Journal of Hydrology 374 (2009) 185–195

Contents lists available at ScienceDirect

Journal of Hydrology

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

A daily weather generator based on a two-stage resampling algorithm

Robert Leander *, T. Adri Buishand

Royal Netherlands Meteorological Institute (KNMI), P.O. Box 201, 3730 AE De Bilt, The Netherlands

ARTICLE INFO

Article history:
Received 30 June 2008
Received in revised form 3 April 2009
Accepted 1 June 2009

This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of Taha Ouarda, Associate Editor

Keywords:
Meuse basin
Time series
Nearest-neighbour resampling
Hydrological modelling
Extreme values

SUMMARY

A two-stage time-series resampling algorithm is presented that is capable of generating daily values of weather variables outside their historical ranges. In this algorithm the simulated daily values are composed of an expected value and a sampled historical residual. The residuals broaden the range of the simulated daily values. Both the estimation of the expected value and the sampling of the residuals are based on the nearest-neighbours concept. In particular the influence of the neighbourhood sizes in both nearest-neighbour searches was studied. The algorithm was tested with data generated by two theoretical time-series models.

Using observed precipitation and temperature data, a 12,000-year series of precipitation and temperature for the Ourthe catchment (Belgium) was simulated and used as input for a rainfall–runoff model to produce a long synthetic sequence of daily discharge. The two-stage algorithm correctly reproduces the mean, standard deviation and lag 1 autocorrelation of daily precipitation. The simulated distributions of 4-day and 10-day precipitation maxima in winter also show good correspondence with those observed, while the largest daily amounts substantially exceed those in the original data. However, the widened range of daily precipitation amounts has no discernible effect on the simulated discharge maxima in winter.

Introduction

The stochastic generation of weather variables relevant to hydrologic simulation has long been of interest. This is reflected by the variety of papers on this topic, in particular on the generation of daily precipitation sequences (the oldest dating from the 1970s). Initially parametric models were used. Quite often the occurrence of precipitation was described by a two-state Markov chain or an alternating renewal process and the amount of precipitation on a wet-day by a positively skewed distribution (gamma or mixed exponential), see Woolhiser (1992) for a review. Transformations of the multivariate normal distribution have been considered for multi-site simulation (e.g., Richardson, 1977; Bárdossy and Plate, 1992; Wilks, 1998). Nonhomogeneous hidden Markov chains have been developed for the conditional simulation of daily precipitation on large-scale weather variables (e.g., Hughes and Guttorp, 1994; Charles et al., 1999). Parametric models require rather restrictive assumptions regarding the probability distributions and the correlation structure. Nonparametric models avoid this difficulty and are therefore gaining in popularity. Examples of this approach are nearest-neighbour resampling (e.g., Young, 1994; Lall and Sharma, 1996; Rajagopalan and Lall, 1999) and methods involving kernel-density estimation techniques (e.g., Harrold et al., 2003a,b). An advantage of nearest-neighbour resampling is that it can easily be extended to multi-site simulation (Buishand and Brandsma, 2001). A significant disadvantage of the nearest-neighbour algorithm, however, is the fact that it cannot produce amounts beyond those present in the sequences used as base material. This limitation may not hamper applications relying mainly on extreme multi-day precipitation. However, in some applications it might be desirable to allow for daily amounts beyond the maximum found in the historical record.

In this study an extension of the nearest-neighbour resampling algorithm is explored that is capable of generating larger daily precipitation amounts than those observed. This is achieved by a two-stage resampling scheme. In the first stage the expected amount is determined, using a nearest-neighbour regression. In the second stage the expected amount is multiplied by a randomly selected residual factor, which is also based on a nearest-neighbour search. The general concept is first tested for two univariate cases using simulated data from theoretical time-series models. Subsequently, this idea is used to generate sequences of daily precipitation and temperature of a river catchment in the Ardennes, Belgium. The properties of the resampled precipitation are studied in detail. A 12,000-year synthetic series of daily precipitation and temperature serves as input for simulations with the semi-distributed rainfall–runoff model HBV (Lindström et al., 1997). The results for extreme floods are compared to the results of a traditional nearest-neighbour resampling algorithm. Additionally, a series resampled with

* Corresponding author. Tel.: +31 30 2206434; fax: +31 30 2210407.
E-mail address: leanderr@knmi.nl (R. Leander).

0022-1694/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2009.06.010

this algorithm is considered in which the largest amounts are replaced by random values from the tail of an exponential distribution.

Method

The simulation procedure is based on the decomposition of the variable X into an expected value M conditional on prior values of X and a non-negative residual factor e with unit mean, i.e.

$$X = Me. \tag{1}$$

The procedure of simulating a value X_j^* for a certain day j consists of two stages, the estimation of its expected value M_j from the simulated values for previous days and the generation of a residual e_j^* (throughout this paper the superscript '*' is used to signify simulated values). This requires information on the distribution of e to be extracted from the historical record.

Prior to the simulation, the expected value M_i of X_i for each day i in the historical sequence is estimated by a nearest-neighbour regression. 'Nearest-neighbour' (Lall and Sharma, 1996) here refers to a day $l \neq i-1$ that is similar to day $i-1$ in terms of $\delta_{l,i-1} = |X_l - X_{i-1}|$ (or a weighted Euclidean distance in the case of multiple variables characterizing day $i-1$). The estimate $\widehat{M}_i$ of M_i can be expressed as a linear combination of the successors of the sorted k_M nearest-neighbours of day $i-1$ with coefficients λ_k:

$$\widehat{M}_i = \sum_{k=1}^{k_M} \lambda_k X_{\mathrm{nnb}(k,i-1)+1}, \tag{2}$$

where nnb(k,$i-1$) refers to the kth closest neighbour of X_{i-1}. The residual $\widehat{e}_i$ of day i then equals the historical value X_i divided by $\widehat{M}_i$.

Simulating daily values proceeds in an analoguous way. For each new day j in the simulation, an estimate $\widetilde{M}_j$ of the expected value is calculated from the historical successors of the nearest-neighbours of the last simulated value X_{j-1}^* as:

$$\widetilde{M}_j = \sum_{k=1}^{k_M} \lambda_k X_{\mathrm{nnb}(k,j-1)+1}. \tag{3}$$

From the k_e historical days of which $\widehat{M}_i$ is closest to $\widetilde{M}_j$, one of the residuals $\widehat{e}_i$ is randomly selected as the simulated residual e_j^*, using the decreasing **1**/**k**-kernel introduced by Lall and Sharma (1996):

$$\alpha_k = \frac{1/k}{\sum_{k=1}^{k_e} 1/k} \quad 1 \leqslant k \leqslant k_e \tag{4}$$

with α_k the probability of selecting the residual of the kth closest neighbour. From the expected value and the residual factor the simulated value for day j then becomes:

$$X_j^* = \widetilde{M}_j \, e_j^*. \tag{5}$$

Based on a suggestion of Lall and Sharma (1996), Prairie et al. (2006) followed an analoguous approach to simulate values beyond the observed range. They used an additive algorithm instead of a multiplicative algorithm. In particular with a view to simulating precipitation, the latter has the advantage that the generation of negative values is avoided. Another difference with the algorithm discussed here is that in their algorithm the residual was sampled from the successor of one of the nearest-neighbours used for estimating M.

The decomposition of the historical value X_i into an expected value $\widehat{M}_i$ and a residual $\widehat{e}_i$ can be summarized as follows:

1. Find and sort the k_M nearest-neighbours of the historical day $i-1$.
2. Determine the expected value $\widehat{M}_i$ from the historical successors of these nearest-neighbours (Eq. (2)).
3. Obtain the residual as $\widehat{e}_i = X_i/\widehat{M}_i$.

The simulation is then initialized by selecting a historical day at random and values at subsequent days j are simulated:

1. Find and sort the k_M nearest-neighbours of the last simulated day $j-1$.
2. Determine the expected value $\widetilde{M}_j$ of the new day j as a weighted average of the historical successors of these nearest-neighbours (Eq. (3)).
3. Find and sort the k_e historical days of which $\widehat{M}$ is nearest to $\widetilde{M}_j$.
4. Sample the residual $\widehat{e}$ of one of these days as e_j^* using the **1**/**k**-kernel (Eq. (4)).
5. Multiply $\widetilde{M}_j$ and e_j^* to simulate the value X_j^* for day j.

For the hydrological application in this study this algorithm has been extended in order to generate daily precipitation and temperature simultaneously, thereby preserving the dependence between both variables.

