Environmental Modelling & Software 86 (2016) 264–276

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Missing data simulation inside flow rate time-series using multiple-point statistics

Fabio Oriani [a, e, *], Andrea Borghi [b, c], Julien Straubhaar [a], Grégoire Mariethoz [d], Philippe Renard [a]

[a] Centre for Hydrogeology and Geothermics, Université de Neuchâtel, Neuchâtel, Switzerland
[b] École Nationale Supérieure de Géologie, Vandoeuvre-lès-Nancy, France
[c] Swiss Federal Office of Topography, Wabern, Switzerland
[d] Institute of Earth Surface Dynamics, Université de Lausanne, Lausanne, Switzerland
[e] Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 8 December 2015
Received in revised form
19 August 2016
Accepted 6 October 2016
Available online 21 October 2016

Keywords:
Time-series
Flow rate
Missing data
Non-parametric
Resampling
ARMAX
Multiple-point statistics

ABSTRACT

The direct sampling (DS) multiple-point statistical technique is proposed as a non-parametric missing data simulator for hydrological flow rate time-series. The algorithm makes use of the patterns contained inside a training data set to reproduce the complexity of the missing data. The proposed setup is tested in the reconstruction of a flow rate time-series while considering several missing data scenarios, as well as a comparative test against a time-series model of type ARMAX. The results show that DS generates more realistic simulations than ARMAX, better recovering the statistical content of the missing data. The predictive power of both techniques is much increased when a correlated flow rate time-series is used, but DS can also use incomplete auxiliary time-series, with a comparable prediction power. This makes the technique a handy simulation tool for practitioners dealing with incomplete data sets.

Software availability

The following information is about the software implementation of the simulation technique used in this paper:

Algorithm name: Direct Sampling (Mariethoz et al. (2010)).

Implementation name: DeeSse (Straubhaar (2015)).

Program language: C.

Developer: University of Neuchâtel, Julien Straubhaar (julien.straubhaar@unine.ch).

Year first available: 2015.

Minimal requirements: Windows/UNIX OS.

Availability: free license on request for research purposes, available on purchase for commercial use – for any request please contact Philippe Renard (philippe.renard@unine.ch).

A tutorial of the application shown in the paper is available upon request.

* Corresponding author. Department of Hydrology, Geological Survey of Denmark and Greenland, Øster Voldgade 10, DK-1350 Copenhagen K, Denmark.
E-mail address: fabio.oriani@protonmail.com (F. Oriani).

http://dx.doi.org/10.1016/j.envsoft.2016.10.002
1364-8152/

1. Introduction

The reconstruction of missing data portions inside time-series is a critical topic in applied hydrology since a large number of the numerical simulation techniques, used to model the hydrological processes, need continuous data records as input. Sometimes, technical failures of measurement instruments produce missing or unreliable data for long time periods for which the uncertainty about the observed phenomena is high. For this reason, a technique capable of generating realistic simulations of the missing data, reflecting the complex structures of the signal, and possibly making use of auxiliary information, is needed.

Many different approaches have been proposed for time-series gap filling in earth sciences: techniques based on mean diurnal variation or regression (Falge et al., 2001; Moffat et al., 2007), autoregression (Bennis et al., 1997; Wang, 2008), singular spectrum analysis (Schoellhamer, 2001; Kondrashov et al., 2014), self-organizing maps (Wang, 2003; Lamrini et al., 2011), look-up tables (Bamberger et al., 2014), rough sets (Dumedah et al., 2014), and artificial neural networks, widely used in recent years (Aminian

and Ameri, 2005; Dastorani et al., 2009; Diamantopoulou, 2010; Nkuna and Odiyo, 2011; Bahrami et al., 2011; Nourani et al., 2012; Dumedah et al., 2014). In this paper, we propose a non-parametric method to simulate missing data inside flow rate time-series based on the Direct Sampling (DS) technique (Mariethoz et al., 2010) belonging to multiple-point statistics (MPS). Already tested on gap filling in multivariate data sets representing natural heterogeneities (Mariethoz et al., 2012, 2015) and on rainfall time-series simulation (Oriani et al., 2014), DS can simulate the outcome of a complex natural process by reproducing similar patterns to the ones found in the available data without imposing a specific statistical model. More particularly, missing data are simulated by sampling the available data set where a sufficiently similar pattern is found. High-order statistical relations in the variable of interest are preserved by respecting the similarities in the neighborhood at multiple scales. The approach is almost entirely data-driven and fairly simple, but its efficiency largely depends on finding the good ensemble of auxiliary variables suitable to the current application. We present a multivariate standard setup for missing data simulation inside hydrological flow rate time-series using a correlated time-series as auxiliary variable. The setup is tested on the gap filling of a high-resolution karst flow rate time-series using different auxiliary variables. To make the test systematic and relevant for application, a gap size varying from a few hours to 20 days and total missing data percentage up to 30% are considered. Finally, a last group of tests focuses on the comparison of the proposed technique with a classical time-series model of type ARMAX. The general methodology, the setup, as well as the data set used, are illustrated in Section 2, the results are presented in Sections 3 And 4, while Section 5 is dedicated to the conclusions.

2. Methodology

2.1. The data set

The data set used to test the proposed technique is the 1990–2013 flow rate record from two karst springs of the Jura mountains (Switzerland) provided by the Swiss Federal Office for the Environment (FOEN). This paleozoic karst system is characterized by flashy spring discharges (Painter et al., 2008). Three high-resolution (10-min) time-series are used: the Areuse creek measured at St. Sulpice station (Ar) is used as a target variable, while the same water flow measured at Boudry station (Ar2) and the Seyon creek measured at Valangin station (Se) are used as auxiliary variables. The two river basins are contiguous (Fig. 1) and their regimes have been both classified as Jurassian pluvial and nivo-pluvial (FOEN). Ar station (443 m a.s.l.) lies at a distance of about 20 km from Ar2 (750 m a.s.l.) and 30 km from Se (628 m a.s.l.). Measuring from the same river, Ar and Ar2 are highly correlated (Pearson's correlation coefficient PCC = 0.96), whereas Ar and Se show a medium to weak correlation (PCC = 0.72). The considered time-series do not contain any missing data, but Ar and Ar2 show isolated sharp fluctuations around the local trend due to instrumental errors. To remove this kind of artifact, the following preprocessing treatment is applied (Oriani, 2015): given a time-series $Z(t)$ and computing the differential operator $\delta Z(t) = Z(t) - Z(t-1)$, the artifacts are identified with the portions of $Z(t)$ presenting $\sigma(t,a) > b$, where $\sigma(t,a)$ is the local standard deviation of $\delta Z(t)$, computed on the time interval $[t \pm a]$ and b is a user-defined threshold. The appropriate value for a and b depending on the smoothness of the signal and the magnitude of the artifacts, can be manually set by visually checking the results. In this paper, the chosen values are $a = 19$, $b = 0.3$ for $Z(t) = \mathrm{Ar}$ and $b = 0.05$ for $Z(t) = \mathrm{Ar2}$. The data detected as artifacts are replaced by a cubic spline interpolation.

2.2. The Direct Sampling technique

Multiple-point statistics (MPS) techniques are based on the concept of training data set (TI): a representative sample of the target variable or conceptual model which is used to estimate the probability of occurrence of each event inside the simulation. MPS methods (Guardiano and Srivastava, 1993; Strebelle, 2002; Allard et al., 2006) generally consider a catalog of neighboring data patterns found in the TI to impose high-order conditioning in the simulation and thus reproduce similar structures to the ones found in the TI. This requires the estimation of the conditional probability

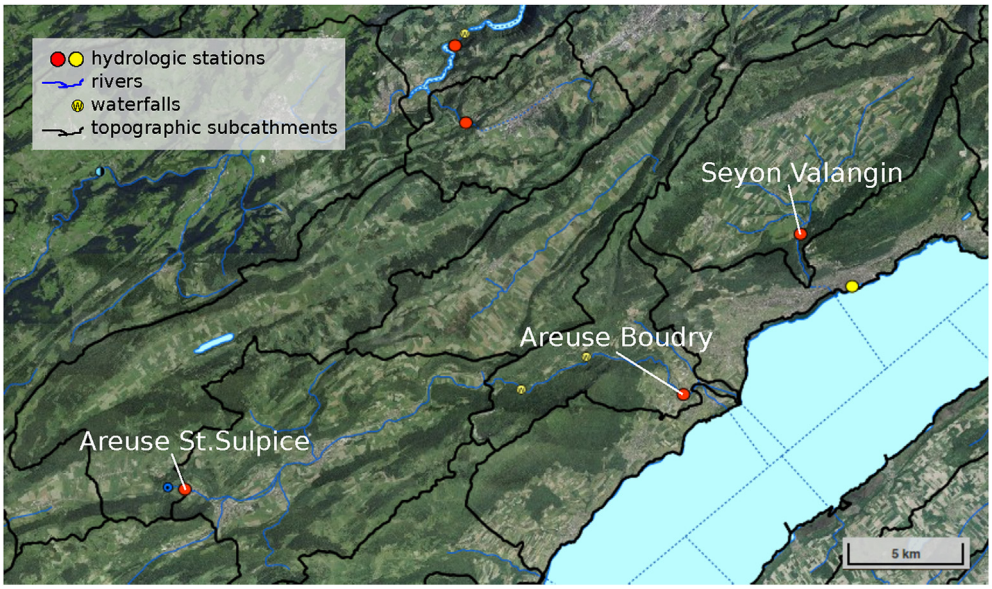


Fig. 1. Aerial photo of the study region (Jura mountains and Neuchtel lake), with location of the measure stations and topographic basin subdivision (modified from Swiss Federal Office of Topography, map. geo.admin.ch).

