (1)$$

with θ_1, θ_2 and θ_3 being the location, scale and shape parameter, respectively. T is the return period in years associated to the quantile Q_T; other probability distributions have been considered in the sensitivity analysis (see section 3.3). For a set of parameters, the normalized frequency, or growth factor, K_T can be obtained as

$$K_T = \frac{Q_T}{\bar{Q}} \quad (2)$$

where $\bar{Q} = \theta_1 + \theta_2[1 - \Gamma(1 + \theta_3)]/\theta_3$ is the mean value of the distribution (the index-flood), $\Gamma(\cdot)$ being the Gamma function.

This work is focused on the estimation of K_T, congruently with typical applications in the field of regional flood frequency analysis based on the index-flood approach [*Dalrymple*, 1960]. In fact, the mean value is usually considered easier to be estimated than the growth factor, as it can be assessed locally using short samples or through regression models [*Laio et al*., 2011; *Rosbjerg et al*., 2013, section 9.3.4].

An example of virtual environment is reported in Figure 1a, where the 11 gauging stations are placed over the unitary domain. The figure reports a typical simulation generated by sampling a GEV parent distribution whose parameters are summarized in Table 1 (with θ_3 referred to the LIN-H scenario); grey dots in the figure denote the normalized annual series drawn from each station distribution (virtual records), while the true growth factor (solid line), computed over the spatial domain using the exact parameters of the parent distribution, is compared to the estimated growth factor (circles) obtained by the at-site fitting of the simulated records.

2.2. Regional Modeling

The second step of our analysis is the application of the regional approaches to the simulated data sets to provide estimates of K_T at different locations, referred to as target sites.

The modeling framework developed in this section has been designed to be representative of the main regional approaches currently used in real applications, although, as already mentioned, each regionalization method can be customized in different ways. Three approaches have been selected in this work, each one implemented in order to be unsupervised and reproducible, allowing their systematic application to many sets of simulated data without requiring manual case-specific tuning of the model.

The value of K_T is obtained by fitting the GEV distribution using the method of L-moments in all cases; L-moments are widely used statistics describing the shape of a probability distribution (analogously to the moments), based on linear combinations of order statistics. The L-moment of order 1, or λ_1, is equivalent to the mean value of the distribution, whereas the dimensionless L-moments ratio τ (or L-CV) and τ_3 (or L-skewness) respectively represent the variability and the skewness of the distribution. With λ_1, τ and τ_3 the three parameters of the GEV distribution can be estimated. For an introduction to L-moments the reader is referred to *Hosking and Wallis* [1997].

Since we are interested in the regional prediction of K_T rather than Q_T, it is sufficient to make (regional) predictions of τ and τ_3: the K_T value can in fact be obtained directly from equation (1) with $\lambda_1 = 1$ and the

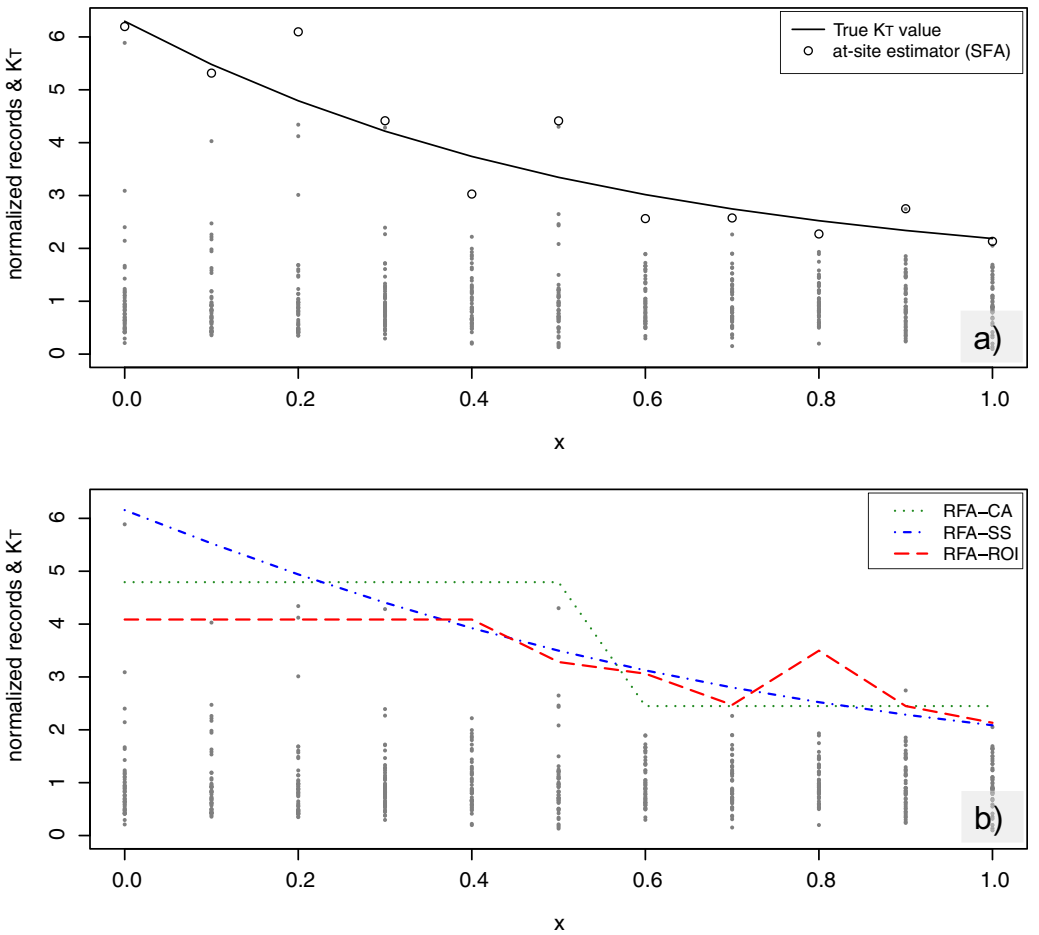

Figure 1. Example of simulation with 11 virtual gauging stations. Grey dots are the simulated data (i.e., the normalized annual maximum values) drawn from the GEV distribution divided by the sample average (with $\theta_1 = 1$, $\theta_2 = 0.5$, $\theta_3 = -0.4 + 0.6 \cdot x$). (a) The true value and the at-site estimator of K_T; (b) the regional estimators. A return period $T = 200$ years is considered.

regional τ and τ_3 values, as K_T from equation (2) is equivalent to Q_T where the mean of the distribution has a unitary value.

All methods implement a regional frequency analysis (RFA): based on a cluster analysis (RFA-CA) of the gauging stations; identifying similar gauging stations through the region of influence approach (RFA-ROI); or using a spatially smooth (RFA-SS) estimator of the variable of interest.

The RFA-CA method splits the descriptor space in a number of fixed, contiguous and nonoverlapping subregions. The number and size of the subregions (also a single region is a possible outcome) is defined by combining a hierarchical clustering algorithm and a test for homogeneity. Each subregion is characterized by a unique regional pair of τ and τ_3 values, thus resulting in a single K_T value. Prediction is performed by associating the target site to a specific region according to its location in the descriptor space.

The RFA-ROI method creates an ad hoc subregion for each target site by grouping together gauging stations by proximity in the descriptor space. The pooling group is tested for homogeneity. Prediction is performed by computing τ and τ_3 values using only the samples belonging to the group, and then computing K_T. Differently from the CA approach, the subregions dynamically change for different target sites. Different examples of application of the CA and the ROI methods can be found in *Rosbjerg et al.* [2013, section 9.2.3].