Theoretical models

The algorithm described above was first tested with data from theoretical time-series models which both generate only positive values. The advantage of considering a theoretical model lies in the fact that its statistical properties can usually be derived exactly from the model formulation. Two first-order autoregressive (AR1) models were considered, a lognormal AR1 model and an exponential AR1 model. The lognormal model was chosen because it is multiplicative, similar to the algorithm. The exponential model is useful to detect possible effects of non-multiplicativity on the performance of the algorithm. Both models have commonly been used within a hydrologic context, usually for the stochastic simulation of streamflows.

Lognormal AR1 model

A sequence $\{X_i\}$ of correlated standard lognormal variables can be generated by transforming the values $\{Y_i\}$ of a normal AR1 process:

$$Y_i = \rho Y_{i-1} + \varepsilon_i \sqrt{1-\rho^2}. \tag{6}$$

$$\begin{aligned} X_i = \exp(Y_i) &= \exp\left[\rho \log(X_{i-1}) + \varepsilon_i \sqrt{1-\rho^2}\right] \\ &= X_{i-1}^{\rho} \exp\left(\varepsilon_i \sqrt{1-\rho^2}\right), \end{aligned} \tag{7}$$

where the $\{\varepsilon_i\}$ are independent standard-normal variables. The $\{X_i\}$ have mean $\sqrt{\mathrm{e}} \approx 1.65$ and standard deviation $\sqrt{\mathrm{e}(\mathrm{e}-1)} \approx 2.16$. The lag 1 autocorrelation coefficient ϱ_1 of the lognormal process $\{X_i\}$ can be derived from the lag 1 autocorrelation coefficient ρ of the underlying AR1 process $\{Y_i\}$ using (Mejía and Rodríguez-Iturbe, 1974):

$$\varrho_1 = \frac{\exp(\rho) - 1}{\mathrm{e} - 1}. \tag{8}$$

The expectation M_i follows from Eq. (7):

$$\begin{aligned} M_i = E(X_i|X_{i-1}) &= X_{i-1}^{\rho} \exp\left[\frac{1}{2}(1-\rho^2)\right] \\ &= \exp\left[\rho \log X_{i-1} + \frac{1}{2}(1-\rho^2)\right]. \end{aligned} \tag{9}$$

Note that M_i is nonlinear in X_{i-1}. For this AR1 process the residuals

$$e_i = \frac{\exp\left(\varepsilon_i \sqrt{1-\rho^2}\right)}{\exp\left(\frac{1}{2} - \frac{1}{2}\rho^2\right)} \tag{10}$$

have a lognormal distribution with

$$E(e_i) = 1, \quad \mathrm{Var}(e_i) = \exp(1-\rho^2) - 1. \tag{11}$$

Since the distribution of $\log X_{i-1}$ in Eq. (9) is standard normal, the distribution of M_i is lognormal with mean and variance

$$E(M_i) = \exp\left(\frac{1}{2}\right) \approx 1.65, \quad \mathrm{Var}(M_i) = \mathrm{e}\left[\exp(\rho^2) - 1\right]. \tag{12}$$

Here $\rho = 0.5$ was chosen, for which $\mathrm{Var}(M_i) \approx 0.77$, $\mathrm{Var}(e_i) \approx 1.12$ and $\varrho_1 \approx 0.378$. Furthermore, M_i and e_i are independent, because X_{i-1} and ε_i are independent.

Exponential AR1 model

An exponential AR1 model (EAR1) was presented by Gaver and Lewis (1980). Here this model is used to generate a sequence $\{X_i\}$ of correlated standard exponential variables. The EAR1 process has the same additive form as a normal AR1 process:

$$X_i = \rho X_{i-1} + \varepsilon_i. \tag{13}$$

The innovation ε_i equals zero with probability ρ and is positive with probability $1-\rho$, in which case it is sampled from the standard exponential distribution, i.e. $\Pr(\varepsilon_i > x) = (1-\rho)\exp(-x)$. A generalization of this process to gamma-distributed random variables (Gaver and Lewis, 1980; Lawrance, 1980) is related to the shot-noise models of Weiss (1977) for the generation of daily streamflow data. If the innovations ε_i are zero, the X_i decay exponentially, resembling streamflow recessions during dry periods. The autocorrelation stucture is the same as that of a normal AR1 process, i.e. the lag j autocorrelation coefficient ϱ_j equals ρ^j. A value of $\rho = 0.5$ was chosen. Though a nonlinear multiplicative first-order exponential autoregressive process is known in the literature (McKenzie, 1982; Fernandez and Salas, 1986), the EAR1 process is considered here to detect possible limitations of non-multiplicativity to the resampling procedure.

The conditional means $\{M_i\}$ and the corresponding residuals $\{e_i\}$ for this model are given by

$$M_i = \rho X_{i-1} + (1-\rho) \quad \text{and} \quad e_i = \frac{X_i}{M_i} = 1 + \frac{\varepsilon_i - (1-\rho)}{M_i}. \tag{14}$$

Table 1
Sensitivity of the mean, standard deviation s_d and lag 1 autocorrelation coefficient to the neighbourhood sizes k_M and k_e for resampling from lognormal data (left) and exponential data (right). The theoretical values, those extracted from the data (size 2000) and those of several resampling simulations (size 20,000) are listed.

k_M	k_e	Lognormal			Exponential		
		Mean	s_d	r_1	Mean	s_d	r_1
Theory		1.65	2.16	0.378	1.00	1.00	0.500
Data		1.69	2.22	0.376	1.02	1.01	0.500
50	100	1.69	2.24	0.281	1.04	1.02	0.406
100	100	1.66	2.23	0.288	1.03	1.01	0.422
200	100	1.71	2.29	0.298	1.03	1.01	0.440
400	100	1.68	2.23	0.329	1.02	1.02	0.458
800	100	1.69	2.25	0.308	1.02	1.00	0.454
400	50	1.67	2.16	0.321	1.04	1.01	0.440
400	200	1.70	2.26	0.309	1.05	1.02	0.441
400	400	1.70	2.25	0.315	1.03	1.02	0.439
400	800	1.72	2.33	0.323	1.02	1.03	0.439
400	1600	1.67	2.28	0.315	1.01	1.08	0.412

The means of M and e are both equal to one and the variance of M equals ρ^2. Contrary to the lognormal model, the variance of e_i depends on M_i:

$$\mathrm{Var}(e_i|M_i) = \frac{1}{M_i^2}\,\mathrm{Var}(\varepsilon_i) = \frac{1-\rho^2}{M_i^2}. \tag{15}$$

This dependence may put a restriction on k_e in the second step of the resampling algorithm. The $\{M_i\}$ follow a shifted exponential distribution with location parameter $1-\rho$ and scale parameter ρ, i.e.

$$\Pr(M \leqslant x) = 1 - \exp\left[-\frac{x-(1-\rho)}{\rho}\right], \quad x \geqslant 1-\rho \tag{16}$$

Simulation results

With each AR1 model a sequence of 2000 values was generated (from here on referred to as 'data'). From these sequences, simulations with a length of 20,000 values were obtained by applying the two-stage resampling algorithm. For the coefficients λ_k in Eqs. (2)

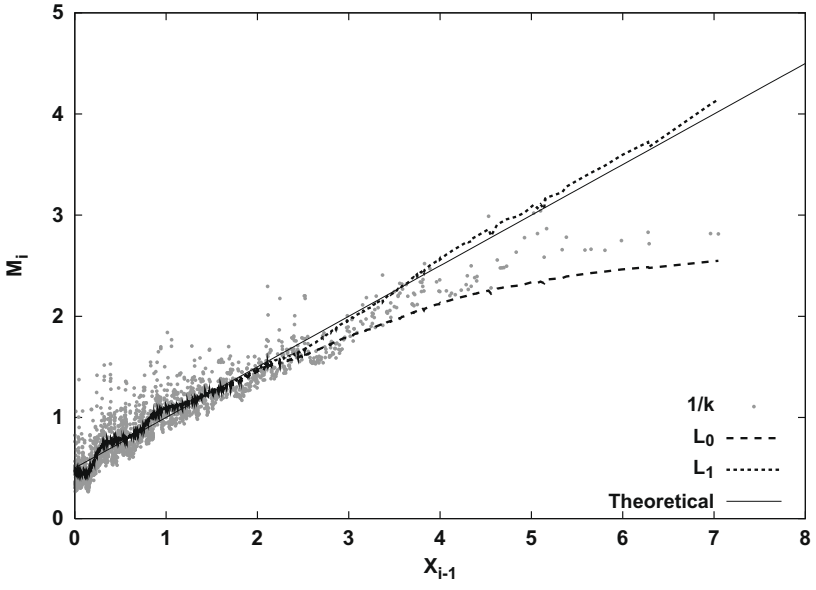

Fig. 1. Expected values $\widehat{M}_i$ versus the 'historical' predecessor X_{i-1} for the 2000 values from the exponential model, estimated as a weighted average with **1**/**k**-weights and by the zeroth- and first-order LOESS smoother (L_0 and L_1) with $k_M = 400$. The solid line represents the theoretical value from Eq. (14).

and (3) the **1**/**k**-kernel in Eq. (4) was used, but with k_M instead of k_e. The neighbourhood sizes k_M and k_e were expected to influence the statistical properties of the simulated series. Table 1 lists the mean, the standard deviation s_d and the lag 1 autocorrelation r_1 of resampled sequences from the data of the lognormal model (left) and the exponential model (right) with various settings of k_M and k_e. For each model the theoretical values and the empirical estimates derived from the data are given in the first two rows.