density function for each pattern and limits the application of the method to categorical variables. The Direct Sampling technique (Mariethoz et al., 2010) avoids this preliminary step by sampling the TI where a sufficiently similar neighborhood occurs in the TI. This principle also extends the application to continuous variables and multivariate data sets. In the case of missing data simulation, only the uninformed time steps are simulated and the rest of the data set is used as conditioning (CD). We refer to Straubhaar et al. (2011) and Oriani et al. (2014)for a detailed description of the algorithm implementation. We present here the main workflow of the algorithm, in the case of multivariate simulation, i.e. when the variable of interest (target variable) is simulated together with one or more given auxiliary time-series. These can be fully or partially informed. The required inputs for the simulation are a data set used as TI and the simulation grid (SG), a time vector hosting the CD together with the simulated data. Both the TI and SG are multivariate, containing the target and the auxiliary variables. In case of missing data simulation, the TI and the CD may be the same data set, meaning that the gaps are filled using the data already present in the SG at the beginning of the simulation. In this case, only the SG is filled with simulated data, not the TI, that will preserve the original data and gaps. The algorithm proceeds as follows:

1 A random permutation of the index vector is done to obtain a random simulation path inside the SG.
2 Each variable is linearly normalized to a range of [0,1].
3 Following the random simulation path, and uninformed time step t of the SG is chosen for simulation.
4 A pattern of neighboring data of t is retrieved independently for each variable and according to a search template defined by a radius R_k and a maximum number of considered time steps N_k for each k-th variable. For example, if $R_k = 20$ and $N_k = 10$, the pattern is composed by the 10 informed time steps closest to t inside the time span $[t \pm 20]$. This local data subset sampled for the SG is called data event ($d_{k,t}$), and constitutes the pattern on which the simulation of the k-th variable at t is conditioned. Since $d_{k,t}$ is composed by the closest available data, it does not require the variable to be completely informed. R_k and N_k are user-defined parameters.
5 A random time step y of the TI is randomly scanned to retrieve a data events $d_{k,y}$. The time-steps in $d_{k,y}$ have the same time lag as the ones in $d_{k,t}$.
6 A distance $D_k(d_{k,t},d_{k,y})$, i.e. a measure of dissimilarity, is computed between the two data events for each variable. D_k is the fraction of non-matching elements for categorical variables and the absolute mean error for continuous variables.
7 If D_k is below a prescribed threshold T_k for all k, the datum in y is assigned to t for all uninformed variables. Otherwise the procedure is repeated from step 5 to 7 until a suitable $d_{k,y}$ is found or a prescribed TI fraction F is scanned.
8. If no time step y presenting $D_k < T_k$ for all k is found, the one which minimizes $\sum_{k=1}^{K} D_k$, with K simulated variables, is assigned to t.
9. The procedure from step 3 to 8 is iterated until the SG is completely informed.
10. The variables are linearly back transformed to their original range.

To summarize, the main DS parameters, related to each simulated variable are: i) the search neighborhood radius R, that defines the time interval $t \pm R$, used to retrieve the conditioning pattern for the simulation at time step t; ii) the maximum number of neighbors N used to form the conditioning pattern; and iii) the distance threshold T, a scalar value used to accept or reject the pattern scanned inside the TI. These parameters can take different values for each variables in the multivariate case. One last parameter defined once for all variables is the maximum TI fraction (F) scanned at each algorithm iteration. The value $F = 0.5$ is adopted for all the tests presented in this paper. This is a standard value used in previous time-series applications (Oriani et al., 2014) that, in case of a representative training data set, generally allows scanning a sufficient training data amount at each iteration. The scanning of the total TI is avoided since it may lead to oversampling of the same TI regions and the reproduction of entire data set portions.

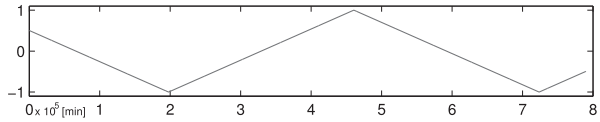
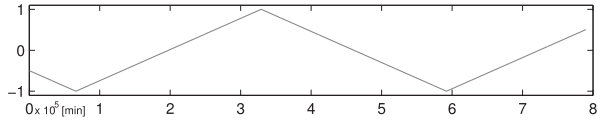
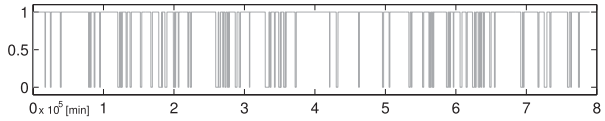
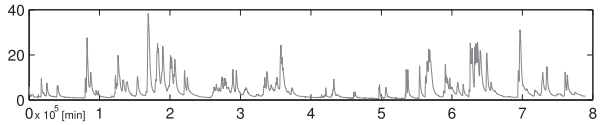
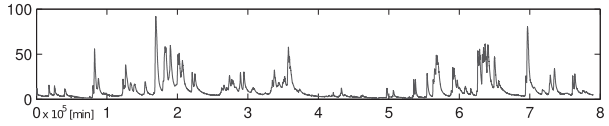
As explained in Oriani et al. (2014), the main difference of this approach with respect to the existing resampling techniques for time-series simulation, e.g. Rajagopalan and Lall (1999); Buishand and Brandsma (2001); Wojcik and Buishand (2003); Clark et al. (2004), is the combined use of i) a random simulation path and ii) a variable conditioning scheme, using the N informed neighbors closest to the simulated time step. These two elements allow considering large-scale patterns at the beginning of the simulation and denser small-scale patterns toward the end of the simulation. For instance, by setting $R = 100$ and $N = 10$, the conditioning pattern for the simulation of the first random time steps will be formed by 10 or less sparse neighbors in the time $t \pm R$, while, for the last simulated time steps, it will be composed by 10 time steps much closer to the simulated one, since at this stage, the SG is more densely informed. This imposes a variable time-dependence, which allows preserving the statistical structure at multiple scales without the formulation of a high-dimensional prior statistical model. Moreover, when multiple variables are simulated together, their statistical correlation is preserved in the multivariate data set. For example, one can use a sufficiently informed variable to guide the simulation of large missing data portions inside the target variable (see Section 2.3). It is worth remembering that since DS samples the data found in the TI, the use of a representative TI is crucial to obtain a reliable simulation (see Section 3).

2.3. The DS setup for flow rate time-series

In this section, we present a standard multivariate DS setup for the simulation of missing flow rate time-series data composed by a series of variables and the main DS parameter values (Table 1). A flow rate time-series ($Z(t)$) is simulated in its missing data parts together with a group of auxiliary variables. This allows the preservation of the temporal structure contained in the original data set. The river flow processes are characterized by an annual seasonality and spatio-temporal correlation. For this reason, the proposed auxiliary variables include two periodic functions describing the annual seasonality ($A_1(t)$ and $A_2(t)$), a correlated flow rate time-series ($Q(t)$) measured from a nearby location (if available) and an indicator variable describing the hydrographic structure ($H(t)$). This multivariate data set is defined at the same temporal resolution of $Z(t)$, that may vary according to the considered data set. A more detailed description of these variables and the respective DS parameterization are given in the following. We removed the temporal reference from the notation where possible.

- The flow rate time-series Z is the target variable, presenting missing data portions that are generated in the simulation and informed time-series portions that are used as conditioning data. A variable high-order conditioning is applied to Z by extending the search neighborhood radius (R) to 10,000 10-min time steps and considering a maximum (N) of 15 neighbors. The distance threshold value $T = 0.002$ includes 0.2% of total variation on the conditioning pattern. The chosen values for R and N for all variables can be related to the correlation length of their temporal structure. The user is recommended to change the

Table 1
DS setup proposed for flow rate time-series simulation. The parameters for each variable are: search window radius *R*, maximum number of conditioning neighbor data *N* and distance threshold *T*.

Variable	*R*	*N*	*T*	TI example
1) A_1	1	1	0.07	
2) A_2	1	1	0.07	
3) *H*	10,000	20	0.05	
4) *Q*	10,000	15	0.002	
5) *Z*	10′000	15	0.002	

value of these parameters according to the time-series resolution. Conversely, the proposed *T* values, lacking a physical meaning related to the variable, are manually set up by trying a limited set of values (0.002, 0.01, 0.05, 0.07) and using the indicators presented in Section 2.6 as optimization criteria. A sensitivity analysis of the DS parameters (Meerschman et al., 2013) showed that, for the majority of the application cases considered until now, the optimal *T* values lie within [0.001,0.1], with lower values suitable for highly autocorrelated, smooth signals and higher values for low-correlated, more noisy signals. In case of lower resolution or more noisy data sets, a higher *T* value may be more appropriate.