In the RFA-SS method a linear regression (i.e., a mapping function [*Wagener et al.*, 2013]) in the form $y = \alpha_0 + \alpha_1 x$ allows the interpolation of the L-moments at any target site along the spatial domain x, where y is in turn the sample τ and τ_3 (α_0 and α_1 are parameters to be estimated). K_T varies smoothly according to

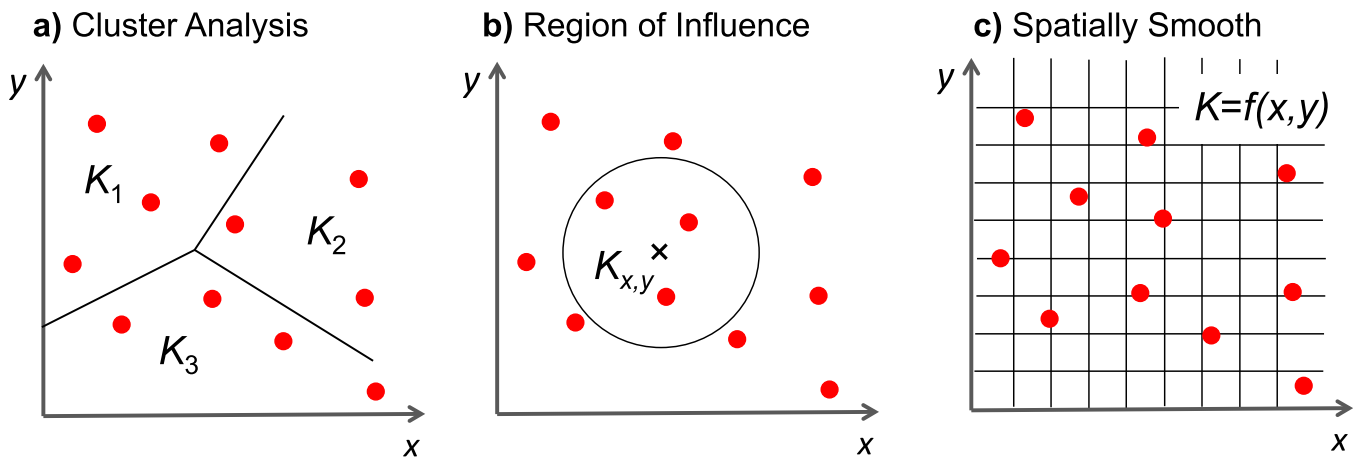

Figure 2. Sketch of the estimation procedure for the three regional models involved in the study: (a) Fixed subregions with a constant K_T value (RFA-CA); (b) variable pooling groups for different target points (RFA-ROI); (c) estimates K_T as a smooth continuous function in the descriptor space (RFA-SS). Red circles represent the gauging stations. The coordinates x and y represent a generic bi-dimensional descriptor space, but the models can work on lower or higher dimensional spaces.

changes in τ and τ_3. Significance tests are used to check if the parameters are statistically different from 0. An application following this approach can be found in *Laio et al.* [2011] with a smooth prediction of the first three L-moments; similar regression-based spatially smooth applications have been used for instance by *Kjeldsen and Jones* [2007] to predict the index-flood, and in *Stedinger and Tasker* [1985] and *Griffis and Stedinger* [2007] to directly predict the quantile for a fixed return period (thus not in the index-flood framework).

The regional approaches described above and the way they implement the regionalization process are sketched in Figure 2, while further technical details are reported in Appendix A. For the sake of clarity, the sketches refer to a two-dimensional spatial domain, while the algorithms can be applied to one-dimensional space as well as to cases with more dimensions. Furthermore, the application of the three regional methods to the virtual environment of Figure 1a is reported in the Figure 1b of the same figure. In the example the CA method detects two homogeneous subregions, providing a constant value $K_T \approx 4.8$ for $0 \leq x \leq 0.5$ and a constant, but different, value in the right part of the domain ($K_T \approx 2.5$ for $0.6 \leq x \leq 1$). The ROI method provides, in general, different K_T values for each target station, although in this case five stations on the left side of the domain show the same value $K_T \approx 4$ because they are associated to the same pooling group. Finally, the SS shows a smooth variability of K_T due to the smooth variability of regional L-moments.

In this work, also a Regional Frequency Analysis with a single Region (RFA-1R) has been considered as a baseline: in this approach a single region is assumed to be representative over the whole domain. The unique regional growth curve is the GEV distribution whose parameters are estimated by using the average sample L-moments obtained from all the available virtual records.

Beyond the regional approaches, at gauged stations also the at-site frequency analysis (SFA) has been performed to estimate the growth factor. The value of K_T is computed by directly using each set of at-site sample L-moments. This approach is again used as a baseline for comparison with other methods.

2.3. Assumptions in the Analysis Framework

The study is based on some simplifying assumptions that help keeping the simulation framework computationally tractable, but that allows one to highlight the main features of the studied models.

In a first instance, both at-site and regional estimates are based on the GEV probability distribution, that is the true parent used in the analysis. This avoids the misspecification of the probability distribution, thus allowing the system to provide results free from epistemic errors [see e.g., *Botto et al.*, 2014] that would be introduced in the analysis by the choice of an erroneous distribution. This hypothesis has been then relaxed to study scenarios in which the fitting distribution differs from the parent one (see the sensitivity analysis).

Moreover, the heterogeneity is defined according to the coordinate x (allowing the shape parameter θ_3 to vary as a function of x only), and the same coordinate is used by all the RFA models to drive the estimation

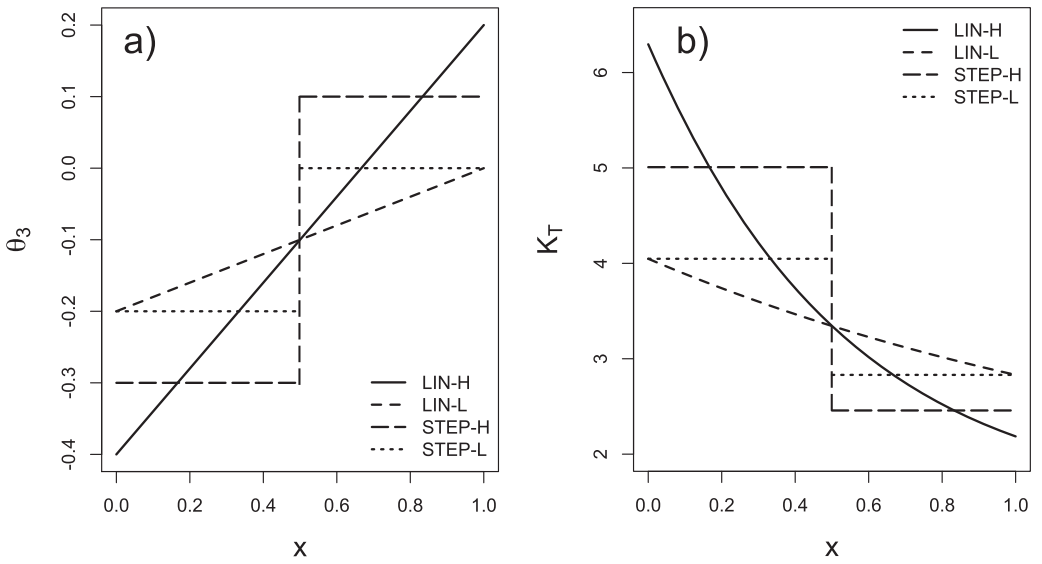

Figure 3. Variability of (a) the shape parameter θ_3 and (b) corresponding normalized quantile K_T in the four scenarios (equation (3)) obtained from the GEV distribution with $\theta_1 = 1$, $\theta_2 = 0.5$ for $T = 200$ years.

of K_T at any station. The regionalization problem, in fact, can be separated into two steps: the first is the identification of the driving variables which (partially) explain the between-site variability of the hydrological variable of interest. This step can be performed regardless of the regional model. In the second step, all regional models can be applied using the same driving variables, although each model uses this information in a different way. The driving variable is thus a common feature of CA, SS and ROI methods, but is used as a predictor variable in a very different way: CA uses x as the basis for clustering the stations; ROI does a similar action, but for creating distinct groups of gauged data for each target site; SS considers x as the independent variable in a regression. Note that, although the driving variable is considered known in this analysis, this does not mean that the estimation by the regional models is straightforward, as this is not equivalent to know the pattern of variation of θ_3. For instance, in the case of a scenario with the shape parameter θ_3 linearly varying along x, the RFA-SS model will search for a linear variability of the L-moments along x that are, however, nonlinearly related to both θ_3 and K_T.

In this study, we focus on two possible patterns of heterogeneity: (i) a linear variation of the shape parameter θ_3 of the GEV distribution along the x-coordinate, and (ii) a step-change in the parameter θ_3 which switches between two constant values. The latter pattern represents a situation in which there are actually two homogeneous subregions. Both the linear and the step-change patterns, referred to as LIN and STEP respectively, have been implemented with a "mild" heterogeneity, in which θ_3 has a low variability range (L), and with a "strong" heterogeneity, in which θ_3 has a high variability range (H). A summary of the parameters used in the different scenarios is reported in Table A4; the patterns are represented in Figure 3a, while Figure 3b of the same figure shows the true K_T value for each point of the domain corresponding to the θ_3 values of Figure 3a. Further analyses with different parameterizations are discussed in section 3.3 (Sensitivity Analysis) while details are provided in the supporting information. Also a truly homogeneous scenario ($\theta_3 = -0.1$ for all x) is considered as a reference situation.

3. Results

A systematic analysis of the behavior of the regional models under different scenarios (i.e., different degrees of heterogeneity) has been implemented with the following procedure:

1. the reference scenario is selected, i.e., the parameters of a GEV parent distribution are defined for each station s, following a specific pattern of heterogeneity;
2. the true K_T value is computed from the parent distribution at each station considering a $T = 200$ years return period; this value will be used as a reference to evaluate models performance;

3. the records are sampled at each station s from the parent distribution; each sample is of length n;
4. the at-site estimate (SFA) of K_T is computed from each record using the GEV distribution and $T = 200$ years; this value will be used as a benchmark result;
5. the RFA-CA model is calibrated using the whole set of records, then it is applied at each station s to provide the predicted K_T value (with GEV and $T = 200$ years);
6. analogously, the RFA-ROI, the RFA-SS and the RFA-1R models are applied to the simulated landscape;
7. error statistics are computed at each station for each model by comparing the true and estimated K_T values;
8. steps 3–7 are repeated for $N = 500$ different simulations (N has been defined as a compromise between results stability and computational load);
9. overall error statistics are obtained by averaging the errors obtained in the individual simulations.