The theoretical mean and standard deviation of both models are reasonably well reproduced by most simulations and seem insensitive to the choice of k_M and k_e. Both parameters are slightly overestimated in most simulations. In both cases (lognormal as well as exponential) r_1 is underestimated for all combinations of k_M and k_e. Since persistence is introduced into the simulation through M, its estimation from the nearest-neighbours has been investigated in some detail. Fig. 1 compares the values of $\widehat{M}_i$ from the exponential data, using $k_M = 400$ (grey dots) with the theoretical value from Eq. (14) (solid straight line). The values of $\widehat{M}_i$ are considerably scattered. This scatter was suspected to be the source of the negative bias in r_1. To reduce the scatter, the LOESS smoother (Cleveland, 1979) was studied as an alternative for the estimation of M_i. In this method a polynomial is fitted to the successors of the nearest-neighbours of X_{i-1} by means of weighted least-squares with weights

$$w_k = \left[1 - \left(\frac{\delta_k}{\delta_{k_M}}\right)^3\right]^3, \tag{17}$$

where δ_k is the distance between X_{i-1} and its kth nearest neighbour and δ_{k_M} the largest distance within the neighbourhood. The estimate of M_i is calculated as the value of the fitted polynomial at X_{i-1}. The LOESS smoothers of degree zero L_0 (local constant) and degree one L_1 (local linear relation) are considered. In both cases the estimate of M_i can be written as a linear combination of the successors of the nearest-neighbours, as in Eq. (2). For L_0 the weights λ_k reduce to $w_k / \sum w_k$. For L_1 the expression of λ_k is more complex.

Table 2
Same as in Table 1, but now for two modifications of the original algorithm with respect to the estimation of M, respectively, based on the zeroth-order (L_0) and first-order (L_1) LOESS smoother.

k_M	k_e	Lognormal			Exponential		
		Mean	s_d	r_1	Mean	s_d	r_1
Theory		1.65	2.16	0.378	1.00	1.00	0.500
Data		1.69	2.23	0.377	1.02	1.01	0.500
***L*$_0$**							
50	100	1.68	2.23	0.335	1.00	0.98	0.460
100	100	1.69	2.20	0.345	1.01	0.97	0.481
200	100	1.65	1.99	0.359	0.98	0.98	0.497
400	100	1.70	2.17	0.387	0.95	0.95	0.508
800	100	1.69	2.22	0.359	1.00	0.99	0.489
400	50	1.70	2.25	0.340	0.97	0.97	0.503
400	200	1.70	2.17	0.373	0.95	0.96	0.510
400	400	1.66	2.11	0.356	0.97	0.99	0.497
400	800	1.68	2.19	0.339	0.98	0.98	0.482
400	1600	1.71	2.23	0.359	0.97	1.00	0.462
***L*$_1$**							
50	100	1.68	2.19	0.332	1.04	0.98	0.462
100	100	1.73	2.27	0.368	1.01	0.97	0.478
200	100	1.75	2.24	0.338	1.03	0.97	0.481
400	100	1.67	2.15	0.348	0.97	0.99	0.510
800	100	1.73	2.33	0.350	1.00	0.99	0.494
400	50	1.67	2.14	0.366	0.98	0.97	0.494
400	200	1.68	2.17	0.373	1.00	1.00	0.503
400	400	1.73	2.35	0.356	0.97	0.98	0.509
400	800	1.70	2.26	0.359	0.99	1.00	0.508
400	1600	1.64	2.12	0.357	0.98	1.05	0.505

As is seen in Fig. 1, the use of L_0 considerably reduces the scatter of $\widehat{M}_i$. However, it is also seen that $\widehat{M}_i$ based on the **1**/**k**-kernel or L_0 falls below the theoretical line for $X_{i-1} > 3$. Method L_1, on the contrary, does not underestimate M_i at large X_{i-1}. The mean, s_d and r_1 for methods L_0 and L_1 are listed in Table 2. Especially the reproduction of r_1 improves substantially, compared to the values in Table 1 for the **1**/**k**-kernel. The best results are found for $k_M > 100$. This could be related to the scatter of $\widehat{M}_i$, which decreases with k_M. In the case of the lognormal data a negative bias in r_1 still remains in most simulations, probably induced by the long tail of the distribution. Furthermore, the standard error of the estimated lag 1 autocorrelation is much larger in the lognormal case than in the exponential case, due to the influence of fourth order moments (Bartlett, 1946).

The upper panels of Fig. 2 show probability plots of X_j^*, resulting from resampling the lognormal data (left) and the exponential data (right) with method L_1 using different values of k_M and $k_e = 100$. The lower panels show the corresponding distributions of the simulated means $\widetilde{M}_j$. In the simulations with the lognormal data, the values hardly exceed the highest value in the data, irrespective of the choice of k_M. These simulations are also unable to reproduce the upper 0.5% of the distribution of M. For the exponential data in the upper right panel, the tail of the distribution of X^* shows a better agreement with the theoretical distribution when k_M increases. This can solely be ascribed to the fact that in this case the reproduction of the distribution of M is improved by increasing k_M. The difference between the results for both models may be due to the long tail of the lognormal distribution.

For the simulations in Fig. 2, a relatively small value of k_e was used. The $(\widetilde{M}_j, e_j^*)$-pairs will then generally be close to those in the data. Increasing the value of k_e enhances the simulation of new combinations of M and e and therefore leads to higher simulated values, which is demonstrated in Fig. 3. In the case of the lognormal data the upper tail of the simulated distribution approaches that of the underlying distribution for large k_e. However, in the case of the exponential data the distribution for $k_e = 1600$ clearly overshoots the theoretical distribution. This effect should be ascribed to variance heterogeneity: if k_e is too large, the residuals associated with the set of k_e nearest-neighbours can no longer be considered as identically distributed random variables, due to the dependence of the variance of e on M (Eq. (15)). The effect of variance heterogeneity becomes more pronounced if a uniform kernel is used instead of the **1**/**k**-kernel.

From the foregoing results it is concluded that k_M should be chosen sufficiently large in order to avoid an underestimation of r_1 in the simulations. The value of k_e has a direct influence on the range of simulated values, in particular the highest value. To simulate values that are substantially larger than those in the data, k_e should be chosen sufficiently large. However, too large values of k_e should be avoided, because of possible dependence of the distribution of e on M, which deteriorates the reproduction of the distribution of x. The optimum choice of these parameters depends on the characteristics of the underlying data.

