SERRA manuscript No.
(will be inserted by the editor)

Stochastic evaluation of simple pairing approaches to reconstruct incomplete rainfall time series

Daniele Pedretti · Roger D. Beckie

Received: date / Accepted: date

Abstract Two-station pairing approaches are routinely used to infill missing information in incomplete rainfall databases. We evaluated the performance of three simple methodologies to reconstruct incomplete time series in presence of variable nonlinear correlation between data pairs. Nonlinearity stems from the statistics describing the marginal peak-over-threshold (POT) values of rainfall events. A Monte Carlo analysis was developed to quantitatively assess expected errors from the use of chronological pairing (CP) with linear and nonlinear regression and frequency pairing (FP). CP is based on *a priori* selection of regression functions, while FP is based on matching the probability of non-exceedance of an event from one time series with the probability of non-exceedance of a similar event from another time series. We adopted a Generalized Pareto (GP) model to describe POT events, and a t-copula algorithm to generate reference nonlinearly correlated pairs of random temporal distributions distributed according with the GP model. The results suggest that the optimal methodology strongly depends on GP statistics. In general, CP seems to provide the lowest errors when GP statistics were similar and correlation

D. Pedretti, R. Beckie
Earth, Ocean and Atmospheric Sciences University of British Columbia (UBC) 2207 Main Mall, Vancouver BC V6T1Z4 E-mail: dpedretti@eos.ubc.ca

became linear; we found that a power-2 function performs well for the selected statistics when the number of missing points is limited. FP outperforms the other methods when POT statistics are different and variables are markedly nonlinearly correlated. Ensemble-based results seem to be supported by the analysis of observed precipitation at two real-world gauge stations.

Keywords Missing data · Rainfall · Chronological pairing · Frequency pairing · Nonlinearity · Copulas

1 Introduction

The analysis of missing rainfall data has been long addressed in the literature (e.g. [12,27,30]). An increasing number of sophisticated approaches have blossomed in the recent years in the attempt to obtain optimal methodologies to reconstruct missing data. Examples are kriging algorithms, artificial intelligence or neural networks (e.g. [14,15,20,21]), which allow for multiple stations to be simultaneously used to infill missing points at a specific location. Yet, it seems that simpler approaches such as two-station chronological and frequency pairing are still widely adopted among practitioners (e.g. [24]). The popularity of these methods may stem from the limited number of available and complete gauge stations, and the practicality of these algorithms, which are often quickly programmable in a spreadsheet.

The most common and widely known approach is possibly chronologically pairing (CP) with linear regression. It consists of regression-based estimation of missing measurements at a primary location (showing missing data) using observations at a nearby secondary, more complete location. Least-squares or other best-fitting techniques are applied to obtain regression coefficients. A schematic example of linear regression is depicted in Figure 1-1.

There are several potential weaknesses associated with this approach. We highlight for instance that reconstructed data series from regression-based approaches often underestimate the variance of the original data series [13] . A suite of methods, including the maintenance of variance extension (MOVE) techniques, the Generalized MOVE (GMOVE), the modified Kendall-Theil Robust Line (KTRL2) and the Robust Line of Organic Correlation (RLOC), do not suffer from this problem (e.g. [11–13,32,17,18]. Another difficulty is that variables may be nonlinearly correlated. This intrinsically renders a linear regression estimation error-prone. The paired data from rainfall time series are also typically much more scattered (i.e. more noisy and poorly correlated) than those shown in Figure 1-1 (e.g. [24]). In small mountainous catchments, this can be also linked to topographic control of local convective cells, which strongly affects precipitation even at short spatial scales [10].

We show in this work that the nonlinear correlation between variables stems from the difference in the probabilistic distribution of rainfall events recorded at each individual station. In general, rainfall distributions are positively skewed and display heavy-tailed distributions which are typically described in terms of the probability of rainfall peaks exceeding a fixed threshold (in short, peak-over-threshold or POT). Examples of POT distributions proposed in the past are Generalized Extreme Values (GEV), Gamma or Generalized Pareto (GP) [19,22,24,26,31]. Below the thresholds, no specific distribution is usually observed.

To deal with nonlinearity, one may opt to embed nonlinear regression models into CP estimations. Identifying optimal nonlinear functions is often not trivial, as a link between nonlinear regression functions and rainfall statistics may be non trivial. Moreover, scattering may hide the presence of nonlinear structures. These aspects are explored in detail in the first part of this work. More complex approaches like variable transformations or data regularization

have been proposed in the past (e.g. [3]). However, it seems that the actual implementation of these concepts is still not common in routine applications.

An appealing alternative methodology to effectively deal with nonlinearity and that avoids the selection of a regression model is frequency pairing (FP). The approach is derived from engineering applications such as prediction of flooding return periods or rainfall depths at ungauged locations. Most of these techniques were developed at the end of the 1980s (e.g. [1,23,29]) but only recently applied to the estimation of missing rainfall (e.g. [24]). FP is based on the probability of non-exceedance (P_E) of an event at the primary location compared to P_E of another event at a secondary location. The methodology is graphically explained in Figure 1-2. In each dataset, an event of given rainfall intensity is uniquely associated with a P_E value; a missing datum at a specific time at a primary location is assigned a value corresponding to P_E of the same event measured at a secondary location. In the provided example (Figure 1-2), station 2 records an event $r_2 = 2.5$ at a generic time; if station 1 is missing an event at the same time, the latter is assigned a value $r_1 = 0.25$ since r_1 and r_2 match the same probability ($P_{E1} = P_{E2} = 0.6$). It is worth mentioning that reconstructed datasets using FP preserve all the distributional moments of the original distribution, including the variance.

There are also potential weaknesses associated with the FP methodology. This approach relies on a direct relationship between probabilities at different stations; as such, a first requirement may be to obtain reliable empirical distributions from finite datasets, a task that can be complicated in case of limited data. This is especially true when the data are power-law distributed and the observations only cover a limited range and do not represent the full variability of the distribution, specifically the tails (e.g. [7]).

In this work we addressed two relevant questions regarding two-station pairing approaches. (a) What errors should we expect from simple approaches

to infill missing databases? (b) How reliable are such predictions? We attempted to provide an answer by first developing a Monte-Carlo (MC) framework, where the ability of CP and FP is tested against synthetic reference time series. Benchmarking data are generated using a t-copula approach to follow statistical distributions commonly associated with rainfall databases. We assessed each method by evaluating the ensemble average of the difference between reference and estimated missing values as a function of the number of missing data to be estimated and for different statistical rainfall distributions. In addition we evaluated the methods through two real-world case studies with a similar geographical configuration and comparable daily rainfall statistics.

Our purpose was to provide a direct indication for practitioners aiming to quantitatively evaluate the efficiency of unsophisticated two-station pairing approaches when data pairs can be affected by nonlinearity and the number of missing points is variable. We emphasize that the tested methodologies may not be the best available algorithms to optimally reconstruct incomplete time series. Indeed, we intentionally did not test more sophisticated approaches (including multiple stations or modified CP methods using MOVE or similar corrections). While an exhaustive study of all methods goes beyond the scope of this analysis, we also note that the comparison of different infilling methods have been already evaluated (e.g. [3, 4, 8, 14, 16, 20, 25]).

The paper is structured as follows. Section 2 introduces initial relevant aspects of this analysism such as the presence of nonlinearity in the correlation between data pairs and theoretical aspects regarding the behavior of the GP model. Section 3 presents the methodologies to be evaluated; it first addresses CP methods and then the FP methods. Section 4 describes the MC-based analysis, including the generation of synthetic realizations based on the t-copula approach. Section 5 presents the results for the ensemble-based analysis, emphasizing the difference between estimates made in presence of linear

correlation and those based on nonlinear correlations. Section 6 describes the application of the tested methods to two real-world case studies and examines the results. The paper ends with the general conclusions drawn from this analysis.

2 Nonlinearity between data pairs associated with POT statistics

The role of different POT statistics to generate nonlinear correlations between data pairs is illustrated. Compiled reviews of POT statistics revealed that the Generalized Pareto (GP) model is able to fit daily precipitation observed in several experimental sites in the world [31]. While other studies indicated that alternative models can be applied to fit POT statistics [26], we adopted the GP as a working model to develop our analysis. This is done without loss of generality, as any power-law model can be applied to generate nonlinear correlations.

Understanding the behavior of the GP function and a few theoretical aspects about convergence criteria related with subsampling errors is required before addressing the case of partially sampled distributions and the use of infilling methodologies.

2.1 Behavior of the GP model

The three-parameter GP model is part of the family of power-law parametric models (e.g. [2,6,28]). The probability density function for a rainfall event (r) can be written as

$$g(r|k,\sigma,\theta) = \left(\frac{1}{\sigma}\right)\left(1+k\frac{(r-\theta)}{\sigma}\right)^{(-1-\frac{1}{k})}, \quad (1)$$

for $\theta < r$, when $k > 0$ or $\theta < r < \theta - \sigma/k$ when $k < 0$. The special condition $k = 0$ has not been explored in this work. According to common formalisms, k is termed the "shape" parameter, σ is the "scale" parameter and θ is the "location" parameter.