For each simulation, time series are generated independently at each station, meaning that there is no correlation between concurrent values at different spatial locations. In fact, all the regional models under analysis explicitly require the samples to be uncorrelated to avoid including redundant information in the model, that would reduce the model robustness and results reliability. A discussion about the effects of inter-site correlation on quantile prediction in the RFA-CA approach can be found in *Hosking and Wallis* [1997, section 7.5.6], where the authors observe that moderate correlation is not a major concern in cluster-based regionalization. Inter-site correlation can also affect the reliability of the heterogeneity test of Hosking and Wallis, but corrections to the testing procedure are available [*Castellarin et al.*, 2008]. If the regression-based model is used, a generalized least square procedure should be adopted, thus requiring the estimation of the covariance matrix (an example of computation of the covariance structure between flood series in a regression-based approach, although not in the index-flood context, is reported by *Griffis and Stedinger* [2007]).

3.1. The Homogeneous Scenario

The first scenario under analysis is a truly homogeneous region which can be considered as a baseline. In fact, this is the base hypothesis of many regional applications, and consists of an area where every time series is generated from the same probability distribution, in this case the GEV parent distribution with parameters $\theta_1 = 1$, $\theta_2 = 0.5$ and $\theta_3 = -0.1$. This set of parameters provides a true K_T value 4.5 for all stations. Note that different sets of parameters would provide similar results, with a different dispersion of the prediction errors due to the different variance/skewness of the simulated records.

The application of the regional models to each of the 500 simulation leads to the results reported in Figure 4a, which shows the box plots of the relative error

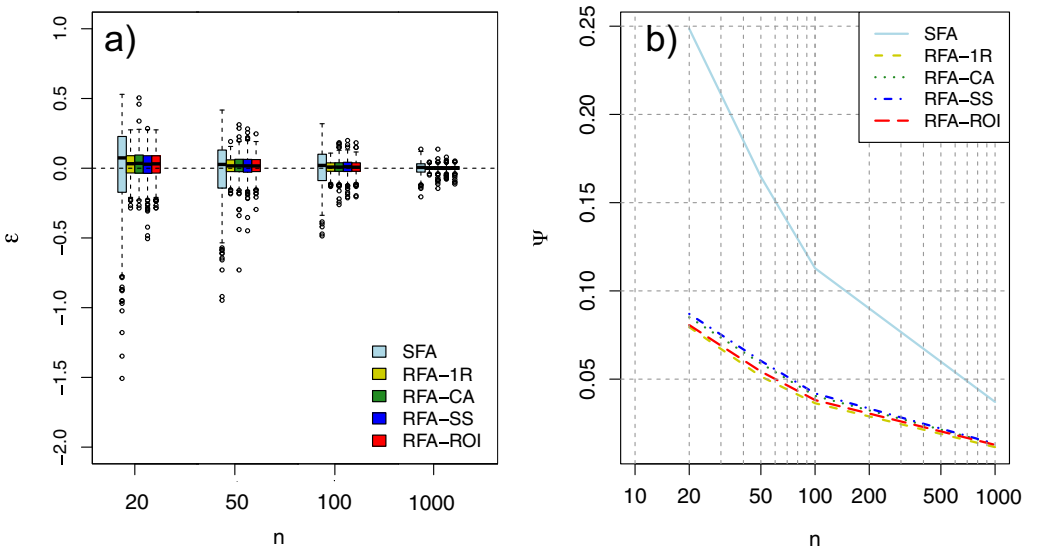

Figure 4. (**a**) Box-plot representation of the prediction error ε, based on 500 simulations and (b) resulting overall error Ψ for all models at one station. The scenario is based on a truly homogeneous region with a GEV parent with $\theta_1 = 1$, $\theta_2 = 0.5$ and $\theta_3 = -0.1$: the length of the simulated time series (n) varies on the horizontal axis; the return period is $T = 200$ years.

$$\varepsilon_{s,i} = \frac{K_{T;s} - \widehat{K}_{T;s,i}}{K_{T;s}}, \tag{3}$$

where K_T is the true normalized quantile, $\widehat{K}_T$ the normalized quantile obtained from one of the models, s indicates the station and i the simulation run. Each box plot represents a modeling approach and summarizes the variability of the relative error across the simulations. Box plots are grouped according to the sample length n of the simulated time series; four different homogeneous scenarios with different sample lengths have been investigated. Being the region homogeneous, all stations are characterized by equivalent results (except for small differences due to the sampling variability).

All models appear to be unbiased, and the at-site fitting (SFA) always provides larger variance than the regional models (RFA); this is expected as the RFA better exploits the information than the SFA, since RFA estimators are based on the joint use of multiple records. These results confirm the basic assumption of regional models, i.e., that more reliable results can be obtained by grouping together data from multiple stations, when applied to a genuine homogeneous region. A complementary result is that all the regional models provide equivalent results, being the difference between their error distributions not significant. Sample length does not affect these conclusions, but just acts on the variance of the results by reducing the variance with increasing n. Note that values of the sample length in the range $20 \leq n \leq 100$ can be considered realistic, with $n = 50$ a typical value representative of real regional applications; instead, $n = 1000$ is reported just for comparison to show the quasi-asymptotic behavior of the estimators.

Although the homogeneous case is quite straightforward to understand, a couple of error statistics can be conveniently introduced to facilitate the comparison of the results obtained for different scenarios. The first index is the mean absolute relative error, computed at each station s, as:

$$\Psi_s = \frac{1}{N}\sum_{i=1}^{N} |\varepsilon_{s,i}| \tag{4}$$

where N is the number of simulations performed in the analysis (e.g., $N = 500$). The second statistic is the mean absolute relative error computed for each simulation i:

$$\Omega_i = \frac{1}{n_s}\sum_{s=1}^{n_s} |\varepsilon_{s,i}| \tag{5}$$

where n_s is the number of stations considered in the analysis. The two error indices are useful to investigate in more detail the model efficiency (Ψ) and the model robustness (Ω).

Figure 4b, shows the variability of Ψ as a function of the sample length n. RFA approaches are basically indistinguishable, while SFA is characterized by significantly larger errors.

A different way to analyze the results is presented in Figure 5, which shows the statistic Ω_i computed for each simulation, thus resulting in 500 points on each plot; this representation allows one to directly compare any pair of models. Among RFA approaches, there is a large number of realization in which the Ω values fall on the bisector of the plot; this means that both models provide the same results, which happens because they are both able to properly detect that the region is homogeneous. Correct detection of homogeneity occurs (for $n = 50$) in 89.9% of the simulations for RFA-CA, 90.0% for RFA-SS and 85.4% for RFA-ROI, while the RFA-1R is inherently homogeneous by definition in all simulations.

Results provided by the regional models are similar despite the differences in methods for detecting homogeneity: the RFA-CA starts by considering a unique cluster and tests it for homogeneity; in most of the cases the cluster results homogeneous and the creation of subregions is thus inhibited; the RFA-ROI behaves like the RFA-CA although possible subregions are created with a different grouping algorithm. Instead, the RFA-SS computes at-site L-moments and performs regressions with the coordinate x as a regressor; in most of the cases the slope of the regression does not result to be significative; the linear model is thus dropped in favor of a constant value (i.e., the regression intercept, equal to the mean L-moment), which is adopted as the unique regional value.