- Two out-of-phase periodic triangular functions (A_1 and A_2) with period 365.25 days, indicate the position of each datum inside the annual cycle. A_1 and A_2 are given as CD to help the simulation respecting the annual seasonality. Since high-order conditioning is not necessary for this purpose, *R* and *N* are set to 1. The distance threshold *T* set to 0.07 allows sampling from the same period of the year with a maximum 7% of the total variation of the variables. As already observed for rainfall and climate variables (Oriani et al., 2014), *T* varying between 0.05 and 0.07 allows imposing the annual seasonality without over-conditioning the simulation.
- A flow rate time-series (*Q*) from a nearby located station is given as CD but it is not necessarily fully informed. Any missing data inside *Q* will be co-simulated with the target variable. If *Q* is correlated to *Z*, its conditioning helps restricting the uncertainty around the missing data, e.g. indicating a flood occurrence if a peak is present in *Q*. The same DS parametrization as *Z* is applied to *Q*.
- An indicator variable called recession indicator (*H*), takes values $H = 1$ to indicate a recessing hydrograph limb and $H = 0$ for a rising hydrograph limb observed in *Q*. *H* is necessary to simulate a more realistic flood pattern in the target variable. Since the flow rate time-series is a complex signal, showing abrupt fluctuations of different magnitude, computing the sign of the derivative in *Q* is not sufficient to identify the effective succession of rising and recessing limbs, corresponding to the main flood pattern. For this reason, *H* is computed with a more complex procedure (Oriani, 2015) summarized in the following. *H* is a deterministic function of time *t* and the user defined parameters (w, v). First, the local extremes (minimum and maximum) of *Q* inside a moving temporal window $[t \pm w]$ are identified. Moreover, each extreme is considered only if: i) it shows a variation greater than v with respect to the previously considered extreme and ii) the next extreme found is not of the same type (minimum or maximum). Finally, *H* is obtained by applying a logical test on the selected local extremes: a local minimum activates a rising limb ($H = 0$) until a local maximum occurs activating a recessing limb ($H = 1$), ensuring a continuous alternation of the two categories. If *Q* is incomplete, *H* also presents missing data at the corresponding time steps. The values ($w = 50$, $v = 2$) for $Q =$ Ar2 and ($w = 50$, $v = 0.3$) for $Q =$ Se have been set up by trial and error to allow an adequate visual representation of the hydrographic structure. The user is recommended to set up these values by visually checking the result, so that the main alternation of rising and recessing limbs can be detected. This may vary significantly according to the regime type. The DS parameter values for *H* are $R = 10'000$ time steps, $N = 20$ and $T = 0.05$.

The proposed DS setup makes use of *Q* as source of additional information. The rest of the variables are in fact derived from it (*H*) or known a priori (A_1 and A_2). If *Q* is not available, the simulation is still possible with *H* computed on the informed part of *Z*. Since the most adequate parameter values for DS and *H* may vary as a function of the flow rate characteristics and the time-series sample rate, the user should not consider the suggested values as fixed but

rather as a starting point for optimization to a specific application.

2.4. Multiple scenario test

In the first test, artificial gaps are created in the multivariate data set and the corresponding missing data are simulated using the proposed DS setup. Ar is considered as target in 5 simulation tests presenting different time-series as Q: 1) In test Ar, no Q variable is used and H is computed on the informed part of Ar. 2) In test Ar-Ar2, Ar2 is used as complete Q variable, highly correlated to Ar. 3) An incomplete version of Ar2 (Ar2*) is used in test Ar-Ar2*. 4) In test Ar-Se, Se is used as Q to represent a case where the auxiliary time-series is poorly correlated with Ar. 5) Finally, in test Ar-Se*, $Q =$ Se*, representing a case where Q is incomplete and poorly correlated with Z. To test the sensitivity of the method performance to different gap sizes and missing data quantities, random groups of untouching and equally sized gaps are generated inside Z according to different missing data scenarios: as shown in Table 2, three different classes of missing fraction up to 30% and three different classes of gap size up to 3000 time steps per gap are considered for a total of 9 fraction-size combinations. Since the time step is 10 min, the generated gap sizes vary between 8 h and 20 days. This time range can represent the data loss due to small mechanical failures or large breakdowns comprising entire wet periods (see Fig. 2). For each fraction-size combination, 10 different missing data series (gap scenarios) and 10 DS realizations per scenario are generated for a total of 900 runs per test. For test Ar-Ar2* and Ar-Se*, gap scenarios for Q are generated independently from those for Z and present always a 20% missing fraction with 300-time-step gaps.

2.5. Comparative test

2.5.1. The ARMAX model

The last group of experiments focuses on the comparison of the proposed direct sampling setup with a time-series model of type autoregressive moving average with exogenous variable (ARMAX). Under the hypothesis of weak stationarity, the ARMAX model (Box and Jenkins, 1976) aims at preserving the temporal structure of the observed process by considering the linear dependence of the simulated time-step $Z(t)$ with a series of past values of both Z and an exogenous (auxiliary) variable Q, with the addition of a noise term that is also correlated with its past values. The result is the following regression model:

$$Z(t) = \varepsilon(t) + \sum_{i=1}^{I} \alpha_i Z(t-i) + \sum_{j=1}^{J} \beta_j Q(t-j+N) + \sum_{k=1}^{K} \gamma_k \varepsilon(t-k) \quad (1)$$

where $Z(t)$ is the target variable simulated at time t, I is the autoregression order for $Z(t)$, J the regression order for $Q(t)$, with a delay time N, and K is the autocorrelation order for the noise term, with $\varepsilon(t)$ being a white-noise. $\vec{\alpha}$, $\vec{\beta}$, and $\vec{\gamma}$ are the regression coefficient vectors. This ones are numerically calibrated on the training data set of the comparison test (2.5.2) with a prediction-error iterative method (PEM, Ljung, 1999), minimizing the quadratic prediction error on the given training (input, output) data set ($Q(t)$,$Z(t)$). To choose the model orders $I, J, N, K \in \mathbb{N}_{\geq 0}$, all the combinations up to the 5-th order have been considered: the values $I = 1$, $J = 2$, $N = 1$, and $K = 1$ have been chosen since they gave the best results in terms of visual comparison with the reference. Moreover, the error (RMSE) between the simulated and reference time-series and their distribution mismatch (RMSE on the quantiles) were among the lowest.

Table 2
Simulation schedule for each test: 9 missing fraction-gap size combinations, 10 gap scenarios per combination, 10 realizations per gap scenario, for a total of 900 realizations.

missing fraction → gap size (num. Time steps) ↓	5%	10%	30%
1) 50 (~8 h)	10 scenarios × 10 real.	10 × 10	10 × 10
2) 300 (~2 days)	10 × 10	10 × 10	10 × 10
3) 3000 (~20 days)	10 × 10	10 × 10	10 × 10

2.5.2. Comparative test design

To compare DS with ARMAX, two of the data sets previously presented in the paper (Section 2.4) are used, namely: test Ar-Ar2, where the auxiliary variable Ar2 is highly correlated to the target Ar, and Ar-Se, where the auxiliary variable Se is poorly correlated to Ar. Since the ARMAX technique, in its original version, is not adaptive to simulate partially informed time-series, the two techniques are compared on a simple test using the first 3 years of the time-series as training data set and the following 3 years as validation data set: the training data set is entirely informed, while the validation data set is entirely simulated. The auxiliary variables used (Ar2 and Se) are also entirely informed. Each simulation ensemble includes 10 realizations.

2.6. Evaluation

The performance of the proposed simulation technique is analyzed separately for each test and fraction-size combination. The visual comparison between the generated and reference time-series as well as a group of statistical indicators are considered. To test the efficiency in simulating the statistical content of the missing data, the probability distribution of the simulated and reference missing time-series portions are compared using quantile-quantile (qq-) plots. To show the behavior of the simulation ensemble, the median, 5th and 95th percentile of the realizations are plotted for each quantile. The predictive power of the technique is tested using classical goodness-of-fit measures: the Pearson's correlation coefficient (PCC) between the simulated and reference missing data, the root mean square error (RMSE) and the Nash-Sutcliffe model efficiency coefficient (NSE). In the comparison test (Section 2.5.2), the percentage bias (PBIAS) is also computed.

3. Multiple scenario test results

In the following, we analyze the results of the first test about the application of the proposed simulation technique to different missing data scenarios (Section 2.4).