Figure 2a depicts the behavior of a few selected analytical CDFs of the GP model characterized by different statistics. We set $\theta = 0$, as this parameter does not affect the shape of resulting distributions. It can be observed that, for a constant σ, the total number of large or extreme values (i.e. the tail of the distribution) is uniquely controlled by k, while for a constant k the tailing depends on σ. When both parameters change, tailing is controlled by the combination of both k and σ. For instance, tailing is more pronounced (i.e. there is more probability in the tail of the distributions) for the pair $k = 0.01, \sigma = 5$ than for the pair $k = 0.5, \sigma = 1$.

Figure 2b illustrates the expected errors related with the statistical subsampling. This is calculated as the ensemble mean of root mean square errors between the analytical solution of the GP models (G) using predefined parameters k and σ and the estimated distributions (G') calculated from best-fit parameters from a sample of n randomly generated values following the same GP distributions. For one individual realization of size n, the error (ε') is calculated as

$$\varepsilon'(m) = \sqrt{\frac{1}{n}\left(\sum G(x|\Omega) - G[x|\Omega'(m)]\right)^2} \tag{2}$$

where Ω' is the vector containing the estimated parameters (k' and σ') and m is the random realization. The ensemble average of individual errors is calculated as

$$\varepsilon = \frac{1}{N_{MC}} \sum \varepsilon'(m) \tag{3}$$

where $m = 1, ..., N_{MC}$. We set $N_{MC} = 10^4$. The GP random numbers were generated using the "gprnd.m" native function in the MATLAB environment.

The results indicate that ε decreases by a factor $\sqrt{n}$. More interestingly, the required number of sampling points for a specific amount of error is strongly controlled by the tailing. For low k, the number of points to obtain a similar ε is independent from σ; however, for large k, the errors are larger for larger σ. For instance, to obtain an average error of $\varepsilon = 10^{-2}$, the sample size should be close to $n \approx 8 \times 10^1$ for $k = 0.5, \sigma = 1$ but increasing to $n \approx 2 \times 10^2$ for $k = 0.5, \sigma = 5$. Rearranging these results, Figure 2c shows the distribution of ε depending on k for different sample sizes and different values of σ. It can be observed that for negative k values the difference between $\sigma = 1$ and $\sigma = 5$ are very small and $\varepsilon \to 0$ as $k \to -\infty$.

2.2 Correlation of synthetic rainfall pairs embedding GP distributions

We now illustrate how the specific combination of k and σ between stations results in a nonlinear correlation between data pairs. We assumed uncorrelated uniformly (U) distributed rainfall events below threshold and correlated, GP distributed POT events among stations. Correlated multivariate variables are generated using a copula-based approach (e.g. [5,9]). Copula approaches provide a suitable method to account for any type of marginal distribution, including non-parametric ones.

The specific methodology used to generate correlated random variables uses t-copulas. Similar to Gaussian copulas, t-copulas also allow for any marginal distributions to be imposed, and involve the definition of a correlation matrix, which expresses the correlation coefficients and cross-correlation coefficients between pairs of variables (ρ_{ij}), and a degree of freedom (ν). For $i = j$ (auto-correlation), $\rho_{ij} = 1$. The larger ρ_{ij} for $i \neq j$, the better correlated becomes

a specific synthetic pair of variables. On the contrary, the lower ρ_{ij}, the more 'noisy' or scattered becomes the correlation. We set $\nu = 1$ for all simulations; the specific value does not affect our conclusions (ν has to be positive to generate non-Gaussian multivariate distributions). In this work, we used the algorithm suggested by Mathworks and easily programmable in a MATLAB environment; details about this implementation can be found in software house web page (http://www.mathworks.com/help/stats/examples/simulating-dependent-random-variables-using-copulas.html).

In Figure 3, the resulting correlation between three hypothetical stations, S1 ($k = 0.5$, $\sigma = 2.5$), S2 ($k = -0.1$, $\sigma = 5$) and S3 ($k = 0.1$, $\sigma = 1$) is shown. These values are similar to those experimentally found in a reference mining site in Peru (described below) and in the range of values observed in typical gauge stations [31]. We set $\rho_{12} = 0.75$ and $\rho_{13} = 0.5$, respectively for pairs S1-S2 and S1-S3, to illustrate the effects of data scattering on the resulting correlation between data pairs.

Three important aspects deserve special attention. (a) Extreme rainfall events are much larger for S1 than for S2 and S3, due to different shape factors, while the range of S2 is larger than S1 due to the difference in scale factor (consistent with the theoretical behavior shown in Figure 2). (b) Scattering is quite small for the pair S1-S2 when compared with the pair S1-S3. This is a consequence of the larger ρ_{12} than ρ_{13}, which determines that S1 has a stronger relationship with S2 than with S3. (c) Using different GP distributions to generate the correlated variables, data pairs show nonlinear correlation - that is correlation that depends upon the value of the variable.

To examine the latter point in detail, we evaluated the coefficients of determination (R^2) for each pair of variable, which helps to identify the shape of this correlation. R^2 is estimated from the application of a linear or nonlinear regression model to these data and defined from the sum of the squared dif-

ference between each fitted and observed value and their mean. For the pairs S1-S2, a coefficient of determination based on linear regression ($R^2_L = 0.61$) provides much lower values than a coefficient of regression based on a power-2 model ($R^2_P = 0.90$). This occurs since the correlation between pairs closely resembles a power-law function with exponent 2. For the pairs S1-S3, scattering is more intense and both linear and nonlinear regression coefficients drop ($R^2_L \approx R^2_P \approx 0.3$). This suggests that the higher noise partially masks the nonlinearity between variables, despite the use of different POT statistics between S1 and S3. These empirical observations may be valid only for the adopted POT statistics. The combination of other statistics may generate a different shape in the correlation, which can be fit for instance by another power-law function.

Although not graphically shown, it is worth mentioning that the use of the same POT coefficients generates a linear correlation between variables. Also for this case, a lower ρ may generate random scattering and thus hide the presence of linearly correlated structures. Note that t-copulas implemented in this work do not generate autocorrelation among values in the individual time series. In this sense, we are implicitly not considering clustering of rainfall events, but independent identically distributed events within each rainfall series. This assumption is made without any loss of generality, since the tested methodology does not require any autocorrelation or clustering in individual time series.

3 Tested pairing methodologies

3.1 Chronological pairing

The CP method is based on a regression model between two time series (Figure 1). We considered two random vectors (r.v.) collecting random rainfall events

(r) at two different stations in the catchment. One station (primary) is missing data, which can be estimated using existing measurements from the secondary station. We assumed for simplicity that in both time series, rainfall events are simultaneously collected (except missing data) and that the time lag between each measurement is constant and equal to 1 day. Each r.v. (r_{s1} and r_{s2}) is composed by N_S samples A general regression equation can be defined as

$$r_{s1}(\tau) = f(r_{s2}(\tau)|\Omega_c) + \epsilon \tag{4}$$

where $\tau = 1, ..., N_S$, f is a generic linear or nonlinear regression function, ϵ is the vector containing the residuals, and Ω_c is the vector containing the regression model parameters. Least-square methods can be applied to obtain the best set of Ω_c minimizing ϵ.

Two specific CP models were used in this analysis. The first is the linear-regression-based approach (CP-L) with two regression coefficients (m and q) and in which the rainfall at the secondary station is estimated from the rainfall measurements at first station as

$$r_{s2}(\tau) = m r_{s1}(\tau) + q + \epsilon_L \tag{5}$$

where ϵ_L is the error associated with the linear model. The second is a power-law-based approach (CP-P), with two regression coefficients (a and b) and in which the rainfall at the secondary station is estimated from the rainfall measurements at first station as

$$r_{s2}(\tau) = a r_{s1}^{b}(\tau) + \epsilon_P \tag{6}$$

where ϵ_P is the error associated with the power-law model. The quality of these regression models can be also checked by calculating R_L^2 and R_P^2.

3.2 FP approach

FP establishes a direct relationship between events displaying the same P_E (Figure 1). Similarly to CP, we assumed that the primary station is missing an event at a time τ, and that a second (for simplicity, simultaneous) event is observed at a secondary station. For each r.v., P_E can be defined as

$$P_E = Pr(r(\tau) < \max(r)), \tag{7}$$

where $\max(r)$ is the maximum recorded event of each r.v. After sorting the two r.v. in increasing order, P_E can be empirically found as

$$P_E(r) = \frac{1}{N_s} \sum I\{r(\tau) \leq r\}, \tag{8}$$

where

$$I\{r(\tau) \leq r\} = \begin{vmatrix} 1 \text{ if } r(\tau) \leq r \\ 0 \text{ otherwise} \end{vmatrix}.$$

Using (8), each event is associated with a unique P_E, specific for each r.v.; hence, the same P_E could define two events ($r_1(\tau)$, $r_2(\tau)$) which could display the same or different rainfall intensity. The FP approach postulates that a missing event in the primary database can be estimated knowing the intensity of the rainfall event at the secondary stations, and knowing P_E at both locations. Thus, imposing $P_E(r_{s1}) = P_E(r_{s2})$, one can easily find r_{s1} from the observation of r_{s2}.