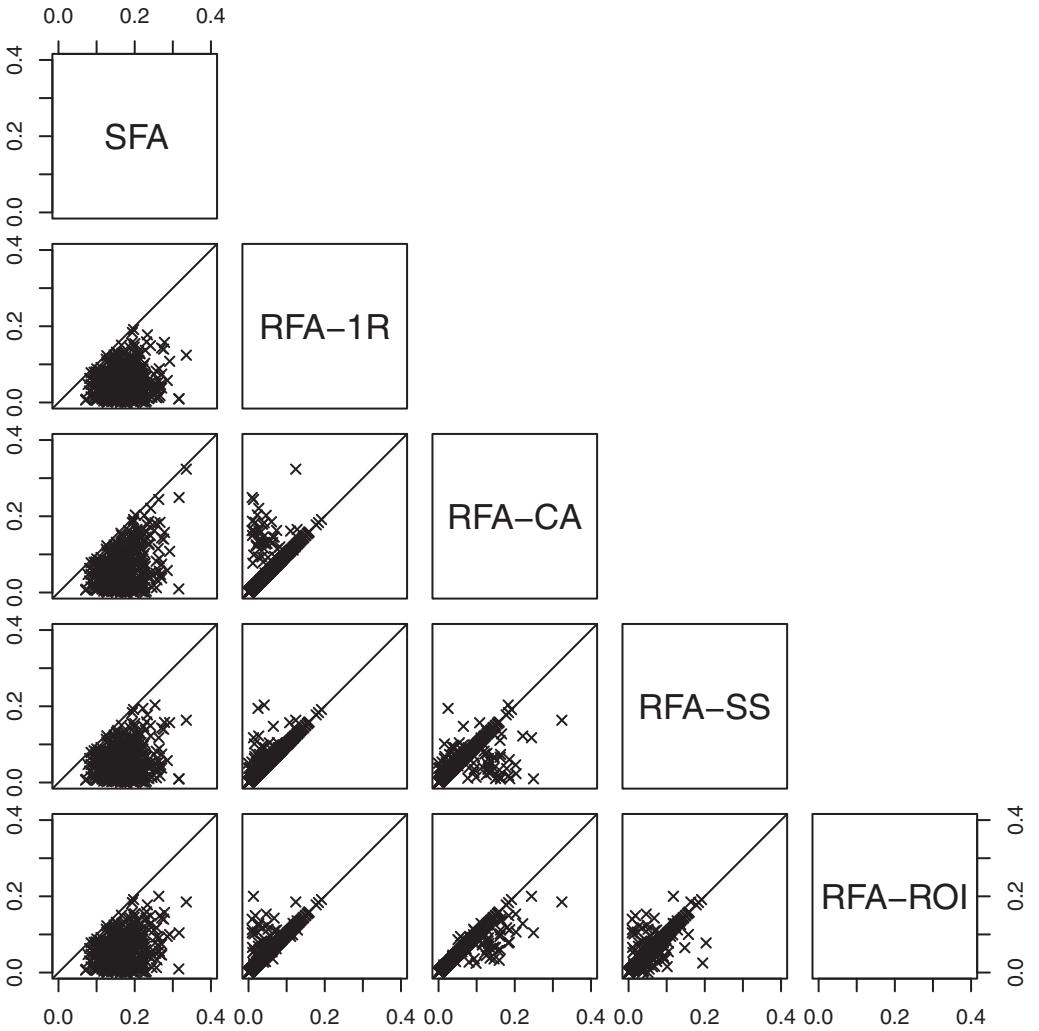

Figure 5. Comparison of couples of Ω values (see equation (5)) of each regional model for 500 simulations; results refer to the truly homogeneous scenario with $n = 50$.

The cases in which a regional model is not able to detect the homogeneity can influence the global performances of the method, summarized in Table 2 as the overall error E obtained by averaging Ψ_s over all stations (or, equivalently, by averaging Ω_i over all simulations),

$$E = \frac{1}{n_s}\sum_{s=1}^{n_s}\Psi_s = \frac{1}{N}\sum_{i=1}^{N}\Omega_i = \frac{1}{n_s N}\sum_{s=1}^{n_s}\sum_{i=1}^{N}\left|\frac{K_{T;s}-\widehat{K}_{T;s,i}}{K_{T;s}}\right|. \quad (6)$$

Results show similar performances of the RFA-CA, RFA-SS and RFA-ROI approaches, with RFA-1R representing the irreducible error value due to sample variability.

3.2. Effect of Heterogeneity

The at-site and regional modeling approaches have been then applied to the four heterogeneity scenarios defined in Table 1; the corresponding boxplots of prediction errors are shown in Figure 6. In contrast to the homogeneous scenario, here the location of the station along x is relevant for the analysis; therefore, all stations have been considered separately. Regional models show rather different performances, and this effect is exacerbated in the high-variability scenarios (plots LIN-H and STEP-H, respectively). In general, none of the regional models provides unbiased results at all stations. Moreover, it is worth noting that RFA methods

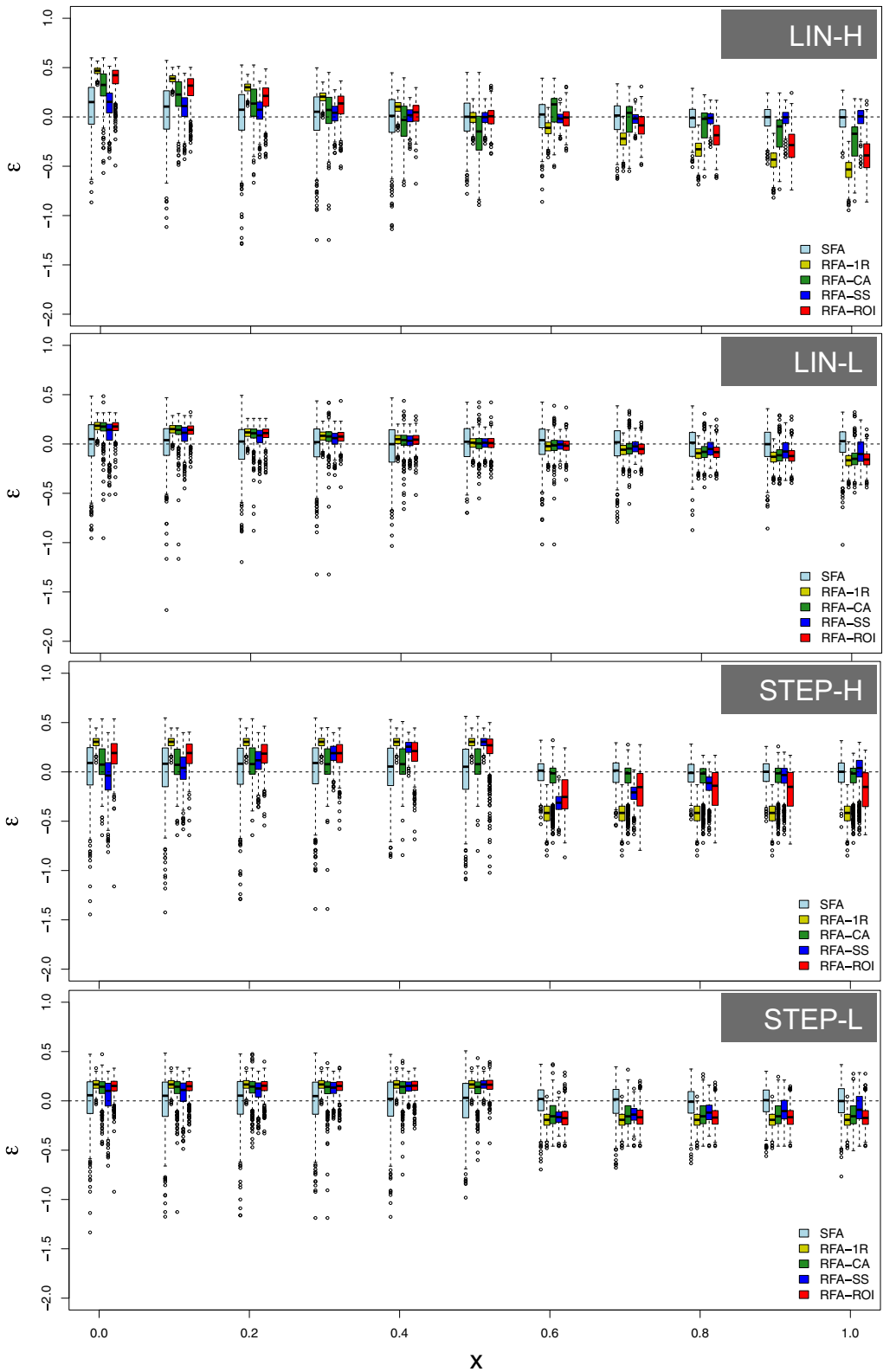

Figure 6. Boxplot of errors ε for SFA and RFAs methods at the 11 stations. Results are based on 500 simulations. The plots refer to: linear strong variability (LIN-H), linear mild variability (LIN-L), strong step-change (STEP-H) and mild step-change (STEP-L). The shape parameter θ_3 varies according to the scenario; the other parameters are $\theta_1 = 1$, $\theta_2 = 0.5$, $T = 200$ years and $n = 50$. Boxes denote the quartiles and the median value; whiskers extend for a maximum of 1.5 times the interquartile range above and below the box.