3.1. Visual comparison

Fig. 2 shows a time-series portion of approximately 100 days presenting two simulated gaps and the corresponding missing data together with the auxiliary variables used. Among the considered gap scenarios, one presenting 30% missing fraction and 3000-time-step gaps has been chosen for visual comparison, since it better illustrates how the signal is reconstructed by the algorithm. In general, the algorithm generates hydrographic structures similar to the one found in the reference by sampling each datum from the TI where a similar neighborhood is found: the simulated flood events are in the same magnitude range and the asymmetric shape of the hydrograph looks realistic although it contains a modest noise. In test Ar (Fig. 2, A), where no Q variable is used and H is computed on the informed part of Z, the flood occurrence in the simulation does

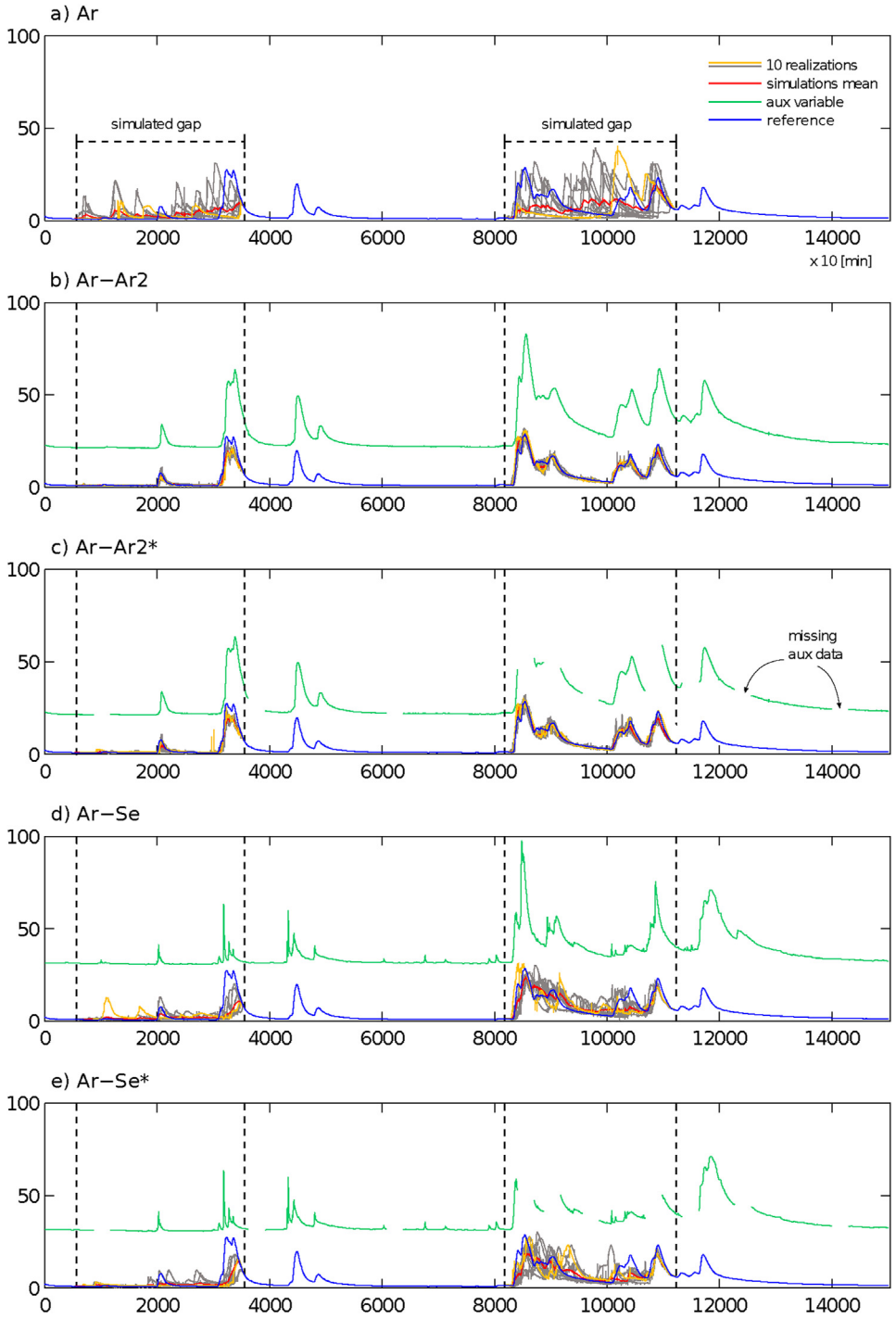


Fig. 2. Multiple scenario experiment: A flow rate time-series portion (m^3/s, 10-min average, approximately 100 days) showing two simulated gaps, the reference and the auxiliary variable used for all the tests. The shown examples belong to one scenario presenting 30% missing fraction and 3000-time-step long gaps. A randomly chosen realization (orange color) is put in evidence over the simulation ensemble (dark gray color). The auxiliary variable (green line) is shifted upward with respect to the vertical scale for illustration purpose. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

not match the reference flood structure but flow rate values are in the same range. This is expected since, within the missing data portion, the simulation is only conditioned by A_1 and A_2 informing about the annual seasonality. Therefore, the algorithm explores a larger variability, generating different types of flood structures. Conversely, when Q is present and highly correlated to Z as in test Ar-Ar2 (Fig. 2, B), the uncertainty is restricted around the reference flood structure. Local extremes are estimated quite accurately as shown by the simulation mean (red line). When Q is poorly correlated to Z (test Ar-Se, Fig. 2, D), the simulation mean follows the main reference shape, but the simulation ensemble shows larger uncertainty. This suggests that Q and its derived variable (H) are highly informative about the hydrograph structure and play an important role in conditioning the simulation. When this auxiliary information is incomplete as in test Ar-Ar2* (Fig. 2, C), the algorithm shows a similar performance: even if some portions of the

auxiliary variables (Q and H) are missing in correspondence to peaks, DS can efficiently simulate the local extremes, if the auxiliary variable is sufficiently correlated to the target. It is not the case for Ar-Se* (Fig. 2, E), where the time-series Se is incomplete and poorly correlated with Ar, so it cannot improve the simulated structure. Normally, this result cannot be achieved with a parametric technique based on fixed time dependence, since it requires the predictor variables to be fully informed. Conversely, DS uses a variable conditioning pattern adapted to the data available in a temporal range defined by the parameter R (see Section 2.3). The missing data are simulated in a random order which allows first defining the large-scale structure of the missing portion. Then, the data sequence is completed by considering the already simulated values as conditioning data. This leads to a realistic reconstruction even if the auxiliary time-series used is incomplete.

3.2. Statistical content

In Fig. 3, the probability distributions of the missing reference and simulated data are compared by means of qq-plots. This allows testing the efficiency of the technique in recovering the statistical content lost with the missing data. The results may vary significantly depending on the missing time-series portions. For this reason, 10 scenarios for each fraction-gap combination and test have been considered (see Section 2.4). For all tests, the median of the simulations (solid lines) mainly lies on the bisector of the graph, indicating that the simulation preserves the reference distribution on average. Nevertheless, a tendency to underrepresent data between 40 and 60 m^3/s is observed when the simulated data quantity is limited (5–10% of the data set, Fig. 3 left and center column). This suggests that the algorithm is not an appropriate tool to represent the extremal behavior of the target variable at the temporal scale of the data since it may underrepresent the extreme values in this case. This happens because, direct sampling it is not capable of generating values not observed in the training data set. Nevertheless, the underrepresented data correspond to very rare events: for example, in this case, they constitute only 0.0014% of the observations and occur sparsely in

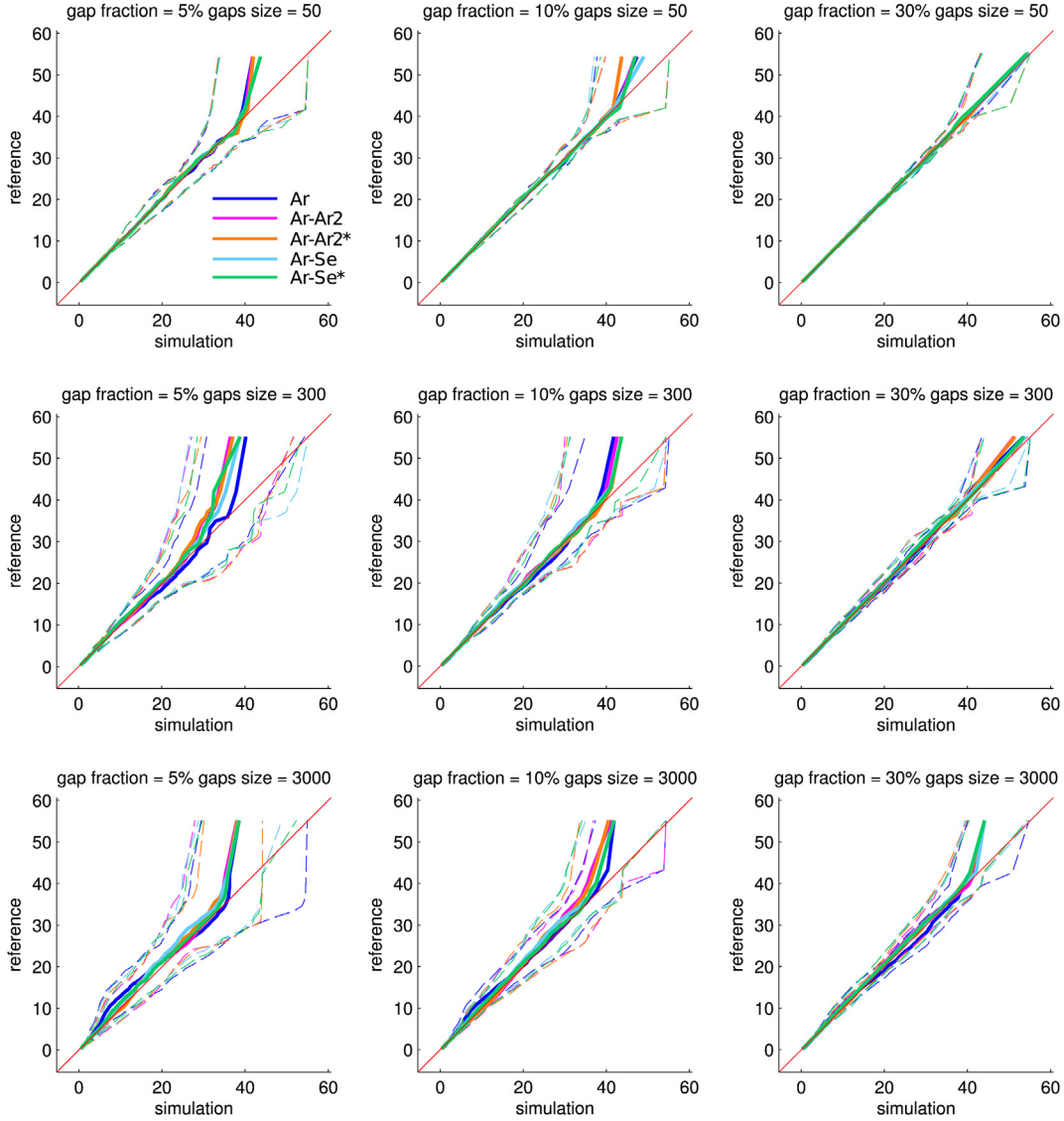


Fig. 3. Multiple scenario experiment: QQ-plot of the reference distribution against the simulation ensemble for all tests and gap scenario classes. Each graph contains the simulation of a specific fraction-size combination (10 realizations by 10 scenarios): the missing fraction increases from left to right and the gap size from top to bottom. Each test is indicated with a different color, solid lines indicate the realization median and dashed lines the 5-95th percentile boundary.