Simulation of precipitation and temperature for the Ourthe catchment

This section compares the performance of different resampling algorithms for the simulation of precipitation and temperature in the catchment of the river Ourthe upstream of Tabreux (1588 km^2), located in eastern Belgium (Ardennes) with altitudes varying between 200 and 650 m above mean sea level. The Ourthe is an important tributary of the river Meuse. Like most of the Belgian Ardennes, the catchment largely consists of steep terrain and a soil of hard, impermeable rock. The storage capacity of this

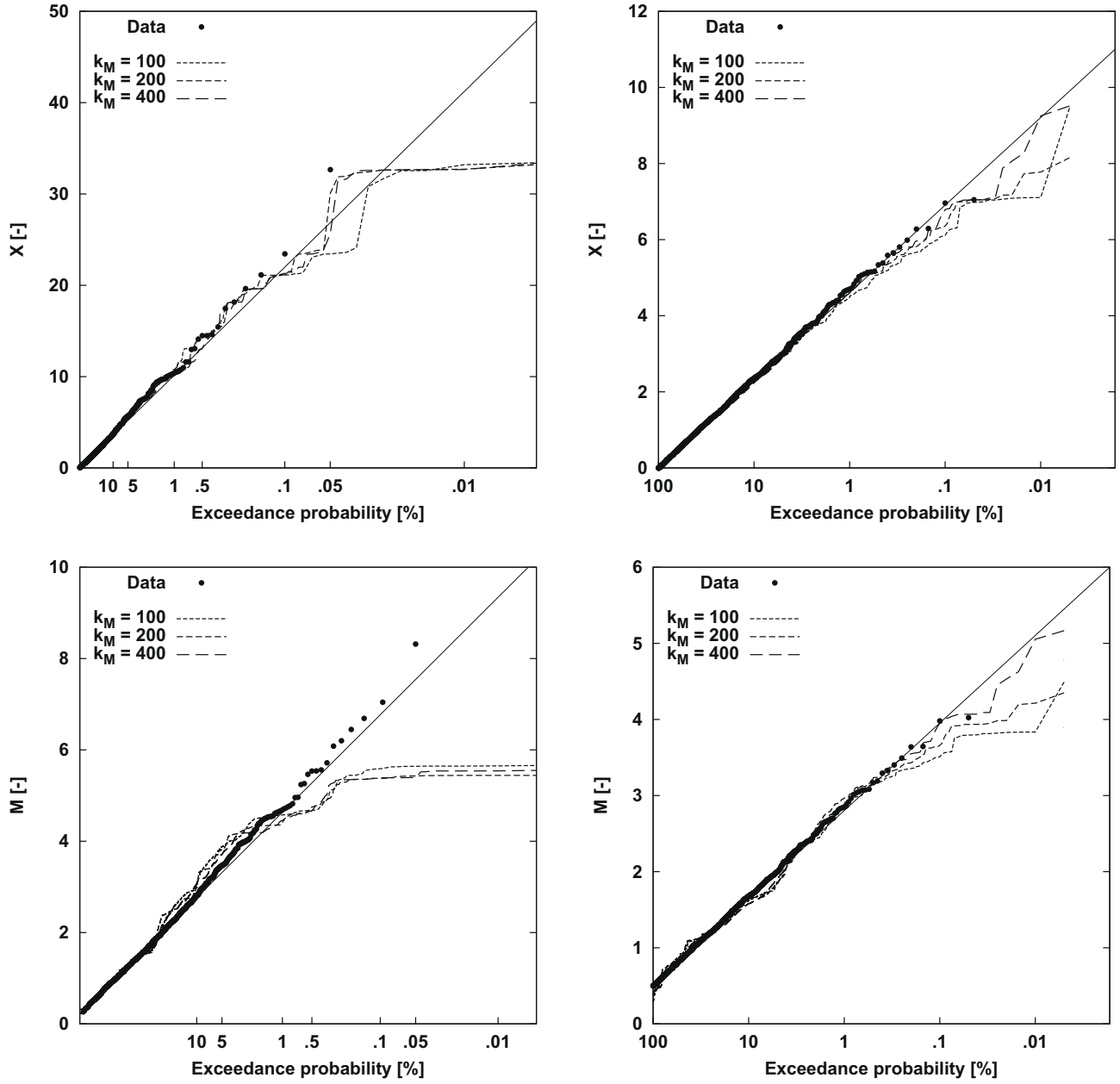

Fig. 2. Probability plots of resampled data X^* and simulated means $\tilde{M}$ for different values of k_M and $k_e = 100$. The upper panels show the probability plots of the data (2000 values, dots) from the lognormal model (left) and the exponential model (right) and the series resampled from these data with method L_1 (20,000 values, curves). The lower panels display the corresponding probability plots of $\tilde{M}$ and those of the empirical distributions (2000 values, dots) calculated from the data through Eqs. (9) and (14) with $\rho = 0.5$. The straight lines represent the theoretical distributions of X in the upper panels and M in the lower panels.

area is therefore low and the response to precipitation relatively fast. Through its position close to the Netherlands border, the Ourthe catchment contributes significantly to the discharge of the river Meuse in the Netherlands during flood waves.

Daily areal average precipitation for this area (derived from station records by the Royal Meteorological Institute of Belgium by means of Thiessen interpolation) and daily temperature from the enclosed station St. Hubert for the 32-year period 1967–1998 were used. The average annual precipitation for this period amounts to 970 mm.

Leander et al. (2005) used nearest-neighbour resampling for the multi-site simulation of precipitation and temperature of the Meuse basin to assess the probability of flood extremes. In that study a three-dimensional feature vector was used to characterize historical and simulated days. The vector of any particular day consisted of the precipitation, the temperature and the precipitation total of the four preceding days. The latter served as a memory element. The similarity of two days t and u is quantified by a weighted Euclidean distance

$$\delta_{tu} = \left[\sum_{j=1}^{3} w_j (v_{tj} - v_{uj})^2\right]^{1/2}, \tag{18}$$

where v_{tj} and v_{uj} denote the jth components of the feature vectors for these days and w_j is a weight, inversely proportional to the variance of the jth component. Mehrotra and Sharma (2006c) compared different methods to determine w_j, which were tested with a variety of data sets. They recommend an optimization of the weights in order to take the relative importance of the components into account. However, the choice and weighting of the feature vector components used here gave satisfactory results in earlier simulations of

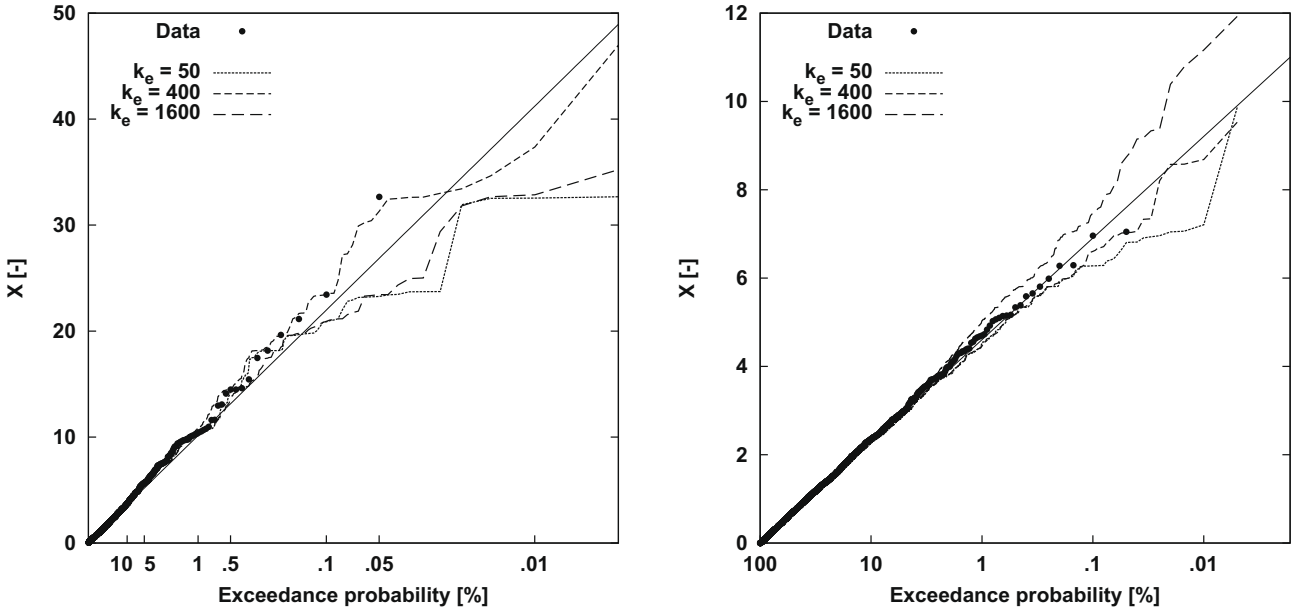

Fig. 3. Probability plots of the data (2000 values, dots) from the lognormal model (left) and the exponential model (right) and the values X^* resampled from these data with method L_1 (20,000 values, curves), using different values of k_e and $k_M = 400$. The straight lines represent the theoretical distributions of X.

precipitation and temperature (Leander et al., 2005), in particular when it comes to simulating multi-day precipitation extremes.

Prior to the simulation, the daily precipitation was deseasonalized with the mean wet-day amount and the daily temperature with the mean and standard deviation. In each step of the simulation, the 10 historical days most similar to the previously simulated day in terms of the weighted Euclidean distance between feature vectors (Eq. (18)) were sought within a moving window of 121 calendar days and one of these historical days was selected at random, using the decreasing kernel in Eq. (4). Its historical successor then provided the precipitation and temperature for the new day in the simulation. Finally, the seasonal cycle of the generated precipitation and temperature sequences was restored. The same resampling algorithm was also used in this study, from here on referred to as 'RG1', with the modification that a 5-day, instead of a 4-day, memory was included in the feature vector.