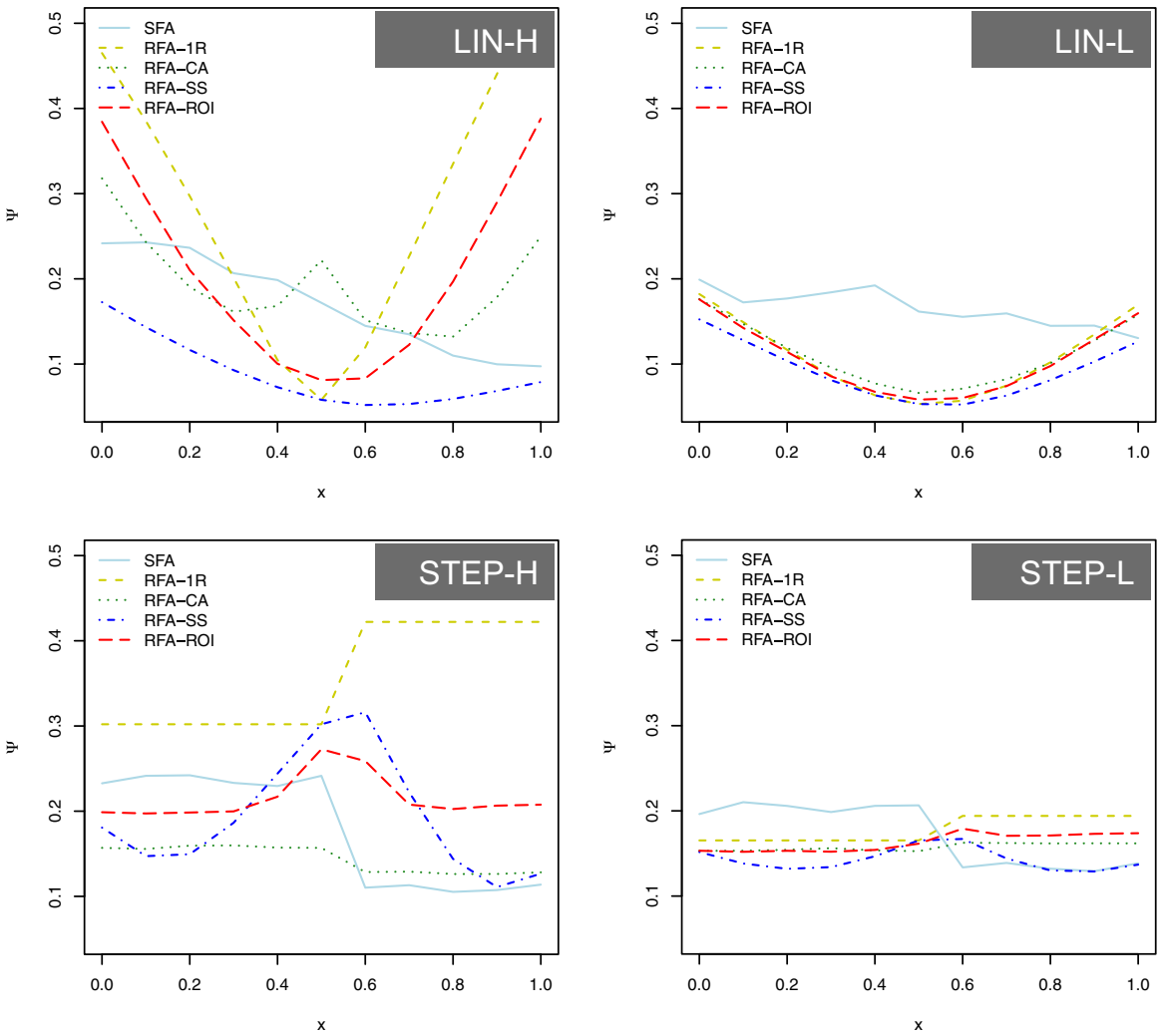

Figure 7. Variability of the overall relative error Ψ along x for each model. The plots refer to: linear strong variability (LIN-H), linear mild variability (LIN-L), strong step-change (STEP-H) and mild step-change (STEP-L). The shape parameter θ_3 varies according to the scenario; the other parameters are $\theta_1 = 1$, $\theta_2 = 0.5$, $T = 200$ years and $n = 50$.

perform generally worse than in the homogeneous case, while results from the SFA become comparable, if not better in some cases, than those of regional models.

While in a truly homogeneous region all RFA approaches result more or less equivalent, Figures 6 and 7 (where the error index Ψ is reported) show that models are quite sensitive to the degree of heterogeneity: when variations are continuous along x (LIN), the RFA-SS method performs significantly better than the others, as it provides smaller errors almost everywhere in the spatial domain. If the region is slightly heterogeneous, the gap between RFA methods performances is reduced, but the RFA-SS still provides better results.

In the step-change scenario, where two homogeneous subregions are hidden in a unique domain, good performances are, as expected, obtained by the RFA-CA method, which is able to clearly identify the

Table 2. Overall Relative Errors E (Equation (6)) Obtained With Different Estimation Methods (Columns) and for Different Sample Size (rows) for the Homogeneous Case With a GEV Parent With $\theta_1 = 1$, $\theta_2 = 0.5$, $\theta_3 = -0.1$ and $T = 200$ Years

n	SFA	RFA-1R	RFA-CA	RFA-SS	RFA-ROI
20	0.256	0.079	0.085	0.083	0.082
50	0.167	0.052	0.061	0.055	0.056
100	0.119	0.036	0.041	0.039	0.038
1000	0.038	0.011	0.013	0.012	0.012

subregions thanks to the cluster analysis procedure implemented in the algorithm. However, it is interesting to observe that while in the STEP-H scenario the RFA-CA approach clearly outperforms the others, with a moderate step change (STEP-L scenario) the RFA-SS method still provides better performances. In general, overall errors Ψ_s are almost constant in the whole domain for the RFA-CA; instead, the RFA-SS works well toward the boundary of the domain, while, as expected, it tends to erroneously "smooth" the sudden change in θ_3 in the central part of the domain because it is not able to represent the jump in the statistical descriptors; in fact, errors are higher near the step-change. Analogously, the RFA-ROI approach is sensitive to the step change in the STEP-H scenario. This is due to the fact that the pooling group is selected according to the distance from the target site; for target points close to the step-change, the pooling group is likely to include gauges from both the subregions, thus providing inadequate predictions. Finally, the 1R method, being based on a unique region, is not able to detect any kind of variability in the underlying distribution and always provides biased results. Global errors, reported in Table 3, summarize the conclusions through the index *E* computed as in equation (6).

A different view of results is shown in Figure 8, where the values of Ω_i are reported for the LIN and STEP scenarios. Each point represents the overall relative error of a single simulation and represents a possible real case study. In the LIN-H case, one can note that the RFA-SS outperforms RFA-CA and RFA-ROI for most of the simulations. RFA-CA show a twofold behavior that creates two clusters in the error plot; from the comparison with RFA-1R it is evident that in a number of cases the RFA-CA does not detect the variability and considers the whole region as homogeneous (points on the bisector in the 1R-CA scatter plot). The LIN-L case shows analogous results, expect for the RFA-ROI model which performs better than in the LIN-H case when compared to other methods.

The STEP-H case confirms the good performances of RFA-CA while compared to RFA-SS; however, in about 23% of the simulations, RFA-CA detects a unique homogeneous region, thus providing comprehensive results worst than RFA-SS. In such cases, also the RFA-ROI typically detects a unique region.

The performances of the SFA are comparable to those of the RFA methods for high levels of heterogeneity, while RFA methods still prevail when the degree of heterogeneity reduces and the scenario is more similar to the homogeneous case. However, a direct comparison between SFA and RFA is not straightforward and will be reconsidered in the Discussion and Conclusions section.

3.3. Sensitivity Analysis

To further support the findings of this work and give guidance for future applications, an extensive sensitivity analysis has been performed by changing many of the characteristics of the original simulation set-up. A summary of the results for the whole set of scenarios studied in the sensitivity analysis is reported in Table 4, while details are reported in the supporting information.

3.3.1. Record Length

The first analysis has involved different values of the sample length of the virtual records, $n = 20$ and $n = 100$ respectively, while the other characteristics have been kept equal to the reference scenarios. In both cases, general conclusions are similar to that of the reference case. Note however that, if the sample length is small ($n = 20$ cases), the RFA-SS outperform the RFA-CA also in the STEP-H scenario as the higher sample uncertainty makes the heterogeneity more difficult to be detected. This case suggests that the RFA-SS is more likely to be reliable when short samples are involved.

3.3.2. Number of Sites and Return Period

Conclusions similar to the reference case can be obtained by changing the number of stations in the domain: with $n_s = 51$ one obtains smaller errors due to the larger available data set, but the relative

Table 3. Overall Relative Errors *E* (Equation (6)) Obtained With Different Estimation Methods (Columns) and for Different Heterogeneity Scenarios (Rows)[a]

scenario	SFA	RFA-1R	RFA-CA	RFA-SS	RFA-ROI
LIN-H	0.171	0.289	0.199	**0.088**	0.204
LIN-L	0.168	0.107	0.110	**0.093**	0.106
STEP-H	0.180	0.358	**0.146**	0.189	0.203
STEP-L	0.169	0.178	0.158	**0.138**	0.162

[a]The parameters of the parent distribution (GEV) are $\theta_1 = 1$, $\theta_2 = 0.5$, while θ_3 is variable according to equation (3). The best performing method is highlighted in boldface.