the time-series with a negligible impact on the hydrological regime. Therefore, their underrepresentation is not a main issue unless the user is studying the extremal behavior of the process at the 10-min scale. Increasing the missing data quantity, this bias tends to disappear (30% of total data set, Fig. 3 right column). In this case, the algorithm scans more deeply the training data set to generate a larger number of data patterns, with a higher chance to sample extreme values. The 5th and 95th percentile boundaries (dashed lines) of the realizations indicate the uncertainty on the recovered statistical distribution. This is larger when a small percentage of the data is missing since it is dependent on the statistical content of the missing data, different for each scenario (Fig. 3 left column). Conversely, with a larger missing data amount (Fig. 3 right column), the missing statistical content varies much less depending on the scenario and also the uncertainty of its estimation is lower. In summary, these results show that the algorithm can efficiently recover the main statistical content even when no auxiliary information is used (test Ar).

3.3. Predictive power

In this section, the predictive performance of the technique is analyzed by means of some goodness-of-fit indicators computed for each test and gap fraction-size combination. The box-plot of the root mean squared error (RMSE) between the simulation and the reference missing data is shown in Fig. 4: the results are grouped hierarchically by test, missing percentage and gap size (indicated by different colors).

The most important influence on the prediction is played by the gap size: for all simulation groups, RMSE is lower than 1 m^3/s in case of small-sized gaps (50 time steps) and does not present any substantial change as a function of the missing percentage and test type. This can be explained by the fact that the variable is highly autocorrelated and does not show big variations in 50 missing time steps (see for example Fig. 2). The variability of the missing data in gaps of this size is very limited and can be efficiently performed without using any auxiliary information. A comparable result may be achieved with a reliable method of interpolation. When the gap size is larger, the variability of the possible data patterns is higher

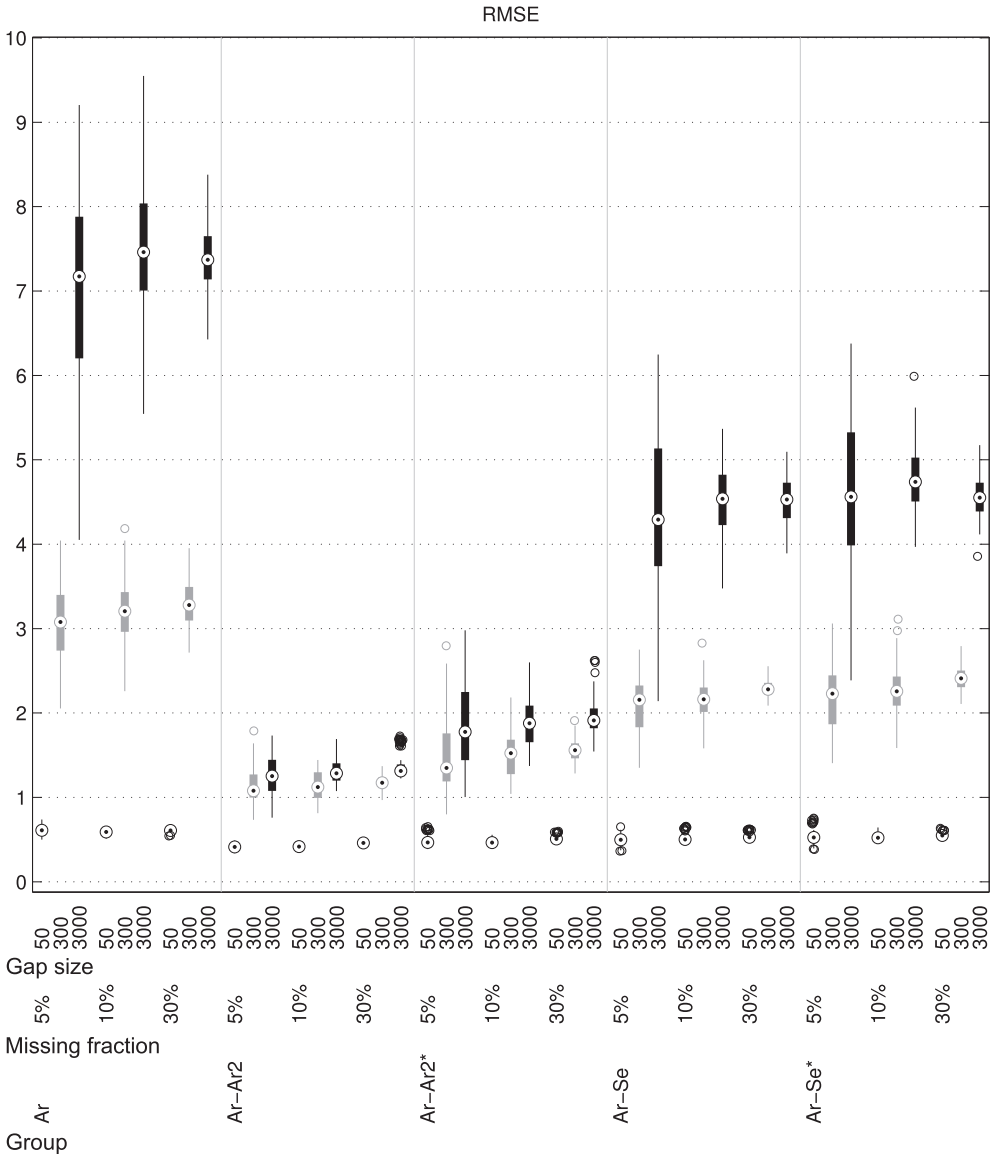


Fig. 4. Multiple scenario experiment: Boxplot of the root mean square error (RMSE) computed on the simulated missing flow rate data [m^3/s] for all tests and gap scenario classes.

and the simulation is much more dependent on the setup used. For test Ar, where Q is not used, RMSE increases rapidly with the gap size: between 2 and 4 m^3/s for 300 time-step gaps and between 4 and 10 m^3/s for 3000 time-step gaps, with the median of the realizations between 7 and 8 m^3/s. For all tests, the variability of the performance is dependent on the gap size and missing percentage: as shown by the interquartile range (thick part of the box-plots), the error variability is maximized for largest gap sizes and smallest missing percentage. In this case, the performance is very dependent on the content of the random missing portion, while, increasing the missing percentage or reducing the gap size, this effect is statistically compensated and the RMSE of the simulation ensemble converges towards its median. The error on large gaps is limited below 2 m^3/s if a highly correlated Q variable is used (Ar-Ar2). This confirms the efficiency of a highly correlated auxiliary variable in reducing the uncertainty of the simulation. Conversely, a moderate RMSE increase is observed when Q is incomplete (Ar-Ar2*). Test Ar-Se and Ar-Se*, where Q is poorly correlated to Z, present a performance in-between tests Ar and Ar-Ar2: the RMSE is between 1 and 3 m^3/s for 300 time-step gaps and between 2 and 6.5 m^3/s with median around 4.5 m^3/s for 3000 time-step gaps.