A method of simulating larger daily precipitation amounts than those observed was introduced by Buishand (2007). In that study, the two largest historical daily amounts in a 60-day season were replaced by a value sampled from a Generalized Pareto Distribution (GPD), whenever they occurred in the resampled sequence. The replacement values were conditioned to exceed the third largest historical value x_3 in the season. The same approach was followed here to perturb a sequence generated with RG1, except that the replacement values were sampled from an exponential distribution fitted to the 10 largest historical amounts in the season of interest. Seasons were defined as bimonthly periods (January–February, March–April, etc.). The used exponential distribution is equivalent to Buishand's GPD with shape parameter $\theta = 0$. The choice of zero shape parameter was justified by a regional analysis of the daily precipitation from different catchments in the Meuse area (Appendix A). From here on, this way of perturbing resampled precipitation from RG1 is referred to as 'RG1p'.

The nearest-neighbour regression algorithm discussed in 'Method' was implemented for the simulation of standardized precipitation and temperature (from here on referred to as 'RG2'). For precipitation the multiplicative model, $X_P = M_P e_P$, was used. The selection of nearest-neighbours of the preceding day, used for estimating both the expected mean standardized precipitation M_P and temperature M_T, was similar to that in RG1. The estimation of M_{Pi} for day i was based on the zeroth-order LOESS smoother L_0 for $X_{Pi-1} < 2$ and on the first-order LOESS smoother L_1 otherwise. In the latter only the daily precipitation amounts for the nearest-neighbours of day $i-1$ and their successors were considered. L_1 was used here in combination with L_0 because precipitation contains zeroes, in contrast with the data of the theoretical examples. Applying L_1 in cases where X_{Pi-1} is small could then lead to a negative value of M_{Pi}. Besides, the linear regression in L_1 is primarily intended to achieve a better estimate of M_{Pi} for large X_{Pi-1}. The standardized temperature was simulated additively, i.e. $X_T = M_T + e_T$, where M_T was estimated with the zeroth-order LOESS smoother L_0. The residuals e^*_{Pj} and e^*_{Tj} for day j in the simulation are both linked to the same historical day. In analogy to the algorithm described in 'Method', the selection of this day should be conditioned on $\widetilde{M}_{Pj}$ and $\widetilde{M}_{Tj}$. However, this resulted in an underestimation of the lag 1 autocorrelation r_1 of daily precipitation. In order to enhance the persistence of the simulated daily precipitation amounts, the last simulated residual e^*_{Pj-1} and its 5-day memory were also taken into account. At the end of each step, X^*_{Pj} was evaluated as $\widetilde{M}_{Pj} e^*_{Pj}$ and X^*_{Tj} as $\widetilde{M}_{Tj} + e^*_{Tj}$. An important contrast between RG1 and RG2 is that in the latter there is no longer a one-to-one correspondence between the simulated values and historical dates.

To assess the sensitivity of RG2 to the neighbourhood sizes, several 320-year sequences of daily precipitation and temperature were simulated with different settings of k_M and k_e. For comparison, also a 320-year simulation was performed with RG1 and RG1p. Only the winter half-year (October–March) was considered, because most floods take place in that season. From all these simulations the mean, standard deviation s_d, lag 1 autocorrelation coefficient r_1 and maximum of the daily precipitation in the winter half-year were compared with those of the 32-year observed record. The results are listed in Table 3. For RG1 and RG1p the mean and s_d show an underestimation, whereas r_1 is close to that observed. There is little difference between both simulations, confirming that the mean, s_d and r_1 are not sensitive to the replacement of the largest values. In the simulations with RG2 the mean is better reproduced than in those with RG1 and RG1p, but most simulations slightly overestimate s_d and underestimate r_1. The latter is comparable to the underestimation of r_1 for the lognormal data in Table 2. In none of the simulations does the bias of

Table 3
Properties of the daily precipitation in the winter half-year (October–March) from 320-year simulations with RG1, RG1p and RG2 with different values of k_M and k_e, compared with those observed (±2 × *se*). The standard deviation s_d, the lag 1 autocorrelation coefficient r_1 and their standard errors *se* were estimated by means of the jackknife method of Buishand and Beersma (1993, 1996). Estimates for the simulations deviating more than 2 × *se* from the historical value are printed in bold. In the last column the largest daily winter amount (Max) is listed.

	k_M	k_e	Mean (mm)	s_d (mm)	r_1	Max (mm)
Obs.	–	–	2.79 ± 0.206	4.59 ± 0.144	0.375 ± 0.033	54.7
RG1	–	–	2.61	4.47	0.375	54.6
RG1p	–	–	2.61	4.48	0.373	66.8
RG2	50	200	2.78	**4.79**	**0.324**	136.5
	100	200	2.70	4.62	**0.330**	64.0
	200	200	2.68	4.62	0.346	62.3
	400	200	2.76	4.71	0.354	67.9
	800	200	2.73	4.69	0.364	62.4
	1600	200	2.76	4.70	0.372	60.1
	200	50	2.67	4.60	0.369	72.5
	200	100	2.72	4.71	0.354	61.5
	200	400	2.69	4.66	0.351	67.3
	200	800	2.77	**4.74**	**0.339**	60.7
	200	1600	2.93	**4.95**	**0.314**	64.3
	400	50	2.73	4.65	0.364	56.5
	400	100	2.71	4.56	0.363	68.5
	400	400	2.74	4.64	0.361	71.9
	400	800	2.76	4.70	0.362	65.3
	400	1600	2.79	**4.85**	**0.341**	76.7
	800	50	2.71	4.63	0.360	55.7
	800	100	2.79	**4.74**	0.365	55.4
	800	400	2.74	4.60	0.371	60.5
	800	800	2.76	**4.76**	0.354	54.8
	800	1600	2.72	**4.78**	0.346	69.4

the simulated mean exceed twice the standard error of the estimate from the historical record. However, in a few simulations the values of s_d and r_1 fall outside their 2 × *se*-intervals. In most of these cases either $k_M \leqslant 100$ or $k_e \geqslant 800$. The value of r_1 generally tends to increase with k_M, in line with the results for the theoretical models. In most RG2-simulations, the largest winter amount is notably larger than that observed.

The distributions of daily precipitation and temperature

With the algorithms RG1, RG1p and RG2 described above, stationary simulations of 12,000 years were performed. These simulations are representative of the historic base period and are not intended to account for climate variability on, for instance, centennial and millennial timescales. For RG2 a value of 400 was selected for k_M and k_e, based on the performance in Table 3 and the experiments with the theoretical models. Setting $k_M = 400$ leads to a satisfactory reproduction of the mean, s_d and r_1 of daily precipitation. Furthermore, it is assumed that $k_e = 400$ is sufficiently large to reasonably approximate the tail of the distribution, while avoiding effects of variance heterogeneity. Fig. 4 compares the simulated distributions of the daily precipitation amounts in the winter half-year with the observed distribution. The simulations show a good agreement with the observations. The figure clearly shows the exponential tail of the distribution in the RG1p-simulation, extending beyond the historically largest amount which limits the values in the RG1-simulation. The plot of RG2 is found roughly in between the plots of RG1 and RG1p, which means that the two-stage resampling algorithm produces a distribution of daily precipitation with a shorter tail than the exponential distribution.

Fig. 5 shows that the RG1- and RG2-simulations reproduce the distribution of the daily temperature in winter quite well. There is only a slight underestimation of the probability of very low temperatures. The highest simulated daily temperatures exceed the highest observed temperature. For the RG1-simulation this can be ascribed to two causes (Buishand and Brandsma, 2001). The values resampled for days in the winter half-year can originate from a historical date outside this season, due to the moving window. Furthermore, differences between the historical and simulated values can arise from the standardization and destandardization.