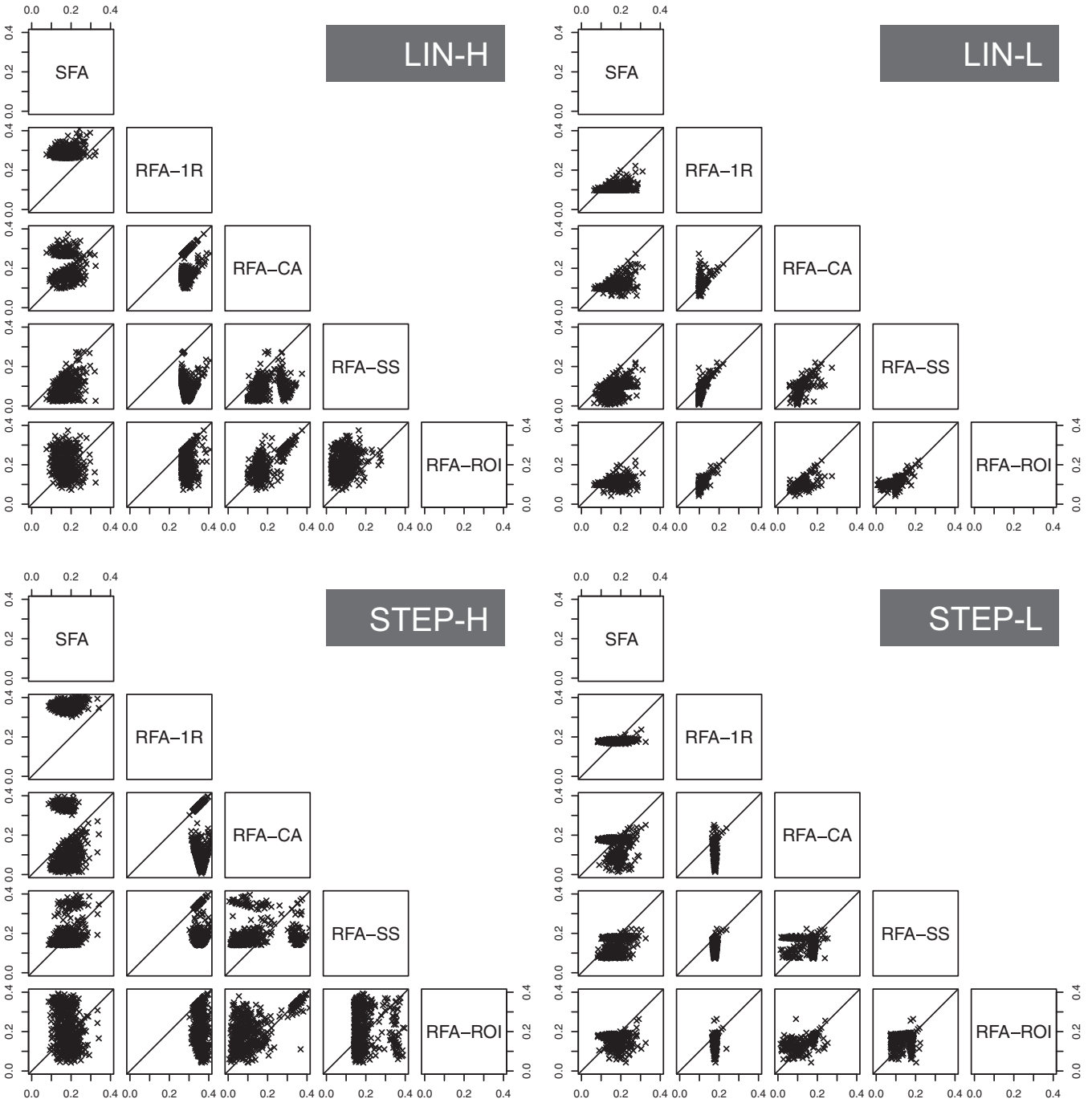

Figure 8. Comparison of couples of Ω values of each regional model for 500 simulations; the plots refer to: linear strong variability (LIN-H), linear mild variability (LIN-L), strong step-change (STEP-H) and mild step-change (STEP-L). The shape parameter θ_3 varies according to the scenario; the other parameters are $\theta_1 = 1$, $\theta_2 = 0.5$, $T = 200$ years and $n = 50$.

behavior of the regional models does not change. Moreover, the use of different return periods, equal to $T = 50$ years and $T = 500$ years respectively, just influences the magnitude of errors that are larger for higher return periods.

3.3.3. Variability of the Scale Parameter

Another possible pattern of heterogeneity has been studied by allowing the scale parameter θ_2 to vary along the x dimension and keeping the location and the shape parameters of the GEV as constant values (equal to 1 and -0.1 respectively). Also in this case, two linear and two step-change scenarios have been investigated (the detailed parameterization are reported in Table 4) which provide results similar to the reference scenarios.

Table 4. Overall Relative Errors E (Equation (6)) Obtained From the Sensitivity Analysis[a]

Sensitivity Target	Tested Conditions	Scenario Type	SFA	RFA-1R	RFA-CA	RFA-SS	RFA-ROI	Notes
n	$n = 20$	LIN-H	0.258	0.289	0.256	**0.158**	0.255	
		LIN-L	0.262	0.123	0.133	**0.121**	0.129	
		STEP-H	0.264	0.349	0.267	**0.254**	0.293	
		STEP-L	0.263	0.180	0.181	**0.167**	0.180	
	$n = 100$	LIN-H	0.126	0.291	0.149	**0.061**	0.147	
		LIN-L	0.119	0.103	0.100	**0.068**	0.098	
		STEP-H	0.132	0.360	**0.070**	0.164	0.117	
		STEP-L	0.123	0.179	0.125	**0.115**	0.140	
n_s	$n_s = 51$	LIN-H	0.169	0.266	0.124	**0.042**	0.124	
		LIN-L	0.167	0.093	0.088	**0.041**	0.087	
		STEP-H	0.174	0.352	**0.089**	0.148	0.112	
		STEP-L	0.170	0.177	0.116	**0.084**	0.130	
T	$T = 50$	LIN-H	0.099	0.167	0.112	**0.055**	0.115	
		LIN-L	0.097	0.062	0.064	**0.053**	0.062	
		STEP-H	0.101	0.202	**0.086**	0.113	0.116	
		STEP-L	0.099	0.101	0.091	**0.081**	0.093	
	$T = 500$	LIN-H	0.227	0.375	0.255	**0.109**	0.262	
		LIN-L	0.219	0.139	0.143	**0.120**	0.140	
		STEP-H	0.236	0.465	**0.196**	0.245	0.269	
		STEP-L	0.226	0.232	0.206	**0.183**	0.212	
θ_2	$\theta_2 = 0.3 + 4x$	LIN-H	0.166	0.129	0.109	**0.058**	0.098	$\theta_1 = 1,\ \theta_3 = -0.1$
	$\theta_2 = 0.4 + 2x$	LIN-L	0.166	0.075	0.088	**0.063**	0.079	
	$\theta_2^L = 0.3;\ \theta_2^R = 0.7$	STEP-H	0.161	0.221	**0.078**	0.112	0.095	
	$\theta_2^L = 0.4;\ \theta_2^R = 0.6$	STEP-L	0.165	0.106	0.081	**0.077**	0.091	
2-D space (x, y)	$y_s \sim U(0, 1)$	LIN-H	0.173	0.289	0.214	**0.090**	0.214	
		LIN-L	0.169	0.108	0.112	**0.096**	0.109	
		STEP-H	**0.183**	0.358	0.216	0.190	0.228	
		STEP-L	0.168	0.178	0.171	**0.139**	0.168	
LN3 distr. (*)	$\theta_3 = 0.86 - 0.54x$	LIN-H	0.153	0.307	0.136	**0.078**	0.118	$\theta_1 = 0,\ \theta_2 = 1$
	$\theta_3 = 0.62 - 0.18x$	LIN-L	0.145	0.108	0.097	**0.070**	0.094	
	$\theta_3^L = 0.73;\ \theta_3^R = 0.38$	STEP-H	0.152	0.363	**0.071**	0.176	0.088	
	$\theta_3^L = 0.62;\ \theta_3^R = 0.44$	STEP-L	0.144	0.184	**0.072**	0.112	0.091	
PE3 distr.	$\theta_3 = 0.73 + 6.77x$	LIN-H	0.099	0.246	0.091	0.113	**0.081**	$\theta_1 = 0,\ \theta_2 = 1$
	$\theta_3 = 1.68 + 2.02x$	LIN-L	0.104	0.099	0.079	**0.051**	0.074	
	$\theta_3^L = 1.12;\ \theta_3^R = 5.34$	STEP-H	0.107	0.368	**0.050**	0.161	0.062	
	$\theta_3^L = 1.68;\ \theta_3^R = 3.7$	STEP-L	0.105	0.179	**0.049**	0.095	0.064	
Parent distr. $\neq$ fitting distr.	Parent = GEV; fitting = LN3	LIN-H	0.163	0.280	0.194	**0.097**	0.205	GEV param. as in Table A.4
		LIN-L	0.149	0.105	0.107	**0.090**	0.104	
		STEP-H	0.173	0.349	**0.144**	0.194	0.208	
		STEP-L	0.152	0.176	0.153	**0.136**	0.160	
	Parent = LN3; fitting = GEV	LIN-H	0.176	0.322	0.154	**0.089**	0.127	LN3 param. as in (*)
		LIN-L	0.168	0.114	0.109	**0.079**	0.101	
		STEP-H	0.174	0.374	**0.084**	0.186	0.103	
		STEP-L	0.168	0.188	**0.085**	0.121	0.100	

[a]More details are reported in the Supporting Information. The best performing method is highlighted in boldface.