P_E can be modelled by a parametric function $G(r|\Omega_{Fi})$, where G is a cumulative distribution function (CDF) of any specific parametric model. If Ω_{Fi} are the parameters for that specific model for the ith station, then the missing value r_{s1} can be estimated in four steps using the FP approach:

1. fit $P_E(r_{s2})$ with a model $G_2 = G(r_{s2}|\Omega_{F2})$ to obtain the parameters describing the distribution of events at r_{s2};
2. find the analytical value of G_2 corresponding the event $r_{s2}(\tau)$;
3. fit $P_E(r_{s1})$ with a model $G_1 = G(r_{s1}|\Omega_{F1})$ to obtain the parameters describing the distribution of events at r_{s1};
4. invert G_1 to obtain $r_{s1}(\tau)$, imposing $P_E(r_{s1}(\tau)) = P_E(r_{s2}(\tau))$.

The last step can be done either analytically (knowing G_1^{-1}) or numerically, for instance using a minimization approach. Although not directly addressed in this work, the methodology could be also extended to non-parametric distributions.

4 Monte-Carlo analysis

A general stochastic approach based on a classical MC analysis was developed to quantitatively test each approach. The testing framework is general, can be applied to any specific parametric model and is independent of the presence of clustered rainfall events. We chose four sets of representative parameters, generating four different scenarios. In scenario 1 and 2, the same statistics (k and σ) are applied to both stations, and the correlation between pairs becomes linear. In scenarios 3 and 4, statistics are different between stations and correlation becomes nonlinear. The specific values for each scenario are listed alongside with the results. The four analyzed scenarios cover a broad range of linear and nonlinear cases which give rise to complex correlations and require detailed analysis to be fully understood and explained.

We used the t-copula approach described above to create a total of $N_{MC} = 10^4$ realizations for each combination of these parameters. For each set of parameters, we fixed the statistical parameters that define the distributions

of the primary and secondary rainfall data, the length of the time series (N_S) and the number of missing data to be estimated (N_R). For each scenario,

1. we generated two correlated random time series of rainfalls, one to be used as a primary time series, and the other as a secondary time series.
2. we selected a random reference time window (RTW) on the primary time series, of size N_R; these values were removed from the time series and used as reference (but unknown) values to compare with estimated values;
3. we estimated the best-fitting regression parameters between the two distributions using two linear and nonlinear regression models, with the RTW values removed from both time series (for the same period of time); subsequently, we applied the CP to reconstruct the missing RTW of the primary station using the regression model and the removed RTW data from the secondary information;
4. we found best-fitting CDF of both distributions without RTW; subsequently, we applied the FP approach between paired stations to estimate the RTW missing data.

Each realization was assumed to contain five cycles of two distinct hydrological seasons (dry and wet). We assumed for simplicity that no rainfall events occurs in the dry seasons, and focused on the reconstruction of missing rainfall events concentrated in the wet seasons and that are characterized by the POT distribution. We rounded the size of the database length to $N_S = 10^3$, and assumed that $N_S/2$ events fell in the range $0 < r \leq \theta$, while the remaining $N_S/2$ events fell within the range $r > \theta$ (POT events) and followed GP models with different k and σ. We set $\rho_{ij} = 0.5$, which led to a scattering between pairs of variables which is similar to the typical observations (Figure 3).

After an initial sensitivity analysis using $\rho_{ij} = 1$, we observed that scenario 1 and 2 led to linear correlation between variables, while scenario 3 and 4 gave

rise to nonlinear correlation resembling a power-2 function (similar to the example in Figure 3a). This was especially true when $\sigma = 5$, for which we observed a larger R_P^2 than for $\sigma = 1$. Based on this empirical observation and recalling (6), we tested a CP-P approach imposing b=2. We did not test other b factors; the results from this specific selection are discussed in the next section.

From the MC analysis, an error is defined as the difference between the estimated value and the true but unknown reference values. For each m-th realization ($m = 1, ..., N_{MC}$), the root mean square errors (RMSE) and sum of absolute errors (SE) were calculated respectively as

$$\mathrm{RMSE}(m) = \sqrt{\frac{1}{N_R}\sum_{i=1}^{N_R}\left(\frac{r_i'(m) - r_i}{r_i}\right)^2} \tag{9}$$

$$\mathrm{SE}(m) = \sum_{i=1}^{N_R} |r_i'(m) - r_i| \tag{10}$$

where r_i' and r_i refer to the estimated rainfall the reference values, respectively. Note that, because of the different normalization in RMSE and SE, the two errors measure two different quantities. In (9), the fluctuations $r_i'(m) - r_i$ are normalized by the local values r_i; thus, RMSE provides the average fraction of error of the set of estimates, which includes both POT events and rainfall below threshold. On the other hand, SE is a useful measure for water balance calculations as it can be interpreted as an average cumulative error of rainfall depths due to the presence of missing values.

5 Results

We initially analyzed scenarios where the selection of POT statistics results in linear correlation between pairs. In scenario 1 ($k = 0.5$ and $\sigma = 1$; Figure

4a and 4e) we observed that the performance of each method depends on the number of missing points: for $N_R < 10$ all methods perform similarly.; for $10 < N_R < 10^2$ CP-L and CP-P tend to outperform FP; for $N_R > 10^2$, CP-P generates very large errors (RMSE>> 2), while the errors for FP and CP-L remain below RMSE=2. The better performance of CP-L may be explained considering that the selection of the same statistics generating copula-based random time series leads to linear correlation between variables. In this case, a linear regression becomes more suitable than nonlinear models.

For small N_R, CP-P also works well. This likely occurs since power-law models can be approximated by linear regression (e.g. by a first-order Taylor expansion) for small errors. However, as N_R exceeds a critical value (which depends on the selected POT parameters) CP-P provides much larger errors compared to CP-L. In this case, the threshold was estimated through MC analysis as $N_R \approx 10^2$.

When N_R approaches the data record size, fewer data points are available for the estimation of the distribution statistics or regression coefficients and their estimates are very uncertain (RMSE $\rightarrow 10^3$). Interestingly, for CP-L and FP this effect is reached only when only a few points remain ($n \approx 20$). It is unlikely that in practice one would try to estimate long time series based on such limited data. Thus, subsampling should not be a problem for this type of approach, at least in the range of values tested here.

The departure of FP from CP-L is not dramatic for this first analyzed case. FP does not suffer from the problem of CP-P on large N_R, explaining why FP errors remain comparable with those of CP-L (ranging between 0.5 <RMSE< 1). This is also confirmed by the observed behaviour of SE (Figure 4e), which is very similar for both cases, being only slightly larger for FP than for CP-L. Consistent with the observation about RMSE, the cumulative errors for CP-P also tend to high values when $N_R > 10^2$. It can be observed that SE increases

linearly with the number of missing points, which is expected considering that this error expresses a cumulative sum of errors that linearly depends on the number of missing points.

In the scenario 2 ($k = 0.5$ and $\sigma = 5$; Figure 4b and 4f), we obtained comparable errors with the previous scenario from both CP-L and CP-P for $N_R < 10^2$. However, both methods tend to more markedly outperform FP, which provides larger errors than in scenario 1. This is especially true observing SE (Figure 4f) and may suggest that a higher σ provides larger errors for FP than in the case of lower σ. Also for this case CP-L performs quite similarly to CP-P in the limit of a reduced number of missing points; this could again refer to the accuracy of a first-order approximation of the power-law model for low number of missing points. Note that the departure of FP from the other methods also occurs for a limited number of missing values ($N_R < 10$).

We then analyzed scenarios where the selection of POT statistics results in a nonlinear correlation between pairs. In scenario 3, we set $k_1 = 0.5$, $k_2 = 0.1$ (suffixes indicate the number of the station) and $\sigma = 1$ for both stations. The results are illustrated by Figure 4c and 4g. We observed that FP clearly outperforms CP-L and CP-P, especially when the number of missing data increases. The poor ability of CP-L is explained considering the nonlinearity between variables; the poor ability of CP-P should be linked to both the selection of ρ and to the fact that a poor correlation hides the effects of nonlinearity associated with the selection of different POT coefficients (as in Figure 3). Resulting correlations have a different power-2 structure than for higher ρ values. Being independent from the selection of the regression model, FP seems therefore recommended for this case.

In scenario 4 ($k_1 = 0.5$, $k_2 = 0.1$ and $\sigma = 5$; Figures 4d and 4h) a different behavior was observed. For a wide range of missing data, both FP and CP-P perform reasonably better (especially when compared to CP-L). Errors in

general are larger than in the third scenario, which is associated with a larger σ (which requires more simulations to converge to similar errors than $\sigma = 1$, as shown in Figure 2c). It should be noted, however, that the relative difference in cumulative errors (SE) for each individual method are slightly reduced compared to the previously analyzed case with lower σ. Interestingly, CP-P performs (somewhat) better than FP, although this is true only for an intermediate range of total missing rainfall data. This is due to the better identified power-2-like correlated variability of synthetic distributions for larger σ when setting $\rho = 1$, and thus statistically more likely to occur for lower ρ values.