The distribution of 4-day and 10-day precipitation maxima in winter

From the perspective of flood risks, the extremes of aggregated amounts of simulated precipitation in the winter half-year are of particular interest. The top panel of Fig. 6 compares the Gumbel plots of the 4-day precipitation maxima in the winter half-year for the different resampling algorithms. Up to a return period of 20 years the differences between the simulation methods are minor. The plots for RG1p and RG1 even coincide. For longer return periods, the plots diverge. The highest 4-day maxima are generated by RG2, followed by RG1p. For the Gumbel plots of the 10-day maxima, displayed in the bottom panel of Fig. 6, the differences between the three algorithms are negligible over the entire range of return periods. This is directly related to the exponential tail of the distribution of daily precipitation in the winter half-year. For such a distribution extreme 10-day totals are mainly due to a cluster of moderately large daily amounts, rather than an isolated very large daily amount (Buishand, 2007). There is a good agreement between the Gumbel plots for the three resampling algorithms and the plot for the observed maxima. The main difference is that the three highest observed maxima are somewhat above the plots for the simulated data. In order to investigate the significance of the deviations of these maxima, the RG2-simulation was partitioned into segments with the same length as the observed record, as was done by Leander et al. (2005). A pointwise 95%-envelope of the Gumbel plots of the individual segments was constructed. It was found that the Gumbel plot of observed maxima lies within this envelope.

Extreme river discharges

To assess the effect of the different resampling algorithms on floods, the generated 12,000-year sequences of daily precipitation and temperature were used to drive the rainfall–runoff model HBV

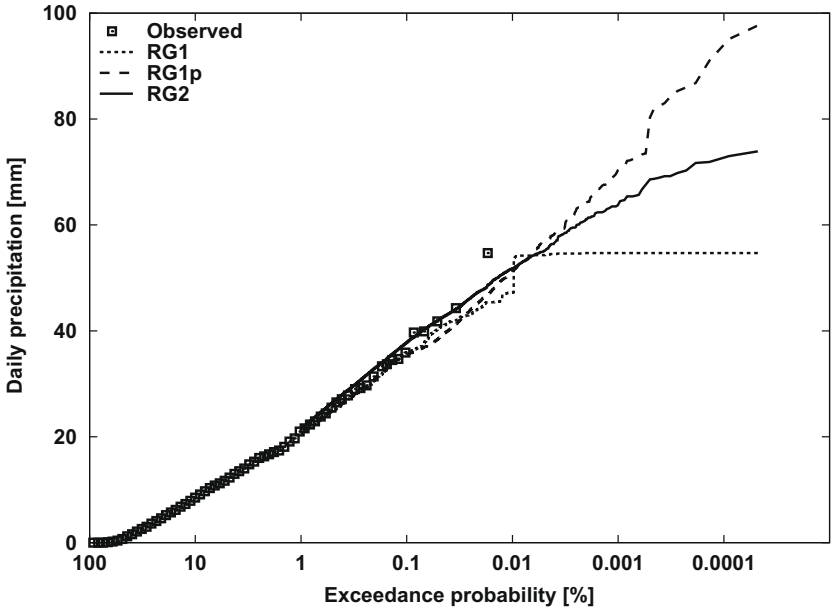

Fig. 4. Exponential probability plots of daily precipitation amounts in the winter half-year for the 12,000-year RG1-, RG1p- and RG2-simulations, compared with the plot of the observed daily precipitation.

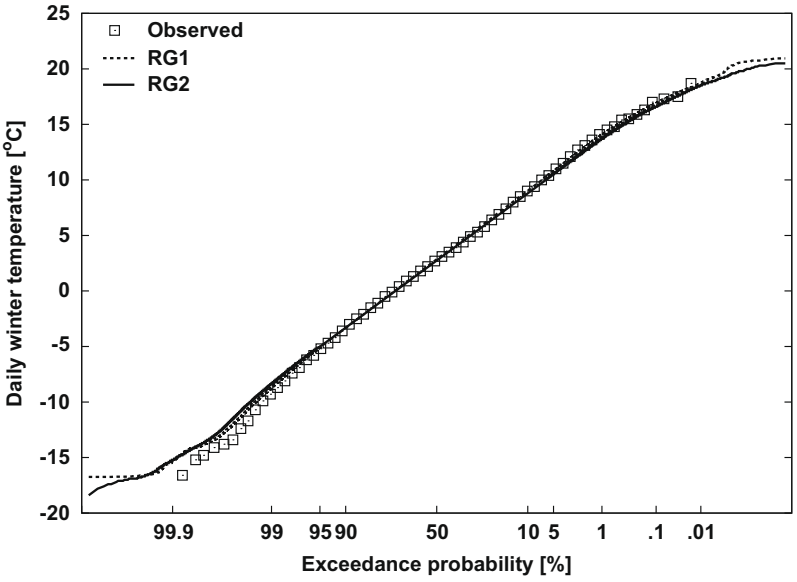

Fig. 5. Normal probability plots of the daily temperature in the winter half-year for the 12,000-year RG1- and RG2-simulations, compared with the plot of the observed daily temperature.

for the river Ourthe at Tabreux. HBV is a conceptual model developed at the Swedish Meteorological and Hydrological Institute (Lindström et al., 1997). The model configuration and the model parameters used here are identical to those for the Ourthe catchment in Leander et al. (2005).

Besides daily precipitation and temperature, the HBV model requires daily potential evapotranspiration (PET) values. These were derived from the daily temperature in the same fashion as in Leander and Buishand (2007). From the simulated daily discharges the maxima of the winter half-years were extracted. Fig. 7 compares the Gumbel plots of these maxima with the plot obtained by driving HBV with observed daily precipitation and temperature (and PET derived from these temperatures). The plots for the generated 12,000-year sequences show a good correspondence with the observed data. Up to a return period of about 200 years there is no distinction between the three simulations. For longer return periods the plots of RG1p and RG2 remain close together and are somewhat above that of RG1. This result points out that being able to simulate daily precipitation amounts beyond the observed range hardly influences the simulation of winter extremes of daily discharge for a tributary of the river Meuse, such as the river Ourthe.

Conclusion and discussion

A two-stage nearest-neighbour algorithm (RG2) is explored that allows for the simulation of daily precipitation amounts and temperatures beyond the range of observed values. The simulation of a new daily value in this algorithm proceeds by the subsequent

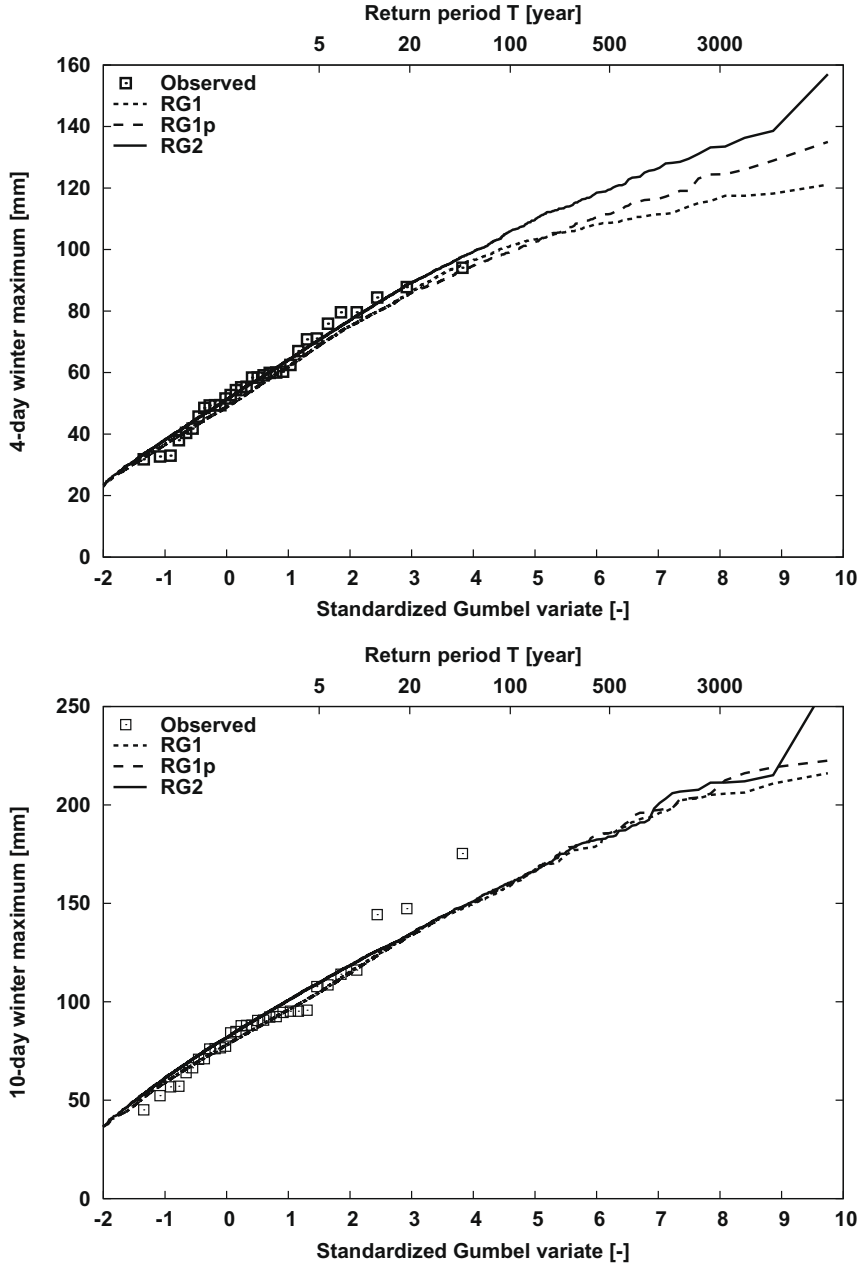

Fig. 6. Gumbel plots of the 4-day (top) and 10-day (bottom) precipitation maxima in the winter half-year for the 12,000-year RG1-, RG1p- and RG2-simulations, compared with those observed.