3.3.4. 2-D Domain

Another way to extend the analysis is to consider a domain with more than one dimension. In this case, the pattern of variability of the parameters still remains a function of the coordinate x only, according to Figure 3; however, each station is now characterized by a pair of coordinates (x, y), with the second one randomly sampled from a uniform distribution between 0 and 1. All the regional models apply to the full 2-D space, i.e., the algorithms (clustering, regression, etc) use a set of two coordinates as descriptors for their estimation procedures. The rationale behind this test is the analysis of the behavior of the regional models when nonideal information is used to support the regionalization process; in fact, the second coordinate is preliminary assumed to be useful, even if it does not have any relationship with the target variable, thus introducing some noise in the analysis. As already mentioned in section 2.3, this condition is very common in real-world applications where the true driver of heterogeneity may not be known or measurable. The present analysis suggest that the 2-D case is very similar to the 1-D scenario; however, although this topic deserves further investigations, preliminary results (not shown) highlight that relevant differences between the 1-D and the 2-D cases arise for longer record, still remarking the fact that sample uncertainty plays a fundamental role with typical record length values.

3.3.5. Alternative Parent Distributions

A further generalization has been obtained using the log-Normal and the Pearson type III distribution (both with 3 parameters) instead of the GEV: each distribution has been considered as the parent distribution and for the fitting of at-site and regional growth curves, analogously to what has been done with the GEV. The shape parameter has been allowed to vary along the x dimension following two linear and two step-change patterns (LIN-H, LIN-L, STEP-H and STEP-L respectively; see Table 4 for numerical values) in a way that K_T values equal to those of the GEV case are obtained at $x = 0$ and $x = 1$. For both distributions the RFA-CA approach turns out to be the most adequate in the step-change scenarios (even with small variability) while the RFA-SS still remains the preferable approach for linear heterogeneities, except in the LIN-H case with the Pearson type III where the RFA-ROI has better performances. Differences with respect to the GEV case may be attributed to the fact that both the log-Normal and the Pearson type III distributions have a lighter right tail than the GEV distribution, thus generally resulting in less skewed sample records. This leads to records that are less likely to include large extreme values and thus, in general, the subregion can be easier detected as homogeneous. Further simulations (not shown) with different sample lengths have highlighted that reducing n, and thus increasing the sample uncertainty, brings the situation back to the reference (GEV) case, with the RFA-SS approach performing better than other models also in the STEP-L case.

3.3.6. Different Parent-Fitting Distributions

Finally, a set of mixed scenarios with different parent and fitting distribution has been analyzed. Virtual records from the reference scenarios, with the GEV as the parent, have been processed with models using the LN3 distribution to fit the growth factor. Viceversa, scenarios generated with a LN3 parent have been fitted using the GEV distribution. Results, reported in the bottom block of Table 4, show that the errors are very similar to those obtained when the correct fitting distribution is adopted. Although these results are preliminary, they show that the ranking of the models does not change for the studied scenarios and thus highlights a negligible role of the choice of the distribution in this context. These findings agree with those obtained by [*Hosking and Wallis*, 1997, section 7.5.9] with similar return periods and sample lengths for the RFA-CA case only.

3.4. Summary of Results

The analysis has been initially applied to a truly homogeneous region, i.e., an area where all the records have been sampled from the same parent distribution. Results, coherently with classic hypotheses on regionalization, show that regional modeling, compared to at-site methods, enhances the prediction quality because it better exploits the available information. The regional approaches show similar performances and none of them turns out to be clearly preferable, even if the spatially smooth approach shows in general slightly better results.

Other interesting results emerge with increasing the degree of variability of the parent distribution: when the heterogeneity is due to continuously varying parameters of the parent distribution, the spatially smooth regionalization approach is always preferable. On the other hand, if a sharp change in the statistical characteristics of the environment occurs, forming two independent homogeneous regions, the choice between spatially smooth and cluster analysis methods depends on different factors: the first is preferable for mild heterogeneity and for high heterogeneity when the sample length is small; the second is preferable when long records are available and for nonheavy-tailed distributions.

4. Discussion and Conclusions

A systematic analysis of different approaches to estimate the normalized flood quantile has been performed in a simulation framework, with the primary aim of evaluating the relative performances, reliability and robustness of different regional models under controlled conditions.

All in all, these results suggest that the spatially smooth regional methods represent a valid regionalization strategy in the view of robustness as they can reliably handle a wider range of conditions. This finding is of practical importance because in real applications it is very difficult, due to data uncertainty and lack of process understanding, to clearly identify if a region is truly homogeneous or not.

In practice, the case of moderate heterogeneity is commonly encountered in real applications, for instance when the homogeneity hypothesis is relaxed to include a larger number of gauging stations in a cluster. This case turns out to be more realistic with respect to the high heterogeneity case, since strongly

heterogeneous situations may be reasonably detected by preliminary inspections of the data or analyzing the spatial variability of the regression residuals.

It is worth noting that other simulation studies have analyzed the effect of heterogeneity on the estimated quantiles and on homogeneity detection [see e.g., *Hosking and Wallis*, 1997; *Viglione et al*., 2007]. However, such studies focus only on the cluster-based regionalization techniques, giving no guidance when the group cannot be considered homogeneous and heterogeneity have to be explicitly accounted for.

In this work a more general framework to address heterogeneity has been developed. But, can these results be considered representative of the real environment? In a first instance, the choice of a rather "simple" underlying scenario structure (i.e., linear or step change) is a reasonable choice: in fact, it can be seen as first-order approximation of more complex structures and it allows one to study the more relevant features, avoiding the risk of introducing sources of noise in the simulations. The considered pattern of heterogeneity, linear and step-wise variation of the parameters of the parent distribution, may also resemble the model structure of the regional models (respectively the spatially smooth SS and the subregion-based CA/ROI). However, this similarity is only apparent as the regional algorithms act on (sample) L-moments and not directly on the parent parameters, which are linked one to the other by nonlinear equations. Results confirm that the RFA-SS model works well also in some STEP environments and, conversely, the RFA-CA/ROI can be preferable in some LIN scenarios. Therefore, the more advisable modeling strategy can be hardly related to the expected pattern of heterogeneity - which, by the way, is not known a priori - but can be more reliably related to the sample length. Moreover, estimation is performed on sample L-moments (and not on exact ones) that, due to sample uncertainty, may not show any evident LIN/STEP pattern (depending on the simulation run).

A further consideration concerns with the rather simple structure of the simulated system (e.g., a 1-D domain, or equally spaced stations) with respect to the complexity of a real hydrologic system. Although this difference could be partially attenuated by creating more complex scenarios (which, on the other hand, make the system much less controllable), this is not expected to really affect the results of the analysis because all the regional models would be based on the same base of knowledge; i.e., the same set of driving variables. Instead, the analyses show that the complexity of the environment has a stronger impact on the relative performances of the at-site frequency analysis (SFA) and regional frequency analysis (RFA). In fact, while in the homogeneous case the RFA always outperforms the SFA, in the heterogeneous cases one can observe that the SFA performances are often comparable to those of the regional models. This results is, however, only qualitative, as we can expect the SFA will have in the reality even better performances, since here the RFA methods are based on the "correct" driving variable, while in real-world applications the driving variable is not precisely known. Therefore, this study allows a comparison of the regional models themselves, but "favors" regional models when compared to at-site estimators. Although this result still requires to be investigated in detail, it clearly highlights the value of at-site data (and local estimation procedures) even in slightly heterogeneous regions, and suggests that future research could be focused on integrated local-regional procedures [e.g., *Kjeldsen and Jones*, 2007; *Ganora et al*., 2013].