In light of these results, the performance of each method seem to be strictly associated with the estimated POT statistics. CP methods may be preferred to FP methods in the case of similat POT distributions between the two stations, which creates linear correlation among data pairs. However, FP would not provide excessive additional errors compared with the other tested methodologies for linear correlation. On the other hand, when the POT statistics between stations are different and the variables are nonlinearly correlated, FP is preferred over the other methods. From our results, this is especially true when the average error per measurement has to be estimated (measured through the RMSE). While the selection of CP-L is implicitly not adequate, a selection of a specific nonlinear regression mode seems hazardous. The power-2-based CP performed reasonably well only when the specific combination of variables determine nonlinear correlations with similar shapes; however, it may be not easy to determine *a priori* which type of regression model performed better depending on the combination of parameters underlying the two POT distributions. This is especially true when ρ is low and data are scattering.

For all tested cases, we found low coefficients of variation of the MC simulations ($CV < 1$). This suggested that the variability among fluctuations is not affecting the ensemble means; as such we did not directly address this aspect

in the present analysis but rather chose to focus on the difference in terms of expected values. It should be noted that our conclusions aimed to emphasize the effect of nonlinearity on the selection of simple pairing approaches. The results may not be universally applicable, as we did not analyze other scenarios (for which the shape of correlated pairs can be different than a power-2-like structure, or the number of missing points is larger than in our analysis).

The impact of the distribution of values below the threshold has not been directly addressed in the MC analysis. Rainfall events below θ are typically characterized by an uncorrelated noise $U = (0, \theta]$, and they should not affect the impact of POT statistics on nonlinear correlation. Although not tested, we argue that the presented MC framework could be also easily extended to incorporate values below this threshold. We finally remark that the tested CP methodologies are intentionally simple and not designed, for instance, to deal with maintenance of variance or other statistical effects on the resulting estimated outputs. Testing of more sophisticated techniques was beyond the scope of this work.

6 Analysis of Two Real-World Scenarios

The MC analysis provided an estimate of the impact of POT statistics on the selection of the pairing approach with the lowest ensemble-based error. We evaluated if these results also apply to the analysis of real-world observations, which can be considered as individual realizations from an ensemble of equally plausible, alternative realizations characterized by similar POT statistics. To this end, we performed a cross-validation to evaluate the behavior of CP and FP approaches to reconstruct missing rainfall measurements in two different sites. The first site is a small catchment at a mine site located in the Peruvian Andes, while the second site is located nearby Vancouver, BC.

6.1 Real-World Scenario 1

The Peruvian site is located at about 4300 m and characterized by two distinct hydrological seasons (Figure 5a). The wet season generally takes place from October to April and the dry season from May to September. In the dry season, precipitation becomes sporadic and the total cumulative rain is negligible compared to the wet season. Daily rainfall events (r) were recorded from 2007 to 2013 at three gauge stations (S-B, S-Y and S-Q). The gauge station S-B suffered from a partial mechanical failure for a period of about 60-70 days during the wet season October 2011 - April 2012. Several gaps also exist since April 2012 and available data are not considered reliable since then (Figure 5a). In the period 2007-2011, which is considered highly reliable, the existing S-B database consists of 1710 daily rainfall values (out of 1825 possible values, thus leading to 115 missing data). 592 days experience no detected rainfall (excluding missing data).

Station S-Y, located approximately 4 km from S-B, was active and continuously collecting rainfall data since 2007, including for the period where gaps occurred at S-B. An additional rainfall station, S-Q, is located approximately 5 km away from S-B, although only collected reliable data from 2007 to 2008. Application of a linear fitting model between daily measurements at S-Y and S-B yields $R_L^2 = 0.35$ (using information from 2007 to 2011) and of $R_L^2 = 0.3$ for S-Q/S-B pair (using information from 2007 to 2008). Although the catchment is quite small (about $5-10\text{km}^2$) the poor R_L^2 can be explained by the site's irregular topography, which influences local rainfall convective cells (S-Y is located at a 200m lower elevation than S-B, while S-Q is located at a 300m higher elevation than S-B). Similar values were obtained using a nonlinear regression model embedding a power-2 model.

Three statistical models (GEV, Gamma and GP) were fit to the experimental data to obtain empirical P_E distributions using a maximum likelihood algorithm and the embedded fitting functions in the MATLAB environment. The results are reported in Figure 5b. Fitting performed to the full distribution of observations (including zeros) are shown by black lines with squares (GP) and dots (GEV). As expected, they performed relatively poorly when compared to the better fits for the POT distribution (rainfall events above a selected threshold). A sensitivity analysis indicated that all models performed reasonably well when setting a minimum threshold of $\theta = 2$mm/d. Note that the Gamma model could not be applied to the full distribution, being defined for non-zero positive values only.

The 679 rainfall values above the 2-mm threshold summed to a total rainfall of about 7560 mm. Of the remaining values, 439 observations were below the threshold, summing to only about 478 mm of cumulative rain, illustrating the dominant role of POT events (hereafter, a short name for rainfall values above the threshold) in the total water balance at this site. No specific statistical distribution and autocorrelation between pairs were found for values below threshold, which was assumed to be uncorrelated noise with uniform distribution $U = [0, \theta]$.

For the POT fitting, GEV model (grey line) tends to underestimate the probability of POT events above $P_E \approx 0.9$, while Gamma (blue line) and GP (red line) models slightly underestimate values over $P_E \approx 0.95$. GP slightly overestimates values above $P_E = 0.999$. The GP model provides a satisfactory fit to the rainfall distributions up to a threshold of $P_E \approx 0.99$ for the three rainfall stations (see Supplementary Electronic Material). However, the reliability of values above $P_E = 0.99$ in the experimental database is considered poor, as the number of samples was too few to provide robust estimates for very rare events. We were not interested here to describe the difference stemming

from the selection of different parametric models, and thus assumed that the GP model can satisfactorily describe the general distribution of POT events at the three sample locations. This was also done to compare the results from the MC analysis with those from the analysis of the experimental database.

For the three experimental stations, best-fitting analysis of the POT distributions (imposing $\theta = 2$) resulted in the following best-fit GP statistics:

- S-B: k=0.449, σ=2.356;
- S-Y: k=-0.082; σ=5.867;
- S-Q: k=0.100; σ=1.160.

It is noted that these statistics are quite similar to those used in the example of Figure 3. However, the maximum rainfall events for the theoretical S1 in Figure 3 are much larger ($r \approx 140$ mm/d) than those found that the experimental site at S-B ($r \approx 35$ mm/d). This occurs despite S1 and S-B showing quite similar k and σ values. This difference is associated with the effects of subsampling, since the population sizes for both experimental and synthetic distributions are not large enough to cover the full variability of possible events associated with these statistics. This may suggest that when CP and FP methods are applied to individual realizations, the efficiency of the tested methods could be different than what is inferred from the analysis of ensemble-averaged results from MC simulations.

We focused on reconstructing the values exceeding $\theta = 2$ mm for the S-B station observed data for the period 2007-2011 using S-Y as secondary information. The results from the direct application of FP and CP-L to the observed database are first analyzed from the visual inspection of the scatter plot shown in Figure 5c. It can be observed that the distribution of CP-L estimated values tends to follow the best-fit line (with equation $r_{S-B} = 0.61 r_{S-Y} + 5.9$), rather than the 1 : 1 distribution line. This issue gives rise to marked scattering

between estimated and observed values. FP is independent of the regression model, and seems to be more able to reconstruct values between $r = \theta$ and $r \approx$5-6 mm/d, being closer to the 1:1 line, although with a wide scattering in the estimation.

A more quantitative estimate of the errors between observed and reconstructed data was calculated in the form of SE. Using a cross-correlation approach, we removed a RTW of different size from existing observations and estimated the departure of estimated values from observation. The procedure was then repeated by changing the initial location of the RTW, for a number of times equals to the number of available time slots. The errors from each RTW were averaged over the number of repetitions, to provide an average cumulative error similar to (10). Figure 5d displays the calculated SE against N_R. The results clearly show that FP outperforms CP-L, especially for larger gaps. This is consistent with our previous MC-based conclusions: since the POT statistics change between experimental stations S-B and S-Y, correlation between pairs becomes nonlinear and FP approaches are more suitable for the reconstruction of missing points. This occurs even if the GP model poorly estimated events above $P_E > 0.95$. This observation may suggest that FP still performs better than CP methods even if distribution tails are not correctly fitted above very high thresholds, as usually occurs in practice for power-law models [7].