The Pearson's correlation coefficient (R, Fig. 5) and the Nash-Sutcliffe model efficiency coefficient (NSE, Fig. 6) confirm the same results shown by the RMSE. R can be interpreted as the fraction of the reference variability predicted by the simulation: for example PCC = 0.7 the model can predict the 70% of the variability. NSE is of less immediate interpretation: NSE = 0 indicates that the simulation has the same predicting power as the estimated mean, NSE = 0.7 indicates that the RMSE is equal to the 30% of the observed variance (Legates and McCabe, 1999) and NSE = 1 corresponds to a perfect prediction. Referring to Fig. 5 and 6, we can consider the predictive performance of the simulation very good (PCC > 0.9 and NSE > 0.8) in case of small gaps for all considered tests, in case of medium gaps when Q is used and in case of large gaps when Q is highly correlated to Z (Ar-Ar2 and Ar-Ar2*). In absence of Q (test Ar), the prediction remains reliable only for gaps up to the medium size. The prediction is not efficient in case of large gaps and absent or lowly correlated Q variable. In these cases, the flood sequence generated by the algorithm may still be a realistic estimation, but it does not correspond to the present one in the

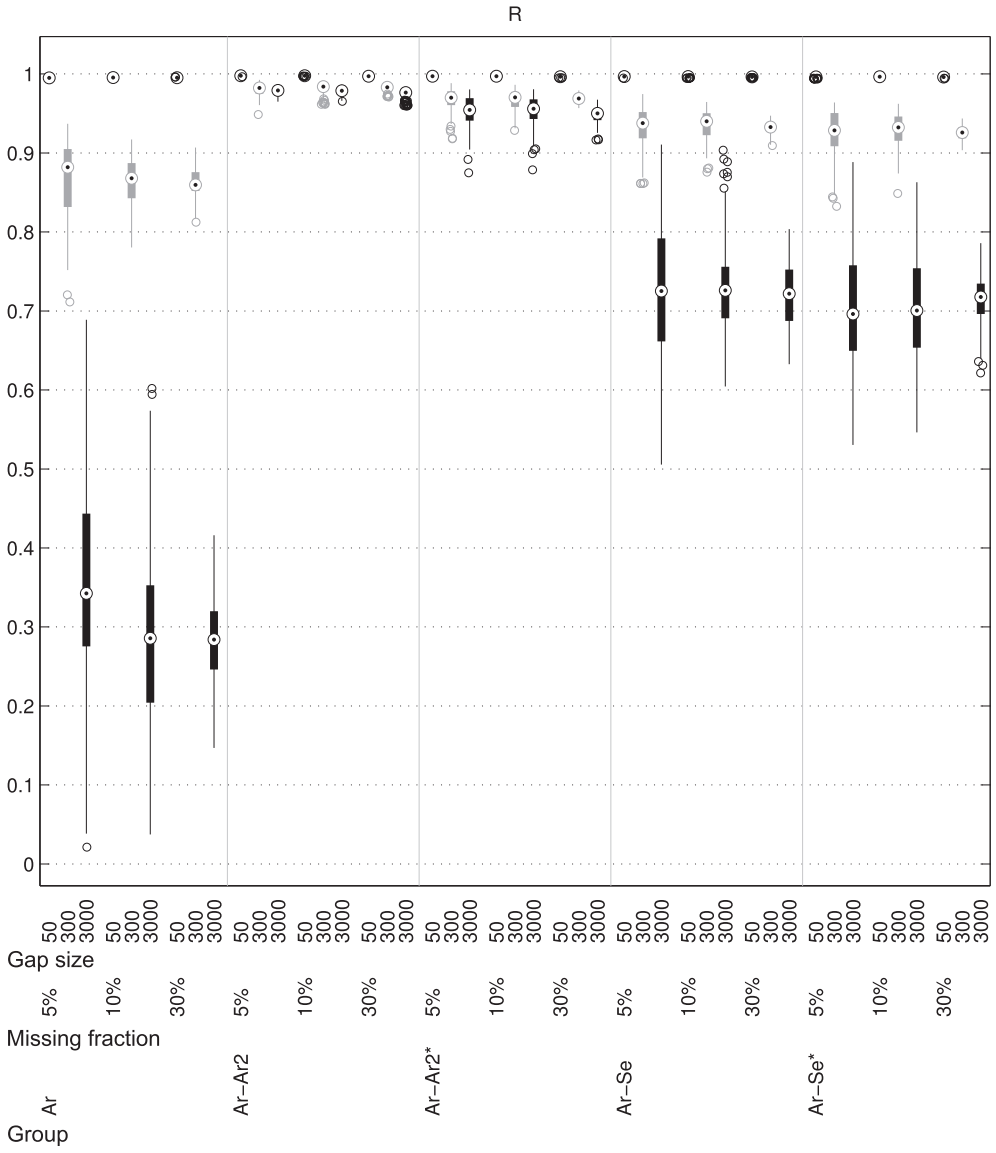


Fig. 5. Multiple scenario experiment: Boxplot of the Pearson's correlation coefficient (PCC) computed on the simulated missing data portions for all tests and gap scenario classes.

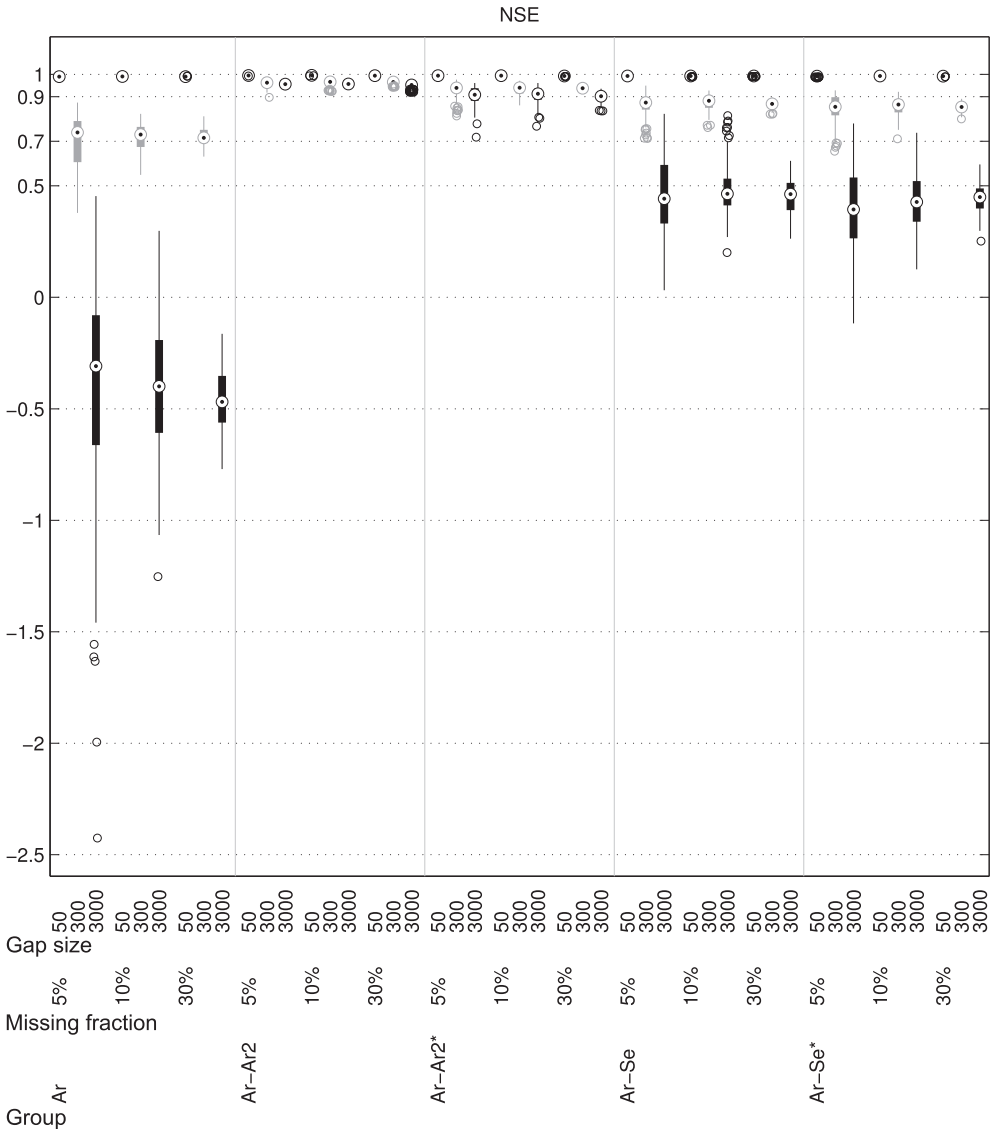


Fig. 6. Multiple scenario experiment: Boxplot of the NashSutcliffe model efficiency (NSE) coefficient computed on the simulated missing data portions for all tests and gap scenario classes.

reference since it is not properly guided by Q and it explores a larger uncertainty space, without preserving the actual flood sequence.

4. Comparative test results

We analyze here the performance of the DS technique compared to the one of the ARMAX model, as described in Section 2.5. Among the data sets shown in the paper, the ones used here are Ar-Ar2 and Ar-Se.

4.1. Visual comparison

Fig. 7 shows a time-series portion simulated by both techniques for the two data sets used. For Ar-Ar2, where the auxiliary variable Ar2 (not shown in Fig. 7) is well correlated to the target Ar, the simulation ensemble generated by both techniques follows the reference time-series quite accurately. The specific flood sequence is efficiently preserved as shown by the simulation mean (solid lines). Since the auxiliary variable is highly informative in this case, the explored uncertainty space (shaded areas) is generally narrow and it includes the reference time-series. Nevertheless, the DS technique presents a higher variability in the simulations in corresponding to hydrograph peaks. This allows a more reliable estimation of the local uncertainty, while, with the ARMAX technique, sometimes the local extremes lie outside the space covered by the simulation ensemble. When the auxiliary variable is poorly informative (Ar-Se test), the uncertainty on the prediction is higher: ARMAX preserves the flood sequence but underestimates the local extremes systematically. Conversely, the DS technique generates a more representative simulation ensemble, without always following the reference flood sequence, but usually including the reference values. Finally, the ARMAX time-series shows a higher small-scale noise with respect to DS.

4.2. Statistical content

The statistical content of the simulation is compared with the reference by means of qq-plots (top of Fig. 8). For the test Ar-Ar2, both techniques can preserve the reference probability distribution, although ARMAX tends to over-estimate the extremal part.