estimation of the value that is expected to follow the preceding simulated values, and resampling a residual associated with one of the historical observations. Different implementations of this algorithm were tested on data from two AR1 processes. The best results were achieved with an algorithm which estimates the expected value by means of a local linear regression (first-order LOESS smoother). The influence of the neighbourhood size for the determination of the expected values, k_M, and that for the sampling of the residuals, k_e, was studied in detail. It was found that k_M should be chosen sufficiently large in order to achieve a satisfactory reproduction of the lag 1 autocorrelation. Furthermore, a small value of k_e limits the potential to simulate larger values than observed. On the other hand, a very large value of k_e may worsen the distribution of the simulated values due to variance heterogeneity.

The two-stage algorithm was further used to generate sequences of daily precipitation and temperature for the Ourthe catchment. Though an underestimation of the lag 1 autocorrelation and a slight overestimation of the standard deviation are found for most settings of the neighbourhood sizes, the bias was only significant in a few cases associated with a high value of k_e or a small value of k_M. With a suitable choice of k_M and k_e, a 12,000-year simulation was conducted with RG2. For comparison an additional simulation of the same length was conducted with the conventional algorithm RG1. A modified version RG1p of this simulation was created by perturbing the highest resampled values in each bimonthly season, in such a way that the distribution of the daily precipitation amounts was extended with an exponential tail. For all three simulations the distributions of the 4-day and 10-day winter maxima were in agreement with the observed data. At return periods longer than 20 years, differences were seen between the Gumbel plots of the 4-day maxima, showing the influence of the larger daily amounts. The plots of the 10-day maxima, however, barely differ. The simulated data were used to drive the HBV rainfall–runoff model for the Ourthe catchment. It was found

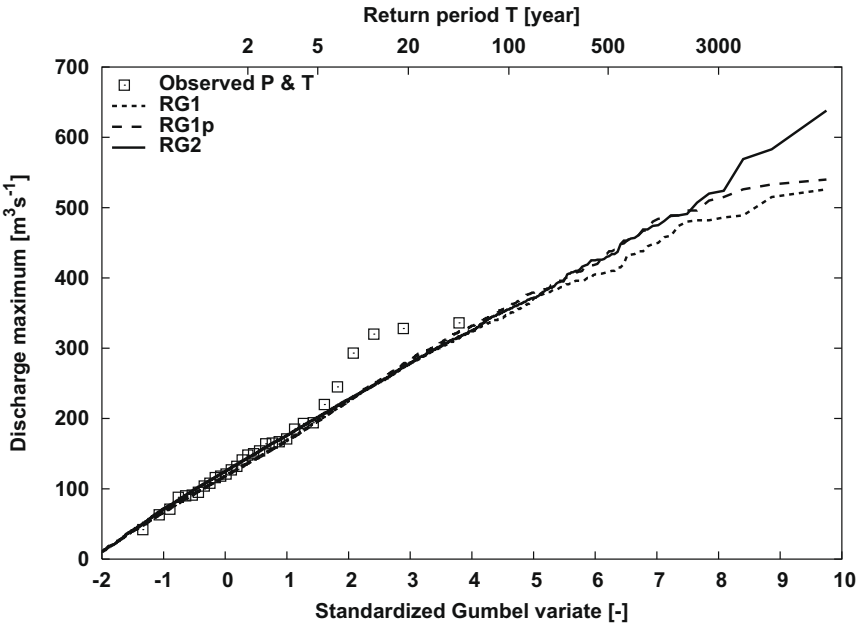

Fig. 7. Gumbel plots of the winter maxima of daily discharge for the river Ourthe at Tabreux from the 12,000-year RG1-, RG1p- and RG2-simulations compared with the plot based on observed daily precipitation and temperature.

that the larger daily precipitation amounts produced by RG1p and RG2 had no discernible effect on the distribution of the winter maxima of daily discharge.

A disadvantage of RG2 is that it is very time-consuming, because a large number of nearest-neighbours has to be sorted twice. A great saving in computer time can be achieved by using the same nearest-neighbours for the estimation of the expected value and the resampling of residuals as in Prairie et al. (2006). However, for the Ourthe data this leads to a substantial positive bias in the standard deviation and a negative bias in the autocorrelation. A possible explanation is that conditioning on the expected precipitation and temperature of the new day yields a more homogeneous set of residuals than conditioning on the characteristics of the previously simulated day.

The RG1 algorithm has been extended to perform multi-site simulations (Buishand and Brandsma, 2001). A similar extension is possible for the RG2 algorithm. There is no need to generate from a multivariate normal distribution as in multi-site versions of the nonparametric weather generator based on kernel-density estimation techniques (Mehrotra and Sharma, 2006a,b, 2007), which involves quite strong assumptions on the spatial dependence structure. A multi-site extension of the RG1p algorithm also needs assumptions on spatial dependence and is not straightforward. The extension of RG2 to multi-site generation is less complicated and requires less assumptions about the data.

There are also possibilities to use the RG2 algorithm in climate-change studies. One is to condition the resampling of daily precipitation and temperature on the large-scale atmospheric circulation. For instance, Beersma and Buishand (2003) performed resampling simulations conditional on three circulation indices. Their algorithm was essentially identical to RG1. The simulated daily values in the case of an altered circulation are then still restricted to the range of the observed values. Due to the additional sampling of residuals, RG2 does not suffer from this restriction. In contrast with the unconditional RG1 simulations in this study, the conditional simulations of Beersma and Buishand (2003) show a significant bias in the lag 1 autocorrelation coefficient of the simulated daily precipitation amounts. This bias might vanish in the RG2 simulations due to the use of the LOESS smoother. By resampling residuals the variance of the predictant in a downscaling relation can be maintained, something which is often necessary in climate-change impact studies (von Storch, 1999). The method used for resampling the residuals in the RG2 algorithm is of particular interest if negative values are to be avoided. Aside from conditioning on large-scale atmospheric variables, resampling can also be applied directly to daily precipitation and temperature from regional climate model runs under future climate conditions, as was demonstrated by Leander et al. (2008). The direct use of climate model data takes all characteristics of future climate change into account. However, bias corrections are needed because of systematic differences between the regional climate model output and the observed data. Anticipating the steady improvement of climate models, the direct use of model data will be of growing interest.

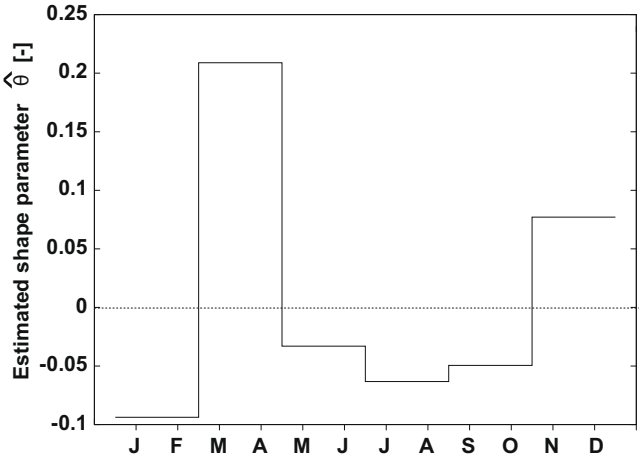

Fig. A1. Regional estimates of the GEV shape parameter θ of bimonthly maxima of daily precipitation for 15 subcatchments of the river Meuse.

Acknowledgements

The work was performed in co-operation with the Institute of Inland Water Management and Waste Water Treatment (RIZA),

Lelystad. The areal precipitation of the Ourthe basin and the temperature of St. Hubert were kindly provided by the Royal Meteorological Institute of Belgium. The authors thank Bart van den Hurk for comments on an earlier version of the paper.