We did not report the results from CP-P, as this method performed remarkably worse than FP and CP-L. Indeed, we found RMSE$> 10^4$ when $N_R > 50$, and very large SE values from a limited number of missing values. The poor performance of CP-P can be ascribed to the wrong selection of the underlying nonlinear regression models, reflecting the difficulty in the adequate selection of an *a priori* nonlinear regression model for CP approaches. It is possible that for this specific combination of POT statistics the shape of nonlinear corre-

lation is different than a power-2 model. We did not test if CP-P provides a better estimate than FP using an alternative station (based on another combination POT statistics and thus another type of nonlinear correlation); the information from the other available secondary station (S-Q) was too limited and incomplete to exhaustively corroborate this hypothesis.

6.2 Real-World Scenario 2

Two other gauge stations nearby Vancouver (BC) were used to further validate the MC-based analysis and assess the actual efficiency of the CP and FP methods against a much longer real-world data set. The Vancouver data included complete daily rainfall values covering a period of 20 years (1995-2015) and can be freely accessed through the Environment Canada webpage (http://climate.weather.gc.ca/). The first analyzed station (Vancouver Harbour, VH) is located nearby the sea side, while the second station is located a few km away from VH and in the proximity of Vancouver Grouse mountain (VG), at an elevation of approximately 1500m.

Likely due to the distinct orographical configuration, the statistics characterizing the data pairs are significantly different. The correlation between the two datasets is $\rho = 0.55$ (which is similar to the Peruvian data set, where the elevation difference is smaller). At VG, the number of records exceeding a threshold of θ=1mm is 2136, which comprised about 30% of the total data series length. The total rainfall volume of associated with POT events is about 71% of total precipitation during the analyzed time. For VH, the POT events were 21% of the total rainfall events, summing up to 55% of the cumulative rainfall collected over the analyzed period.

The results from POT fitting are graphically reported in the Supplementary Online Material. We observed that, for VH, the GP was satisfactorily fitting

the POT values up to $P_E = 0.999$, after which the model slightly overestimated the observed values. For VG, the model was accurate up to $P_E = 0.99$, after which it overestimated the observations. Consistent with the Peruvian-based scenario, the Gamma and GEV models provided less accurate fits than the GP model. Based on these observations and to directly compare this scenario with the previous one, we used the GP as a working model to fit POT values and reconstruct the missing values using the FP approach. The resulting GP best-fit parameters for VH were k=0.07 and σ=10.17, while for VG we obtained k=0.01 and $\sigma = 20.25$.

We evaluated the efficiency of each infilling methodology through a stochastic cross-validation approach similar to that used in the Peruvian scenario. In particular, we used the VG dataset to reconstruct the POT rainfall events at VH. The resulting SE for a number of missing values up to N_R=1000 is shown in Figure 6. The plot shows that FP clearly outperforms the CP-L, consistent with the analysis of the Peruvian site and in line with the stochastic inference from the MC analysis. Indeed, the strong difference in POT statistics determines a clear nonlinearity between data pairs, confirming thus that FP is more suitable than the CP-L for infilling missing data. Once again, the results from the CP-P methods are not reported, strongly overestimating the errors compared to the other tested methodologies, possibly because the nonlinear structure between data pairs significantly departs from a power-2-like behavior.

Conclusions

Missing rainfall values are routinely estimated using simple approaches such as two-station chronological pairing (CP) and frequency pairing (FP). We developed a statistical framework to quantitatively evaluate expected errors

and reliability of predictions from the application of simple versions of these methods in the presence of nonlinear correlation between rainfall events and with different number of missing points.

For CP approaches, we analyzed one formulation with linear regression model (CP-L) and one with a nonlinear power-2 regression model (CP-P), which seemed to resemble the shape of nonlinear correlation between data pairs when scattering is low and POT statistics different. Synthetic realizations were generated with the aid of a copula-based algorithm, in order to specify a correlation between variables characterized by non-Gaussian (Generalized Pareto, GP) marginal distributions. In total, we analyzed four synthetic scenarios and two real world scenarios, each characterized by different GP statistics and different number of missing points (N_R).

The Monte Carlo results showed that when the POT statistics were identical, CP-L and CP-P outperform the FP. This behavior is consistent with the fact that similar statistics give rise to linear correlations between variables. This is especially true for the estimation of a reduced number of values ($N_R < 10^2$, for the specific selection of values used in this work). Despite being based on a nonlinear regression function, CP-P did not provide excessive additional errors compared with CP-L; this may occur because a power-law model can be approximated by a linear function in presence of low expected variability in data fluctuation.

When the POT statistics were different by pairs and variables were nonlinearly correlated, FP more markedly outperformed the other methods. This behavior is justified since the use of CP-L approaches is intrinsically not appropriated when variables are nonlinearly correlated, while CP-P performs well only when data resemble the power-law structure. Due to the high scattering, however, this occurs only for specific selections of POT parameters (e.g. increasing σ in our tested scenarios). Our exercise suggests however that iden-

tifying the optimal parameters *a priori* may be cumbersome and uncertain. Indeed, the application of the three approaches to real-case scenarios seems to confirm that the FP approach may provide more robust results than CP methods, FP being independent of the selection of a regression model.

While we only addressed a limited number of scenarios, we generally conclude that the optimal selection of the simple pairing approach must follow an initial statistical analysis of existing data, which may reveal nonlinearity between variables which can be hidden by poor ρ and high scattering. Only when nonlinearity is properly analyzed, one may obtain reliable quantitative information about expected errors in missing databases conditioned to site-specific empirical POT distributions.

Acknowledgements We acknowledge the useful suggestions provided by the Associate Editor and three anonymous reviewers, who significantly helped to improve the manuscript. All the data and additional information used and cited in this paper can be provided by the corresponding author at specific requests.

Fig. 1 Conceptual examples of chronologically-pairing (CP) and frequency-based pairing (FP) approaches. In CP (1), a linear regression model is applied to fit matching values in time series 1 (r_1) and time series 2 (r_2); m and q are two fitting parameters. In FP (2), each time series is considered as an independent random space; the same probability matches two different data values in the two time series. A missing value in r_2 is calculated using a measurement in r_1 and on the basis of the probability of that specific event in r_2.

Fig. 2 Theoretical behaviour of the GP model. (a) Analytical CDF for six different pairs of parameters. (b) Rate of convergence of expected errors versus sampling population. (c) Rate of convergence of expected errors for different shape parameters and scale parameters. We set $\theta = 0$ in all cases.

Fig. 3 Illustrative individual realizations of synthetic correlated rainfall distributions, characterized by different statistics and correlation coefficients (ρ) and $\theta = 2$. (a) Scatter diagrams for the pairs S1-S2. (b) Scatter diagrams for the pairs S1-S3. (c) Empirical POT rainfall distributions.

Fig. 4 Results of the Monte Carlo analysis, expressing the distribution over time of the mean root mean square errors (RMSE, figure a,b,c,d), and sum of cumulative errors (SE, figures e,f,g,h,) for different pairs of statistics.

Fig. 5 (a) Rainfall timeseries at two stations in the experimental site in Peru (S-B and S-Y) between 2007 and 2013; (b) empirical rainfall distribution of all recorded values, including zeros and values below the threshold (FD) and empirical POT distribution at S-B, along with three different fitting models: Gamma, Generalized Pareto (GP) and Generalize Extreme Values (GEV); (c) scatter plots for the CP-L and FP infilling approaches; (d) cumulative errors (SE) for the two infilling approaches.

Fig. 6 Cumulative errors (SE) for the two infilling approaches for the Vancouver-based scenario.