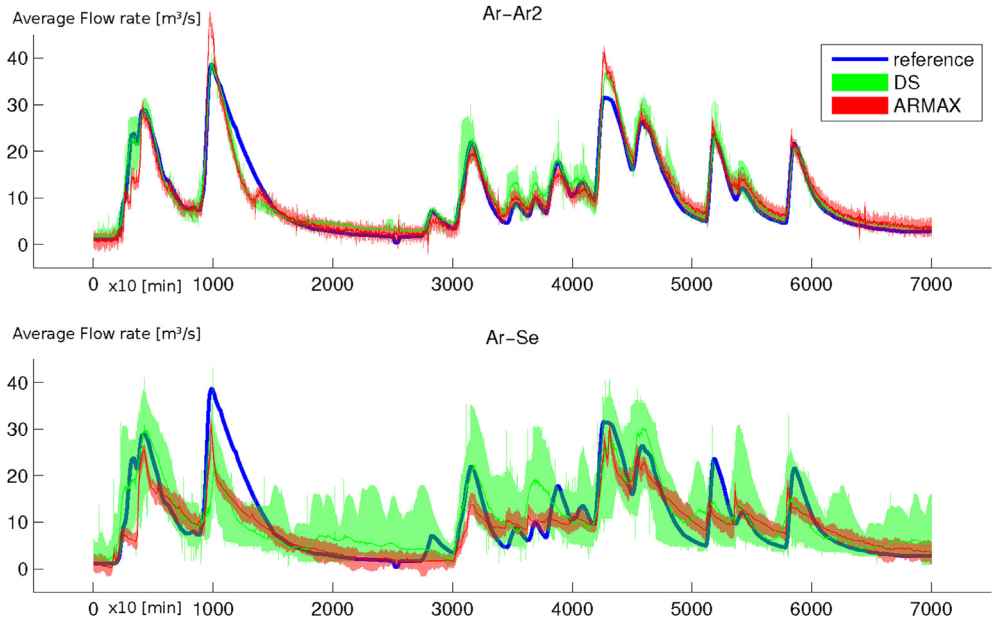


Fig. 7. Comparison experiment: a flow rate time-series portion (m^3/s, 10-min average, approximately 48 days) of the reference and simulated time-series. For both techniques, the transparent color area represents the 5–95% of the simulation ensemble. The solid lines represent the mean of the simulations.

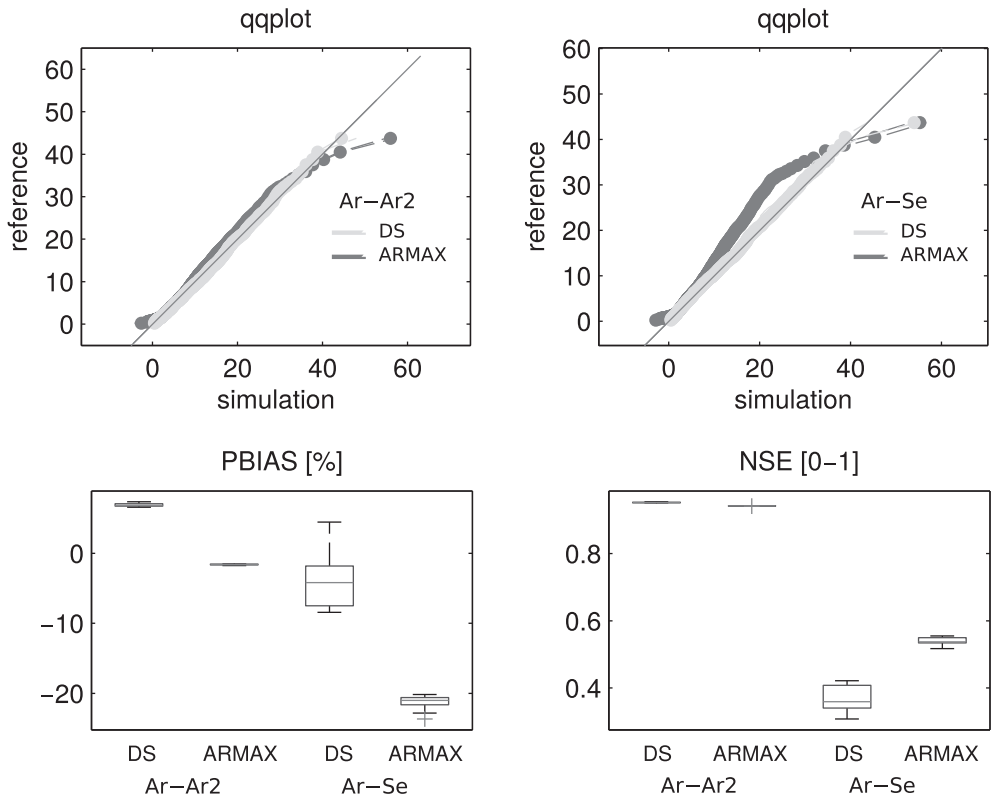


Fig. 8. Comparison experiment: qq-plot of the reference distribution against the simulation ensemble (dots indicate the realization median and dashed lines the 5-95th percentile boundary), boxplot of the percentage bias (PBIAS) and NashSutcliffe model efficiency (NSE) for both techniques and simulation groups.

Conversely, for test Ar-Se, ARMAX shows and under-estimation of the local intense values (30–40 m^3/s), in accordance with the results of Section 4.1. Finally, for both techniques, the simulation ensemble shows a very low uncertainty about the estimated distribution, as shown by the 5–95 percentile boundaries (dashed lines).

4.3. Predictive power

The predictive power of both techniques is compared with two statistical indicators (bottom of Fig. 8): the percentage bias (PBIAS) and the NashSutcliffe model efficiency (NSE). In the test Ar-Ar2 the ARMAX model shows a less biased estimation (PBIAS between −1

and −2%) with respect to DS (PBIAS between 6 and 8%), while, in the test Ar-Se, DS shows a more variable performance (PBIAS between 5 and -9%) and ARMAX significantly increase the bias towards negative values (PBIAS between −20 and −28%). This last result is probably linked to the underestimation of the local extremes as shown in the previous indicators. According to the NSE, both techniques have a good predictive power for test Ar-Ar2, with values higher than 0.8 for ARMAX and 0.9 for DS. Conversely, for test Ar-Se, the performance is poorer for DS (NSE around 0.35) than ARMAX (NSE around 0.55). This last results may be linked to the fact that DS presents a higher variability in the simulation, for example generating peaks where they don't occur (se bottom of Fig. 7), with a subsequent increase in the error with respect to the reference. Conversely, ARMAX, although underestimating the uncertainty, respects better the main flood pattern, resulting in a higher NSE.

5. Discussion and conclusions

The aim of this paper was to propose and test a stochastic methodology for missing data simulation inside hydrological flow rate time-series based on the Direct Sampling technique (DS). Its rationale is fairly simple: without imposing a statistical model for the process of interest, the missing data are simulated by resampling data patterns of the variable of interest together with a group of auxiliary variables. By scanning the available data, a similar pattern is found and the datum at its center is assigned at the simulated time step. The process is repeated until all the data set is complete. Since multiple neighbors and different pattern size are considered for conditioning, realistic structures at multiple scales can be generated.

A standard setup for flow rate time-series is proposed, including the variable of interest (Z) and a series of auxiliary variables: a couple of periodic theoretical functions describing the annual seasonality, a predictor variable (Q), which is a correlated flow rate time-series, and a categorical variable computed on Q describing the hydrographic structure as a succession of rising and recessing flood limbs. The setup can be adapted to any type of flow rate time-series, with or without the use of auxiliary variables, but it may require an adjustment of the parameters.

The model is tested on the gap filling of a high-resolution (10-min) karst flow rate time-series from the Areuse St. Sulpice station (Jura mountains, Switzerland), belonging to a flashy spring discharge system that exhibits abrupt changes in the hydrograph. The performance of the technique is analyzed by considering different auxiliary time-series Q. Moreover, variable missing percentage up to 30% and multiple gaps of size up to 3000 time steps, corresponding to about 20 days. The generated missing data portions show realistic asymmetrical hydrographic structures similar to the one found in the reference even when large gaps are simulated and no Q variable is used. The statistical content lost with the missing data is mainly recovered even when these constitute large portions of the data set, but extreme values may be underrepresented in some cases. For this reason, the training data set (being the incomplete time-series itself or another data portion) should correctly represent the high return time events, since direct sampling is not able to extrapolate values not observed in the available data. Finally, the predictive power of the technique, measured by classical goodness-of-fit measures, is very high when Q, even if incomplete, is highly correlated to Z. If Q is absent or poorly correlated to Z, the prediction is more uncertain since the variability of the possible data patterns within large gaps is much higher.