Finally, while this study gives guidance and suggestions in general for the selection of regional models, the whole simulation framework can also be tailored to specific case studies by considering some known scenario characteristics (e.g., sample lengths, station location, etc.) and by testing the sensitivity of other elements which are not well known or measurable (e.g., degree of heterogeneity based on data observations, empirical correlation among records, etc.).

Appendix A: Regional Models Algorithms

All regional models require fitting the probability distribution parameters to obtain the design flood; this step has been done with the method of L-moments [e.g., *Hosking and Wallis*, 1997; *Grimaldi et al*., 2011]. The analyses have been performed in R [*R Core Team*, 2013] with the use of the contributed packages "lmom" [*Hosking*, 2015].

For RFA, the "regional" L-moments have been estimated at every target site following a specific procedure which exploits the information available in the whole spatial domain, but with different algorithms depending on the considered method.

A1. RFA-1R

This is a reference case in which the regional L-moments are computed as the average of the sample L-moments calculated at each virtual station. Estimated L-moments and, consequently, the quantile K_T are the same for each target site in the region.

A2. RFA-CC

The RFA-CC approach defines a number of fixed nonoverlapping subregions over the spatial domain. Within each subregion, the regional L-moments are constant and are computed as in the RFA-1R case, thus providing a unique frequency curve for each subregion. The delineation of the subregions is performed once, with the following algorithm:

1. an agglomerative custering algorithm is applied using the average distance method [see e.g., *Kaufman and Rousseeuw*, 2009] to create a dendrogram hierarchical structure of the gauged stations; the distance between stations is calculated along the x coordinate (x and y in the 2-D case);
2. the procedure starts considering a unique cluster which includes all the stations;
3. the cluster is tested with the heterogeneity measure H, which compares the observed variability of the at-site sample τ values with its expected value obtained for a truly homogeneous region similar to the tested one (i.e., with the same number of stations, sample lengths and regional L-moments). The variable H is obtained through simulations according to the procedure of *Hosking and Wallis* [1997, section 4.3]. The region is considered homogeneous if $H \leq 1.5$. We have considered the limit 1.5 as an adequate compromise between the strict requirement of homogeneity (which would imply $H < 1$) and the necessity to have an operationally large enough region neglecting a small heterogeneity ($1 \leq H \leq 2$);
4. if the cluster cannot be considered homogeneous, it is divided in two sub clusters according to the dendrogram hierarchy;
5. each sub cluster is separately tested as in point 3 and, if it results heterogeneous, it is split again;
6. points 4 and 5 are iterated and the procedure quits when all the clusters are homogeneous and all the stations have been allocated to a cluster.

A3. RFA-ROI

The RFA-ROI approach associates a group of gauging stations to each target site, using a proximity criterion. The regional L-moments estimated at the target site are computed as in the RFA-1R case, but considering only the records included in the pooling group. As a consequence, the estimated frequency curve is, in general, variable among different target sites. The pooling group is selected according to the following algorithm:

1. the first-try pooling group includes all the available stations;
2. the pooling group is tested with the heterogeneity measure H, and it is considered as homogeneous if $H \leq 1.5$;
3. if the pooling group cannot be considered homogeneous, a new pooling group is obtained by removing the farthest gauged station (distance computed along x in the 1-D case and considering both x and y in the 2-D case);
4. steps 2 and 3 are iterated until the region may be regarded as homogeneous;
5. the procedure is repeated for all the target sites.

A4. RFA-SS

The RFA-SS approach allows the estimation of the L-moments (in this specific case L-CV and L-skewness) at each point of the domain through the use of two regression functions whose predictor variable are the coordinates of the point of interest. The regional L-moments are thus allowed to vary smoothly over the domain (each one varies linearly, independently of the other); as a consequence, the estimated frequency curve is, in general, variable among different target sites. The procedure is calibrated as follow:

1. sample L-moments are computed at the stations;
2. a linear regression is performed for each L-moment, based on the coordinate x as the predictor variable;
3. the regression variable is tested for significance (t-Student test at the 5% level): in case the test is not satisfied, only the intercept is considered (which is equivalent to consider a constant mean value of the L-moment over the whole domain);

4. in the 2-D domain case, three different regressions are analyzed with different combinations of the predictor variables (other than the intercept): only x, only y and both coordinates (x, y). After testing each model, the one with the highest adjusted coefficient of determination R^2_{adj} is adopted (models not passing the t-Student test are dropped).

Acknowledgments
Funding from the Italian Ministry of Education, Universities and Research (FIRB-RBFR12BA3Y project), and the ERC Consolidator Grant 2014 "CWASI - Coping with water scarcity in a globalized world" are acknowledged. The work is based on simulated data. Virtual records can be generated by sampling the selected parent distribution according to the scenarios reported in the text.

References

Blöschl, G., M. Sivapalan, T. Wagener, A. Viglione, and H. Savenije (Eds.) (2013), *Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales*, Cambridge Univ. Press, N. Y.

Botto, A., D. Ganora, F. Laio, and P. Claps (2014), Uncertainty compliant design flood estimation, *Water Resour. Res.*, *50*, 4242–4253, doi: 10.1002/2013WR014981.

Castellarin, A., D. Burn, and A. Brath (2008), Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem?, *J. Hydrol.*, *360*(1-4), 67–76, doi:10.1016/j.jhydrol.2008.07.014.

Dalrymple, T. (1960), Flood frequency analyses, *Water Supply Pap. 1543-A*, U.S. Geol. Surv., Reston, Va.

Ganora, D., F. Laio, and P. Claps (2013), An approach to propagate streamflow statistics along the river network, *Hydrol. Sci. J.*, *58*(1), 41–53.

Griffis, V., and J. Stedinger (2007), The use of GLS regression in regional hydrologic analyses, *J. Hydrol.*, *344*(1–2), 82–95, doi:10.1016/j.jhydrol.2007.06.023.

Grimaldi, S., S. C. Kao, A. Castellarin, S. M. Papalexiou, A. Viglione, F. Laio, H. Aksoy, and A. Gedikli (2011), Statistical hydrology, in *Treatise on Water Science*, vol. 2: The Science of Hydrology, edited by P. Wilderer, chap. 2.18, pp. 479–517, Elsevier, U. K., doi:10.1016/B978-0-444-53199-5.00046-4.

Hosking, J. (2015), *L-moments*, R package, version 2.5. [Available at https://cran.r-project.org/web/packages/lmom/index.html.]

Hosking, J., and J. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge Univ. Press, U. K.

Kaufman, L., and P. Rousseeuw (2009), *Finding groups in data: An introduction to cluster analysis*, vol. 344, John Wiley, Hoboken, N. J.

Kjeldsen, T., and D. Jones (2007), Estimation of an index flood using data transfer in the UK, *Hydrol. Sci. J.*, *52*(1), 86–98.

Kochanek, K., B. Renard, P. Arnaud, Y. Aubert, M. Lang, T. Cipriani, and E. Sauquet (2013), A data-based comparison of flood frequency analysis methods used in France, *Nat. Hazards Earth Syst. Sci. Discuss.*, *1*, 4445–4479.

Laio, F., D. Ganora, P. Claps, and G. Galeati (2011), Spatially smooth regional estimation of the flood frequency curve (with uncertainty), *J. Hydrol.*, *408*, 67–77, doi:10.1016/j.jhydrol.2011.07.022.

Ouarda, T., C. Girard, G. Cavadias, and B. Bobee (2001), Regional flood frequency estimation with canonical correlation analysis, *J. Hydrol.*, *254*(1-4), 157–173.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna.

Renard, B., et al. (2013), Data-based comparison of frequency analysis methods: A general framework, *Water Resour. Res.*, *49*, 825–843, doi: 10.1002/wrcr.20087.

Rosbjerg, D., et al. (2013), Prediction of floods in ungauged basins, in *Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales*, edited by G. Blöschl et al., pp. 189–226, Cambridge Univ. Press, N. Y.

Salinas, J., G. Laaha, M. Rogger, J. Parajka, A. Viglione, M. Sivapalan, and G. Blöschl (2013), Comparative assessment of predictions in ungauged basins–part 2: Flood and low flow studies, *Hydrol. Earth Syst. Sci.*, *17*(7), 2637–2652.

Stedinger, J., and G. Tasker (1985), Regional hydrologic analysis. 1. Ordinary, weighted, and generalized least-squares compared, *Water Resour. Res.*, *21*(9), 1421–1432.

Viglione, A., F. Laio, and P. Claps (2007), A comparison of homogeneity tests for regional frequency analysis, *Water Resour. Res.*, *43*, W03428, doi:10.1029/2006WR005095.

Wagener, T., G. Blöschl, D. Goodrich, H. Gupta, M. Sivapalan, Y. Tachilawa, P. Troch, and M. Weiler (2013), A synthesis framework for runoff prediction in ungauged basins, in *Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales*, edited by G. Blöschl et al., pp. 11–28, Cambridge Univ. Press, N. Y.