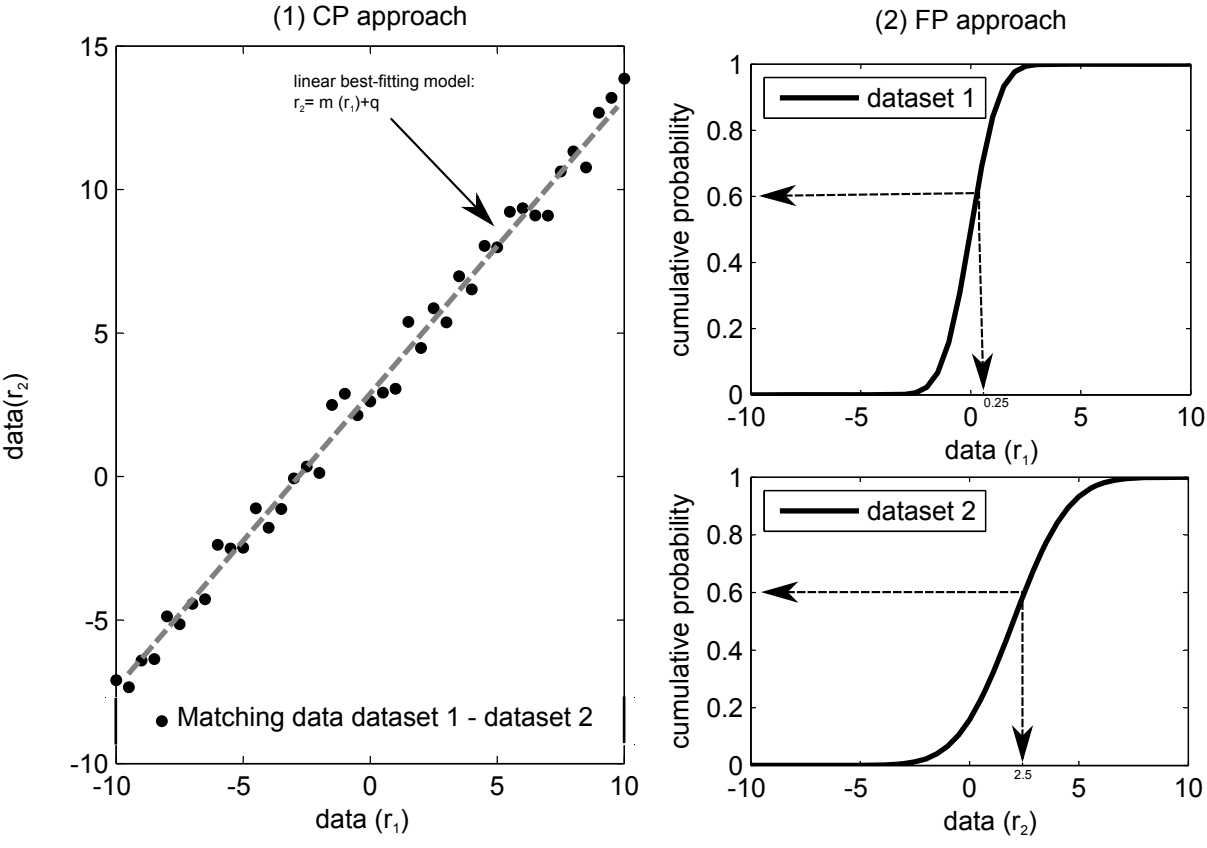
References

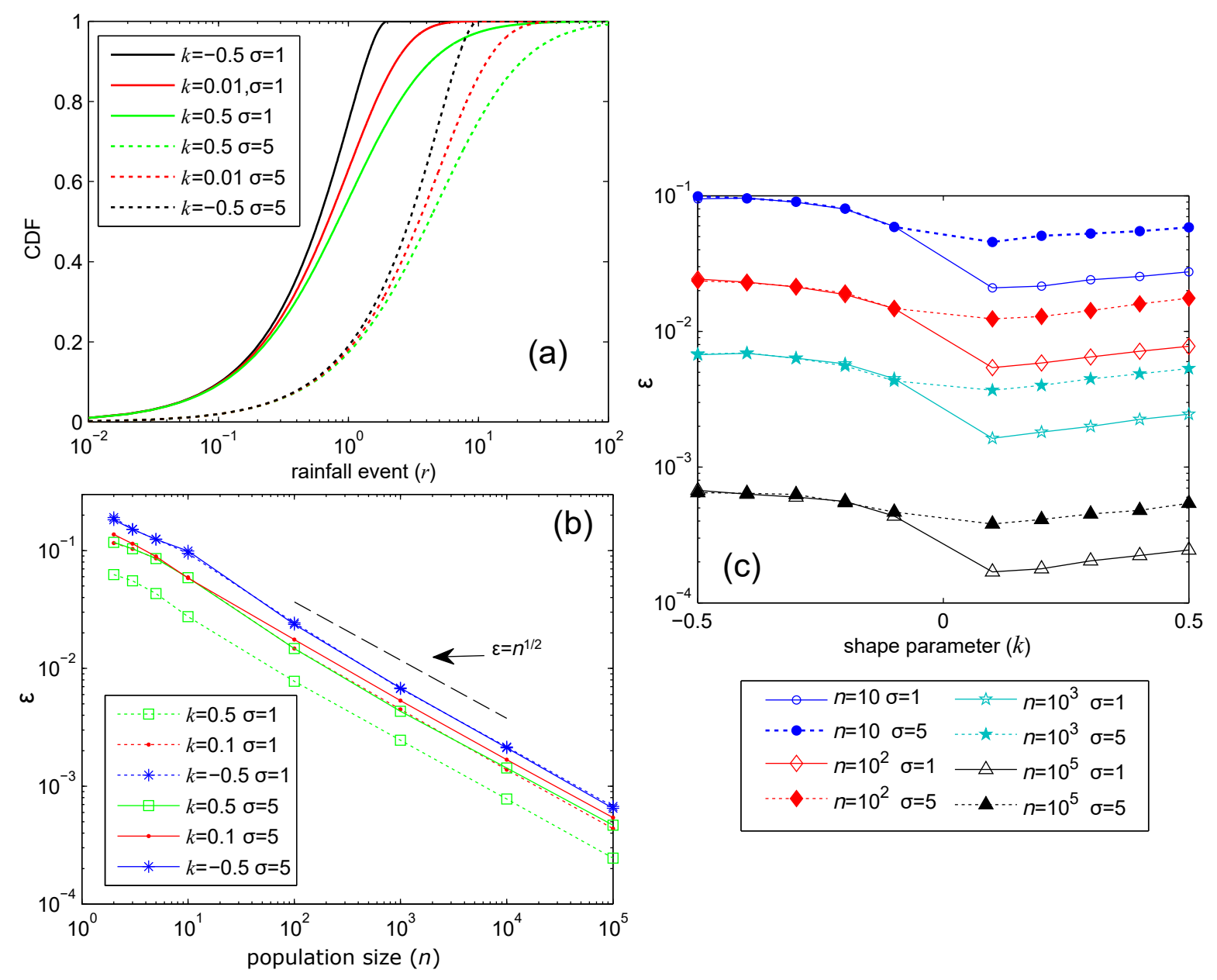
1. Arnell, N.W.: Unbiased estimation of flood risk with the GEV distribution. Stochastic Hydrology and Hydraulics **2**(3), 201–212 (1988). DOI 10.1007/BF01550842
2. Arnold, B.C.: Pareto Distribution. In: Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc. (2004)
3. Beauchamp, J., Downing, D., Railsback, S.: Comparison of Regression and Time-Series Methods for Synthesizing Missing Streamflow Records. Journal of the American Water Resources Association **25**(5), 961–975 (1989). DOI 10.1111/j.1752-1688.1989.tb05410.x
4. Brdossy, A., Pegram, G.: Infilling missing precipitation records A comparison of a new copula-based method with other techniques. Journal of Hydrology **519, Part A**, 1162–1170 (2014). DOI 10.1016/j.jhydrol.2014.08.025
5. Cantet, P., Arnaud, P.: Extreme rainfall analysis by a stochastic model: impact of the copula choice on the sub-daily rainfall generation. Stochastic Environmental Research and Risk Assessment **28**(6), 1479–1492 (2014). DOI 10.1007/s00477-014-0852-0
6. Castillo, E., Hadi, A.S.: Fitting the Generalized Pareto Distribution to Data. Journal of the American Statistical Association **92**(440), 1609–1620 (1997). DOI 10.1080/01621459.1997.10473683
7. Clauset, A., Shalizi, C.R., Newman, M.E.: Power-law distribution in empirical data. SIAM Review **51**, 661–703 (2007). DOI 10.1137/070710111
8. Coulibaly, P., Evora, N.D.: Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology **341**(12), 27–41 (2007). DOI 10.1016/j.jhydrol.2007.04.020
9. Demarta, S., McNeil, A.J.: The t Copula and Related Copulas. International Statistical Review **73**(1), 111–129 (2005). DOI 10.1111/j.1751-5823.2005.tb00254.x
10. Goovaerts, P.: Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology **228**(12), 113–129 (2000). DOI 10.1016/S0022-1694(00)00144-X
11. Grygier, J.C., Stedinger, J.R., Yin, H.B.: A generalized maintenance of variance extension procedure for extending correlated series. Water Resources Research **25**(3), 345–349 (1989). DOI 10.1029/WR025i003p00345
12. Hirsch, R.M.: An evaluation of some record reconstruction techniques. Water Resources Research **15**(6), 1781–1790 (1979). DOI 10.1029/WR015i006p01781
13. Hirsch, R.M.: A comparison of four streamflow record extension techniques. Water Resources Research **18**(4), 1081–1088 (1982). DOI 10.1029/WR018i004p01081