In the last group experiments, DS is compared with a classical time-series model of type ARMAX. The results show that DS is capable to generate more realistic simulated time-series with respect to the concurrent technique, that heavily rely on the linear dependency with recent past time-step of the target and auxiliary variables. In fact, when the auxiliary variable is highly correlated with the target, the two techniques have a comparable prediction power, with DS recovering more efficiently the probability distribution and ARMAX showing a lower bias. Conversely, when the auxiliary variable is poorly correlated and contains information only about the main flood sequence, a technique relying on linear correlation and a simple error structure is not sufficient to represent the entire variability of the process. Although better representing the main flood pattern and resulting in a higher NSE value, ARMAX does not preserve the sufficient variability in the simulations, underestimating the uncertainty and the local extremes. It is also noted that the computation time for ARMAX is much lower than DS depending on the implementation used for both techniques: the latter can be significantly accelerated with parallel-core implementation, but it requires a cpu time significantly higher (about 10 times in this case) than the ARMAX calibration and simulation. Nevertheless, ARMAX requires a more complex parameterization, e.g. to decide the order of the model time dependency, while DS is mainly data driven and adaptive to different data sets: in fact its setup usually needs minimal modification when changing data set (see e.g. Oriani et al., 2014).

Compared to DS, classical time-series techniques like ARMAX remain a more parsimonious option with a similar performance, but only if the variables conditioning the simulation are complete and highly informative. Since they are based on a specific cross-correlation structure, they cannot deal with uninformed conditioning variables. Moreover, their time dependency is usually based on past values, ignoring the data subsequent to each gap. In these cases, DS constitutes a more attractive approach since, as seen in the first group of experiments, it uses the available portions of the auxiliary data to condition the simulation, exploring in a rather realistic way the remaining uncertainty. This makes the proposed method a convenient alternative for gap-filling in the everyday professional practice, with the only requirement of a representative training data set. Future developments may include the use of other types of auxiliary source of information, like for example a rainfall amount time-series, to better represent lagged processes or recurring events that introduce non-stationarity or periodicity in the statistical properties of the target variable.

Acknowledgements

This research work has been funded by the Swiss National Science Foundation (projects #134614 and #162040).

References

Allard, D., Froidevaux, R., Biver, P., 2006. Conditional simulation of multi-type non stationary markov object models respecting specified proportions. Math. Geol. 38 (8), 959–986.

Aminian, K., Ameri, S., Dec. 2005. Application of artificial neural networks for reservoir characterization with limited data. J. Pet. Sci. Eng. 49 (3–4), 212–222 wOS:000234302300009.

Bahrami, J., Kavianpour, M.R., Abdi, M.S., Telvari, A., Abbaspour, K., Rouzkhash, B., Apr. 2011. A comparison between artificial neural network method and nonlinear regression method to estimate the missing hydrometric data. J. Hydroinform. 13 (2), 245–254 wOS:000289376800008.

Bamberger, I., Hoertnagl, L., Walser, M., Hansel, A., Wohlfahrt, G., 2014. Gap-filling strategies for annual VOC flux data sets. Biogeosciences 11 (8), 2429–2442 wOS:000335374200023.

Bennis, S., Berrada, F., Kang, N., Apr. 1997. Improving single-variable and multivariable techniques for estimating missing hydrological data. J. Hydrol. 191 (1–4), 87–105 wOS: A1997XG47200005.

Box, G.E., Jenkins, G.M., 1976. Time Series Analysis, Control, and Forecasting. Holden Day, San Francisco, CA.

Buishand, T.A., Brandsma, T., Nov. 2001. Multisite simulation of daily precipitation

and temperature in the rhine basin by nearest-neighbor resampling. Water Resour. Res. 37 (11), 2761–2776.
Clark, M.P., Gangopadhyay, S., Brandon, D., Werner, K., Hay, L., Rajagopalan, B., Yates, D., Apr. 2004. A resampling procedure for generating conditioned daily weather sequences. Water Resour. Res. 40 (4), W04304.
Dastorani, M.T., Moghadamnia, A., Piri, J., Rico-Ramirez, M., Jun. 2009. Application of ANN and ANFIS models for reconstructing missing flow data. Environ. Monit. Assess. 166 (1–4), 421–434. http://link.springer.com/article/10.1007/s10661-009-1012-8.
Diamantopoulou, M.J., Dec. 2010. Filling gaps in diameter measurements on standing tree boles in the urban forest of thessaloniki, Greece. Environ. Model. Softw. 25 (12), 1857–1865.
Dumedah, G., Walker, J.P., Chik, L., Jul. 2014. Assessing artificial neural networks and statistical methods for infilling missing soil moisture records. J. Hydrol. 515, 330–344 wOS:000338605900030.
Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H., Granier, A., Gross, P., Grnwald, T., Hollinger, D., Jensen, N.-O., Katul, G., Keronen, P., Kowalski, A., Lai, C.T., Law, B.E., Meyers, T., Moncrieff, J., Moors, E., Munger, J.W., Pilegaard, K., Rannik, U., Rebmann, C., Suyker, A., Tenhunen, J., Tu, K., Verma, S., Vesala, T., Wilson, K., Wofsy, S., Mar. 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. Agric. For. Meteorol. 107 (1), 43–69. http://www.sciencedirect.com/science/article/pii/S0168192300002252.
Guardiano, F., Srivastava, R., 1993. Multivariate geostatistics: beyond bivariate moments. In: Geostatistics-Troia, 1, pp. 133–144.
Kondrashov, D., Denton, R., Shprits, Y.Y., Singer, H.J., Apr. 2014. Reconstruction of gaps in the past history of solar wind parameters. Geophys. Res. Lett. 41 (8), 2702–2707 wOS:000335809800006.
Lamrini, B., Lakhal, E.-K., Le Lann, M.-V., Wehenkel, L., Jun. 2011. Data validation and missing data reconstruction using self-organizing map for water treatment. Neural Comput. Appl. 20 (4), 575–588 wOS:000290319100013.
Legates, D.R., McCabe, G.J., Jan. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241 wOS:000078123800021.
Ljung, L., 1999. System Identification: Theory for the User, PTR Prentice Hall Information and System Sciences Series. Prentice Hall, New Jersey.
Mariethoz, G., Linde, N., Jougnot, D., Rezaee, H., Oct. 2015. Feature-preserving interpolation and filtering of environmental time series. Environ. Model. Softw. 72, 71–76.
Mariethoz, G., McCabe, M.F., Renard, P., Oct. 2012. Spatiotemporal reconstruction of gaps in multivariate fields using the direct sampling approach. Water Resour. Res. 48, W10507 wOS:000309608800003.
Mariethoz, G., Renard, P., Straubhaar, J., Nov. 2010. The direct sampling method to perform multiple-point geostatistical simulations. Water Resour. Res. 46, W11536.
Meerschman, E., Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M., Renard, P., Mar. 2013. A practical guide to performing multiple-point statistical simulations with the direct sampling algorithm. Comput. Geosci. 52, 307–324.
Moffat, A.M., Papale, D., Reichstein, M., Hollinger, D.Y., Richardson, A.D., Barr, A.G., Beckstein, C., Braswell, B.H., Churkina, G., Desai, A.R., Falge, E., Gove, J.H., Heimann, M., Hui, D., Jarvis, A.J., Kattge, J., Noormets, A., Stauch, V.J., Dec. 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agric. For. Meteorol. 147 (3–4), 209–232 wOS:000251469900009.
Nkuna, T.R., Odiyo, J.O., 2011. Filling of missing rainfall data in Luvuvhu River Catchment using artificial neural networks. Phys. Chem. Earth 36 (14–15), 830–835 wOS:000296306100013.
Nourani, V., Baghanam, A.H., Gebremichael, M., Mar. 2012. Investigating the ability of artificial neural network (ANN) models to estimate missing rain-gauge data. J. Environ. Inf. 19 (1), 38–50 wOS:000302846300005.
Oriani, F., 2015. Stochastic Simulation of Rainfall and Climate Variables Using the Direct Sampling Technique (Ph.D. thesis). Universit de Neuchtel.
Oriani, F., Straubhaar, J., Renard, P., Mariethoz, G., 2014. Simulation of rainfall time series from different climatic regions using the direct sampling technique. Hydrol. Earth Syst. Sci. 18 (8), 3015–3031.
Painter, S.L., Sun, A., Green, R.T., August 4, 2008. Enhanced Characterization and Representation of Flow Through Karst Aquifers. Chapter 1 introduction. American Water Works Association. ISBN-10: 1843399792, ISBN-13: 978-1843399797.
Rajagopalan, B., Lall, U., Oct. 1999. A k-nearest-neighhor simulator for daily precipitation and other weather variables. Water Resour. Res. 35 (10), 3089–3101.
Schoellhamer, D.H., Aug. 2001. Singular spectrum analysis for time series with missing data. Geophys. Res. Lett. 28 (16), 3187–3190 wOS:000170348100033.
Straubhaar, J., September 2015. DeeSse Technical Reference Guide. Centre d'Hydrogeologie et de Geothermie. University of Neuchâtel.
Straubhaar, J., Renard, P., Mariethoz, G., Froidevaux, R., Besson, O., 2011. An improved parallel multiple-point algorithm using a list approach. Math. Geosci. 43 (3), 305–328.
Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. 34 (1), 1–21.
Wang, Q.J., Apr. 2008. A bayesian method for multi-site stochastic data generation: dealing with non-concurrent and missing data, variable transformation and parameter uncertainty. Environ. Model. Softw. 23 (4), 412–421.
Wang, S.H., Sep. 2003. Application of self-organising maps for data mining with incomplete data sets. Neural Comput. Appl. 12 (1), 42–48 wOS:000186039600006.
Wojcik, R., Buishand, T.A., Mar. 2003. Simulation of 6-hourly rainfall and temperature by two resampling schemes. J. Hydrol. 273 (1–4), 69–80.