Appendix A. Determination of the shape parameter θ

The perturbation of the RG1-simulation (RG1p) requires that the shape of the upper tail of the distribution of daily precipitation is known. To find a suitable representation for the tail, the bimonthly maxima of the daily precipitation amounts for 15 subcatchments of the river Meuse (upstream of the Netherlands) were analysed. It is assumed that the largest amount X in a bimonthly period follows a Generalized Extreme Value (GEV) distribution:

$$\Pr(X \leqslant x) = \exp\left[-\left(1-\theta\frac{x-\mu}{\sigma}\right)^{1/\theta}\right]. \tag{A.1}$$

The location parameter μ and the scale parameter σ vary over the year and over the subcatchments. A common shape parameter θ is assumed for all subcatchments, which varies over the year. The regional L-moments approach (Hosking and Wallis, 1997) was followed to estimate the common θ.

First, the probability weighted moments b_0, b_1 and b_2 were determined from the ordered maxima $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$ as (Landwehr et al., 1979)

$$b_0 = \sum_{j=1}^{n}\frac{x_j}{n}, \quad b_1 = \sum_{j=2}^{n}\frac{(j-1)x_j}{n(n-1)}, \quad b_2 = \sum_{j=3}^{n}\frac{(j-1)(j-2)x_j}{n(n-1)(n-2)}. \tag{A.2}$$

The sample L-moments ℓ_2 and ℓ_3 were derived using (Hosking and Wallis, 1997, p. 26)

$$\ell_2 = 2b_1 - b_0, \quad \ell_3 = 6b_2 - 6b_1 + b_0. \tag{A.3}$$

For each subcatchment, the sample L-skewness $t_3 = \ell_3/\ell_2$ was determined and then averaged over the catchment. The estimated shape parameter $\hat{\theta}$ of the GEV distribution then follows from the average sample skewness t_3^R as (Hosking and Wallis, 1997, p. 196):

$$\hat{\theta} = 2.9554c^2 + 7.8590c \quad \text{with } c = \frac{2}{3+t_3^R} - \frac{\log 2}{\log 3}. \tag{A.4}$$

Fig. A1 shows the bimonthly values of $\hat{\theta}$ calculated from the data for the 15 subcatchments. In November–December a positive $\hat{\theta}$ is found, whereas in January–February the value is negative. In March–April, again, a positive value of 0.21 is seen. Averaging $\hat{\theta}$ over the winter half-year results in a value of about 0.02. Given the fact that this is close to zero, the distribution of the bimonthly maxima of daily precipitation is close to the Gumbel distribution:

$$\Pr(X \leqslant x) = \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] \tag{A.5}$$

in this season, which implies that the exceedances of a high threshold are approximately exponentially distributed. For the summer half-year there is a tendency towards a negative shape parameter. Thence, for this season the exponential distribution underestimates the tail of the distributions. This is, however, not relevant for the simulation of extreme discharges in winter.

References

Bárdossy, A., Plate, E.J., 1992. Space-time model for daily rainfall using atmospheric circulation patterns. Water Resources Research 28, 1247–1259.

Bartlett, M.S., 1946. On the theoretical specification and sampling properties of autocorrelated time series. Journal of the Royal Statistical Society, Series B 8, 27–41.

Beersma, J.J., Buishand, T.A., 2003. Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. Climate Research 25, 121–133.

Buishand, T.A., 2007. Estimation of a large quantile of the distribution of multi-day seasonal maximum rainfall: the value of stochastic simulation of long duration sequences. Climate Research 34, 185–194.

Buishand, T.A., Beersma, J.J., 1993. Jackknife tests for differences in autocorrelation between climate time series. Journal of Climate 6, 2490–2495.

Buishand, T.A., Beersma, J.J., 1996. Statistical tests for comparison of daily variability in observed and simulated climates. Journal of Climate 9, 2538–2550.

Buishand, T.A., Brandsma, T., 2001. Multi-site simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbour resampling. Water Resources Research 37, 2761–2776.

Charles, S.P., Bates, B.C., Hughes, J.P., 1999. A spatiotemporal model for downscaling precipitation occurrence and amounts. Journal of Geophysical Research 104, 31657–31669.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829–836.

Fernandez, B., Salas, J.D., 1986. Periodic gamma autoregressive processes for operational hydrology. Water Resources Research 22, 1385–1396.

Gaver, D.P., Lewis, P.A.W., 1980. First-order autoregressive gamma sequences and point processes. Advances in Applied Probability 12, 727–745.

Harrold, T.I., Sharma, A., Sheater, S.J., 2003a. A nonparametric model for stochastic generation of daily rainfall occurrence. Water Resources Research 39. doi:10.1029/2003WR002182.

Harrold, T.I., Sharma, A., Sheater, S.J., 2003b. A nonparametric model for stochastic generation of daily rainfall amounts. Water Resources Research 39. doi:10.1029/2003WR002570.

Hosking, J.R.M., Wallis, J.R., 1997. Regional Frequency Analysis: An Approach Based on L-moments. Cambridge University Press, Cambridge, UK.

Hughes, J.P., Guttorp, P., 1994. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resources Research 30, 1535–1546.

Lall, U., Sharma, A., 1996. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resources Research 32, 679–693.

Landwehr, J.M., Matalas, N.C., Wallis, J.R., 1979. Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. Water Resources Research 15, 1055–1064.

Lawrance, A.J., 1980. Some autoregressive models for point processes. In: Bartfai, P., Tomko, J. (Eds.), Proceedings Colloquia Mathematica Societatis János Bolyai. Elsevier, Amsterdam.

Leander, R., Buishand, T.A., Aalders, P., de Wit, M.J.M., 2005. Estimation of extreme floods of the river Meuse using a stochastic weather generator and a rainfall–runoff model. Hydrological Sciences Journal 50, 1089–1103.

Leander, R., Buishand, T.A., 2007. Resampling of regional climate model output for the simulation of extreme river flows. Journal of Hydrology 332, 487–496.

Leander, R., Buishand, T.A., van den Hurk, B.J.J.M., de Wit, M.J.M., 2008. Estimated changes in flood quantiles of the river Meuse from resampling of regional climate model output. Journal of Hydrology 351, 331–343.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. Journal of Hydrology 201, 272–288.

McKenzie, E., 1982. Product autoregression: a time-series characterization of the gamma distribution. Journal of Applied Probability 19, 463–468.

Mehrotra, R., Sharma, A., 2006a. A nonparametric stochastic downscaling framework for daily rainfall at multiple locations. Journal of Geophysical Research 111, D15101. doi:10.1029/2005JD006637.

Mehrotra, R., Sharma, A., 2006b. A semi-parametric model for stochastic generation of multi-site daily rainfall exhibiting low frequency variability. Journal of Hydrology 335, 180–193.

Mehrotra, R., Sharma, A., 2006c. Conditional resampling of hydrologic time series using multiple predictor variables: a K-nearest neighbour approach. Advances in Water Resources 29, 987–999.

Mehrotra, R., Sharma, A., 2007. Preserving low-frequency variability in generated daily rainfall sequences. Journal of Hydrology 345, 102–120.

Mejía, J.M., Rodríguez-Iturbe, I., 1974. Correlation links between normal and log-normal processes. Water Resources Research 10, 689–691.

Prairie, J.R., Rajagopalan, B., Fulp, T.J., Zagona, E.A., 2006. Modified k-nn model for stochastic streamflow simulation. Journal of Hydrologic Engineering 11, 371–378.

Rajagopalan, B., Lall, U., 1999. A k-nearest-neighbor simulator for daily precipitation and other variables. Water Resources Research 35, 3089–3101.

Richardson, C.W., 1977. A Model of Stochastic Structure of Daily Precipitation Over an Area. Hydrology Paper 91. Colorado State University, Fort Collins, Colorado.

von Storch, H., 1999. On the use of "inflation" in statistical downscaling. Journal of Climate 12, 3505–3506.

Weiss, G., 1977. Shot noise models for the generation of synthetic streamflow data. Water Resources Research 13, 101–108.

Wilks, D.S., 1998. Multisite generalization of a daily stochastic precipitation generation model. Journal of Hydrology 210, 178–191.

Woolhiser, D.A., 1992. Modeling daily precipitation-progress and problems. In: Walden, A.T., Guttorp, P. (Eds.), Statistics in Environmental and Earth Sciences. Edward Arnold, London.

Young, K.C., 1994. A multivariate chain model for simulating climatic parameters from daily data. Journal of Applied Meteorology 33, 661–671.