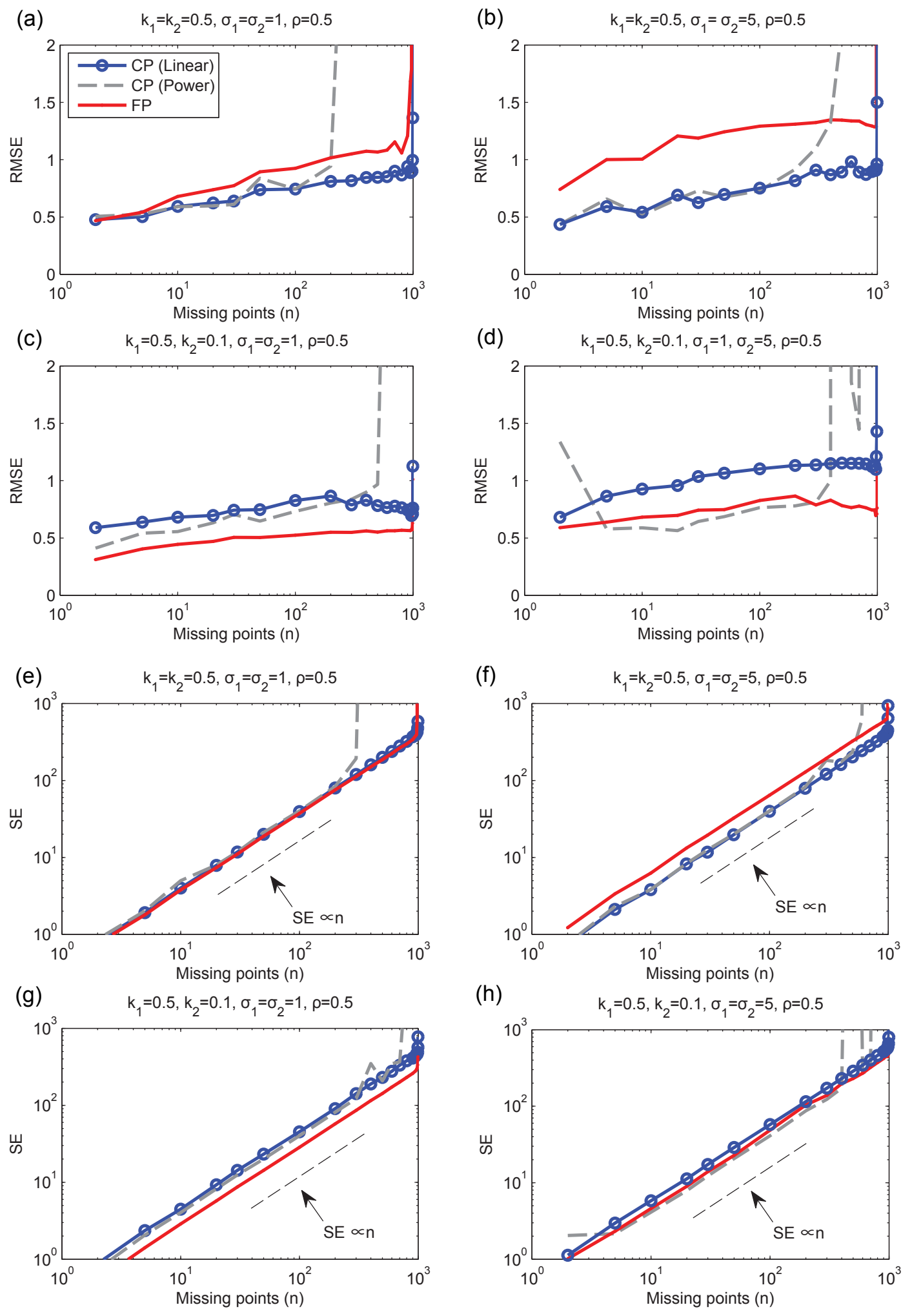
14. Kajornrit, J., Wong, K.W., Fung, C.C.: Estimation of Missing Precipitation Records Using Modular Artificial Neural Networks. In: T. Huang, Z. Zeng, C. Li, C.S. Leung (eds.) Neural Information Processing, no. 7666 in Lecture Notes in Computer Science, pp. 52–59. Springer Berlin Heidelberg (2012)
15. Kashani, M.H., Dinpashoh, Y.: Evaluation of efficiency of different estimation methods for missing climatological data. Stochastic Environmental Research and Risk Assessment **26**(1), 59–71 (2011). DOI 10.1007/s00477-011-0536-y
16. Kebaili Bargaoui, Z., Chebbi, A.: Comparison of two kriging interpolation methods applied to spatiotemporal rainfall. Journal of Hydrology **365**(12), 56–73 (2009). DOI 10.1016/j.jhydrol.2008.11.025
17. Khalil, B., Adamowski, J.: Record extension for short-gauged water quality parameters using a newly proposed robust version of the line of organic correlation technique. Hydrol. Earth Syst. Sci. **16**, 2253–2266 (2012). DOI 10.5194/hessd-9-4667-2012
18. Khalil, B., Ouarda T.B.M.J., and St-Hilaire, A. (2012). Comparison of record-extension techniques for water quality variables, Water Resources Management. Water Resources Management **26**(14),4259–4280 (2012). DOI 10.5194/hessd-9-4667-2012
19. Kim, D., Olivera, F., Cho, H.: Effect of the inter-annual variability of rainfall statistics on stochastically generated rainfall time series: part 1. Impact on peak and extreme rainfall values. Stochastic Environmental Research and Risk Assessment **27**(7), 1601–1610 (2013). DOI 10.1007/s00477-013-0696-z
20. Kim, J.W., Pachepsky, Y.A.: Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. Journal of Hydrology **394**(34), 305–314 (2010). DOI 10.1016/j.jhydrol.2010.09.005
21. Kim, T.W., Ahn, H.: Spatial rainfall model using a pattern classifier for estimating missing daily rainfall data. Stochastic Environmental Research and Risk Assessment **23**(3), 367–376 (2008). DOI 10.1007/s00477-008-0223-9
22. Li, Z., Li, C., Xu, Z., Zhou, X.: Frequency analysis of precipitation extremes in Heihe River basin based on generalized Pareto distribution. Stochastic Environmental Research and Risk Assessment **28**(7), 1709–1721 (2013). DOI 10.1007/s00477-013-0828-5
23. Lye, L.M.: Bayes estimate of the probability of exceedance of annual floods. Stochastic Hydrology and Hydraulics **4**(1), 55–64 (1990). DOI 10.1007/BF01547732
24. Millar, R.: A statistical approach for deriving project design rainfall. pp. 273–276. The Australasian Institute of Mining and Metallurgy, Melbourne, Australia (2013)
25. Musial, J.P., Verstraete, M.M., Gobron, N.: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series. Atmos. Chem. Phys.

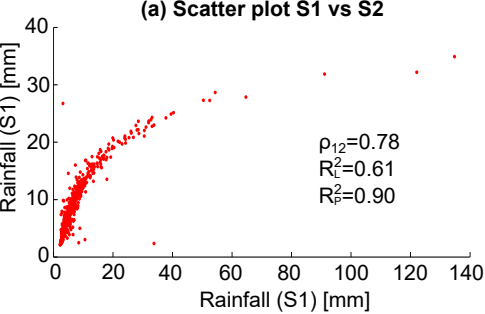
Discuss. **11**(5), 14,259–14,308 (2011). DOI 10.5194/acpd-11-14259-2011.

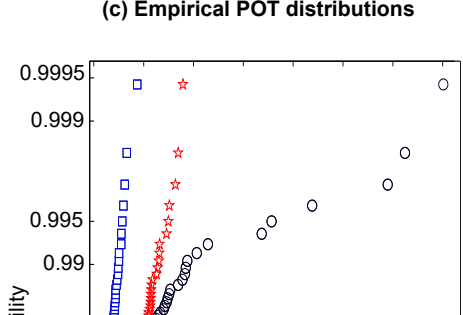
26. Papalexiou, S.M., Koutsoyiannis, D.: Battle of extreme value distributions: A global survey on extreme daily rainfall. Water Resources Research **49**(1), 187–201 (2013). DOI 10.1029/2012WR012557.
27. Paulhus, J.L.H., Kohler, M.A.: {Interpolation of missing precipitation records}. Monthly Weather Review **80**(8), 129–133 (1952)
28. Pickands III, J.: Statistical Inference Using Extreme Order Statistics. The Annals of Statistics **3**(1), 119–131 (1975)
29. Porporato, A., Ridolfi, L.: Influence of weak trends on exceedance probability. Stochastic Hydrology and Hydraulics **12**(1), 1–14 (1998). DOI 10.1007/s004770050006
30. Rubin, D.B.: Inference and Missing Data. Biometrika **63**(3), 581–592 (1976). DOI 10.2307/2335739
31. Serinaldi, F., Kilsby, C.G.: Rainfall extremes: Toward reconciliation after the battle of distributions. Water Resources Research pp. n/a–n/a (2014). DOI 10.1002/2013WR014211
32. Vogel, R.M., Stedinger, J.R.: Minimum variance streamflow record augmentation procedures. Water Resources Research **21**(5), 715–723 (1985). DOI 10.1029/WR021i005p00715

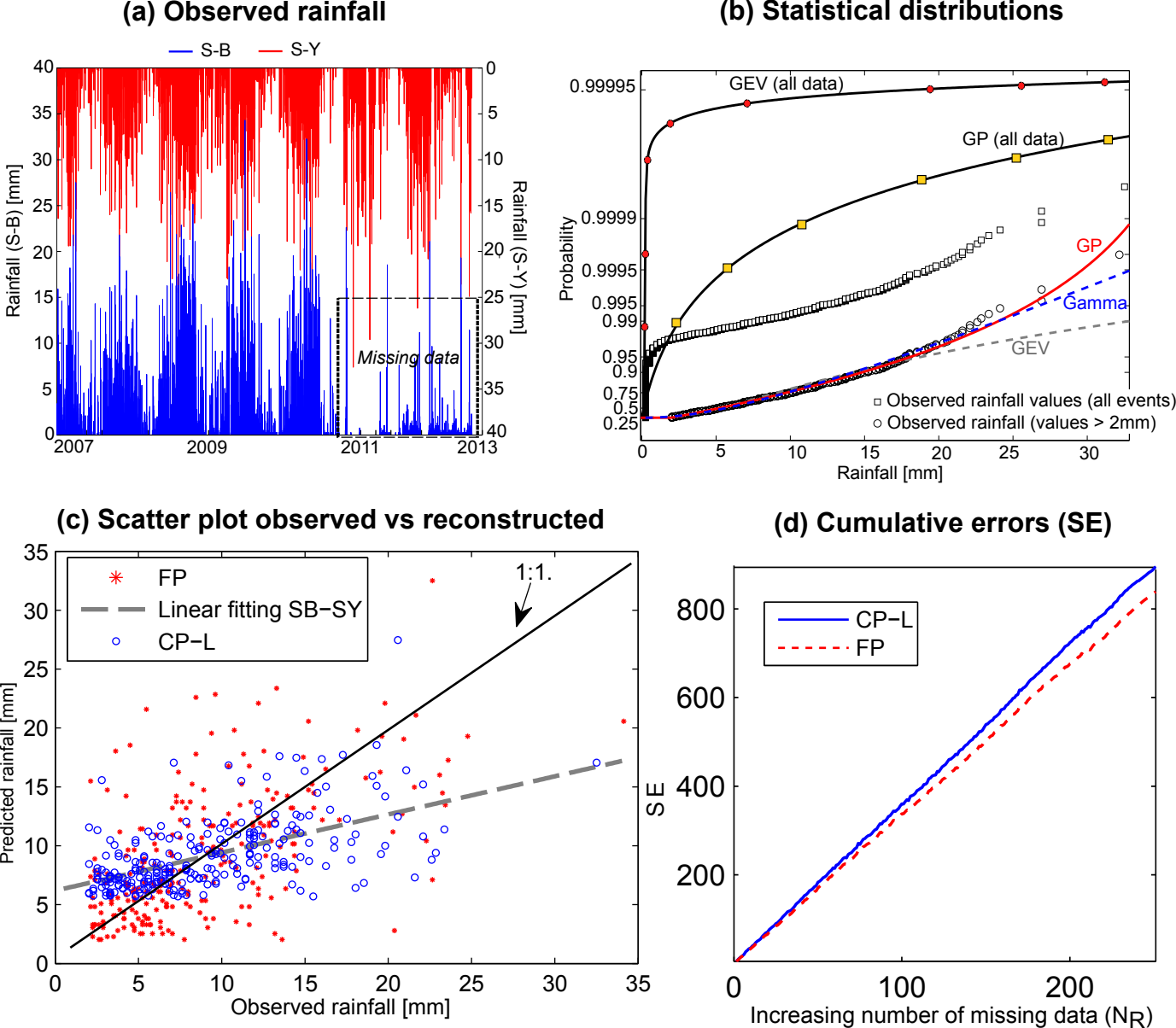

(1) CP approach
linear best-fitting model:
$r_2 = m\,(r_1)+q$
Matching data dataset 1 - dataset 2
data(r_2)
data (r_1)
(2) FP approach
dataset 1
cumulative probability
0.25
data (r_1)
dataset 2
cumulative probability
2.5
data (r_2)

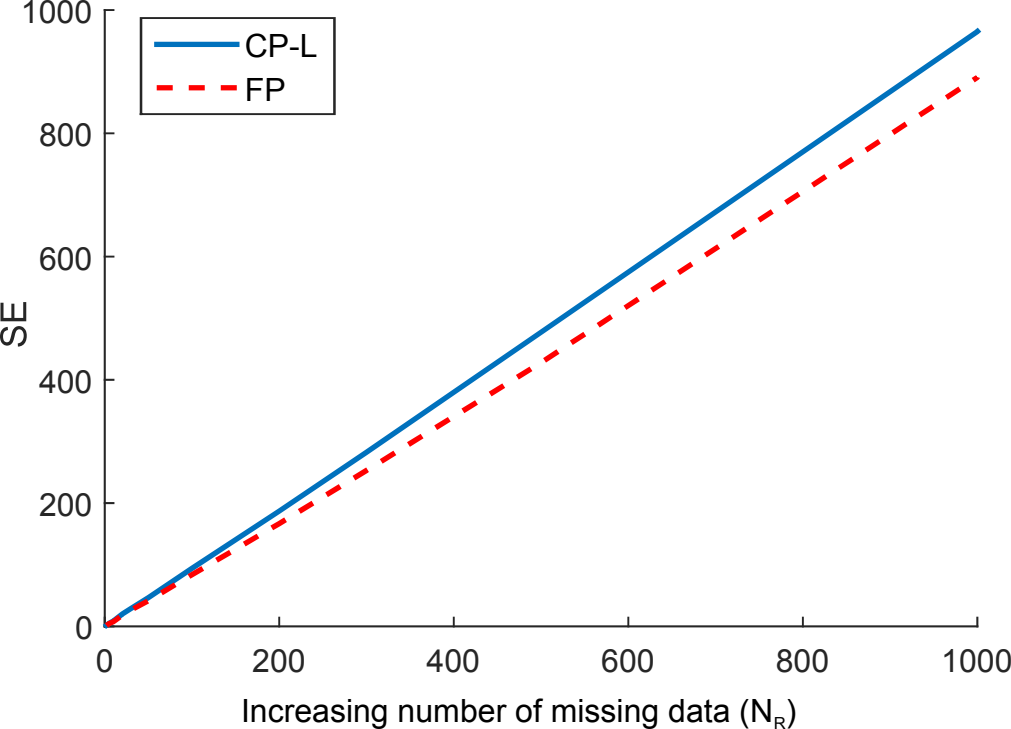

(a)
CDF
rainfall event (r)
k=−0.5 σ=1
k=0.01,σ=1
k=0.5 σ=1
k=0.5 σ=5
k=0.01 σ=5
k=−0.5 σ=5
(b)
ε
population size (n)
ε=$n^{1/2}$
k=0.5 σ=1
k=0.1 σ=1
k=−0.5 σ=1
k=0.5 σ=5
k=0.1 σ=5
k=−0.5 σ=5
(c)
ε
shape parameter (k)
n=10 σ=1
n=10 σ=5
n=10^2 σ=1
n=10^2 σ=5
n=10^3 σ=1
n=10^3 σ=5
n=10^5 σ=1
n=10^5 σ=5


(a)
k1=k2=0.5, σ1=σ2=1, ρ=0.5
CP (Linear)
CP (Power)
FP
RMSE
Missing points (n)
(b)
k1=k2=0.5, σ1= σ2=5, ρ=0.5
RMSE
Missing points (n)
(c)
k1=0.5, k2=0.1, σ1=σ2=1, ρ=0.5
RMSE
Missing points (n)
(d)
k1=0.5, k2=0.1, σ1=1, σ2=5, ρ=0.5
RMSE
Missing points (n)
(e)
k1=k2=0.5, σ1=σ2=1, ρ=0.5
SE
SE ∝n
Missing points (n)
(f)
k1=k2=0.5, σ1=σ2=5, ρ=0.5
SE
SE ∝n
Missing points (n)
(g)
k1=0.5, k2=0.1, σ1=σ2=1, ρ=0.5
SE
SE ∝n
Missing points (n)
(h)
k1=0.5, k2=0.1, σ1=σ2=5, ρ=0.5
SE
SE ∝n
Missing points (n)


(a) Scatter plot S1 vs S2
Rainfall (S1) [mm]
Rainfall (S1) [mm]
ρ12=0.78
R2L=0.61
R2P=0.90
(b) Scatter plot S1 vs S3
Rainfall (S3) [mm]
Rainfall (S1) [mm]
ρ12=0.5
R2L=0.32
R2P=0.29


(c) Empirical POT distributions
Probability
Rainfall [mm]
S1
S2
S3


(a) Observed rainfall
S-B
S-Y
Rainfall (S-B) [mm]
Rainfall (S-Y) [mm]
Missing data
2007
2009
2011
2013
(b) Statistical distributions
GEV (all data)
GP (all data)
GP
Gamma
GEV
Probability
Observed rainfall values (all events)
Observed rainfall (values > 2mm)
Rainfall [mm]
(c) Scatter plot observed vs reconstructed
FP
Linear fitting SB−SY
CP−L
1:1.
Predicted rainfall [mm]
Observed rainfall [mm]
(d) Cumulative errors (SE)
CP−L
FP
SE
Increasing number of missing data (N_R)


CP-L
FP
SE
0
200
400
600
800
1000
Increasing number of missing data (N_R)