Rainfall generator for the Rhine basin: multi-site simulation of daily weather variables by nearest-neighbour resampling

Jules J. Beersma, T. Adri Buishand and Rafał Wójcik

Royal Netherlands Meteorological Institute (KMNI)

Abstract

Nearest-neighbour resampling is used here for the joint simulation of daily rainfall and temperature at 36 stations in Germany, Luxemburg, France and Switzerland all situated in the Rhine basin. The daily temperatures are used to determine snow accumulation and melt in winter. A major advantage of a non-parametric resampling technique is that it preserves both the spatial association of daily rainfall over the drainage basin and the dependence between daily rainfall and temperature without making assumptions about the underlying joint distributions. Both unconditional simulation of daily rainfall and temperature and conditional simulations of these variables on the atmospheric flow are discussed. In particular the unconditional simulations reproduce the standard deviations and autocorrelation coefficients and properties of extreme 10-day rainfall and snowmelt well. The largest 10-day rainfall amounts in 1000-year simulations are up to 40% larger than those in the historical record (1961-1995).

1 Introduction

The Rhine is the most important river in the Netherlands. Large parts of the river, which originates in the Swiss Alps, are situated in Switzerland, Germany, France and the Netherlands. Small parts of Austria, Belgium and almost the whole country of Luxemburg also drain to the river. Protection against flooding is a point of continuous concern. According to safety standards, laid down in the Flood Protection Act, measures against flooding in the non-tidal part of the Rhine in the Netherlands have to withstand a discharge that is exceeded on average once in 1250 years. Traditionally this design discharge has been obtained from a statistical analysis of peak discharges at Lobith, where the river enters the country. Several probability distributions have been fitted to the discharge maxima of that record. The long return period requires an extrapolation far beyond the length of the observed discharge record. Different distributions then usually lead to quite different design discharges.

During a re-evaluation of the design discharge at Lobith, there was a strong feeling that the uncertainties of extrapolation could be reduced by taking the physical properties of the river basin into account. It was suggested to develop a hydrological/hydraulic model for the whole basin. The development of a stochastic rainfall generator was also required in order to produce long-duration rainfall series over the basin. The use of synthetic rainfall series in combination with a hydrological/hydraulic model does not only provide the peak discharges but also the durations of these extreme events. This may lead to a better insight into the profile of the design flood.

In this paper nearest-neighbour resampling models are considered for simulation of multi-site daily precipitation and temperature time series in the Rhine basin. Temperature is needed to account for the effects of snow(melt) and frozen soils on large river discharges. The reproduction of second-order moment statistics of temperature and precipitation and properties of extreme winter precipitation and snowmelt are examined. Results of 1000-year simulations with these models are presented. In these simulations much larger multi-day precipitation maxima occur than the historical ones. More details, including several alternative nearest-neighbour resampling models, can be found in Wójcik *et al.* (2000).

2 Nearest-neighbour resampling

Nearest-neighbour resampling was originally proposed by Young (1994) to simulate daily minimum and maximum temperatures and precipitation. Independently, Lall and Sharma (1996) discussed a nearest-neighbour bootstrap to generate hydrological time series. Rajagopalan and Lall (1999) presented an application to daily precipitation and five other weather variables. Basically the same method is used for generating daily precipitation and temperature in the Rhine basin. Especially for multi-site simulations summary statistics are needed to avoid problems with the high dimensional data space (Buishand and Brandsma, 2000).

In the nearest-neighbour method weather variables like precipitation and temperature are sampled simultaneously with replacement from the historical data. To incorporate autocorrelation, resampling depends on the simulated values for the previous day in the works of Young (1994) and Rajagopalan and Lall (1999). Therefore, one first searches the days in the historical record that have similar characteristics as those of the previously simulated day. One of these nearest neighbours is randomly selected and the observed values for the day subsequent to that nearest neighbour are adopted as the simulated values for the next day t. A feature vector (or state vector) $\mathbf{D}_t$ is used to find the nearest neighbours in the historical record. In Rajagopalan and Lall (1999) $\mathbf{D}_t$ was formed out of the standardized weather variables generated for day t - 1. The k nearest neighbours of $\mathbf{D}_t$ were selected in terms of a weighted Euclidean distance. For two q-dimensional vectors $\mathbf{D}_t$ and $\mathbf{D}_u$ the latter is defined as:

$$\delta(\mathbf{D}_t, \mathbf{D}_u) = \left(\sum_{j=1}^{q} w_j \left(v_{tj} - v_{uj} \right)^2 \right)^{1/2} \tag{1}$$

where v_{tj} and v_{uj} are the jth components of $\mathbf{D}_t$ and $\mathbf{D}_u$ respectively and the w_j's are scaling weights.

A discrete probability distribution or kernel is required to select one of the k nearest neighbours. Lall and Sharma (1996) recommended a kernel that gives higher weight to the closer neighbours. For this decreasing kernel the probability p_n that the nth closest neighbour is resampled is given by:

$$p_n = \frac{1/n}{\sum_{i=1}^{k} 1/i}, \qquad n = 1, \ldots, k \tag{2}$$

From the above description it is clear that apart from creating a feature vector (see Section 5), the user has to set the values of the number k of nearest neighbours and the weights w_j. A sensitivity analysis in Brandsma and Buishand (1999) showed that k = 5 usually works well. In this study we also use this value of k. A more difficult issue is the choice of the weights w_j. Tuning the weights can be very time consuming especially if the dimension of the feature vector is high. In Wójcik *et al.* (2000) an alternative approach is introduced that avoids specification of the weights by using the *Mahalanobis distance*.

3 Data description

Daily temperature and precipitation data from 36 stations were used. The stations are distributed all over the Rhine basin: 25 stations in Germany, 1 station in Luxemburg, 4 stations in France and 6 stations in Switzerland (see Fig. 1). For the 35-year study period (1961-1995) the data were provided by the "International Commission for the Hydrology of the Rhine Basin".

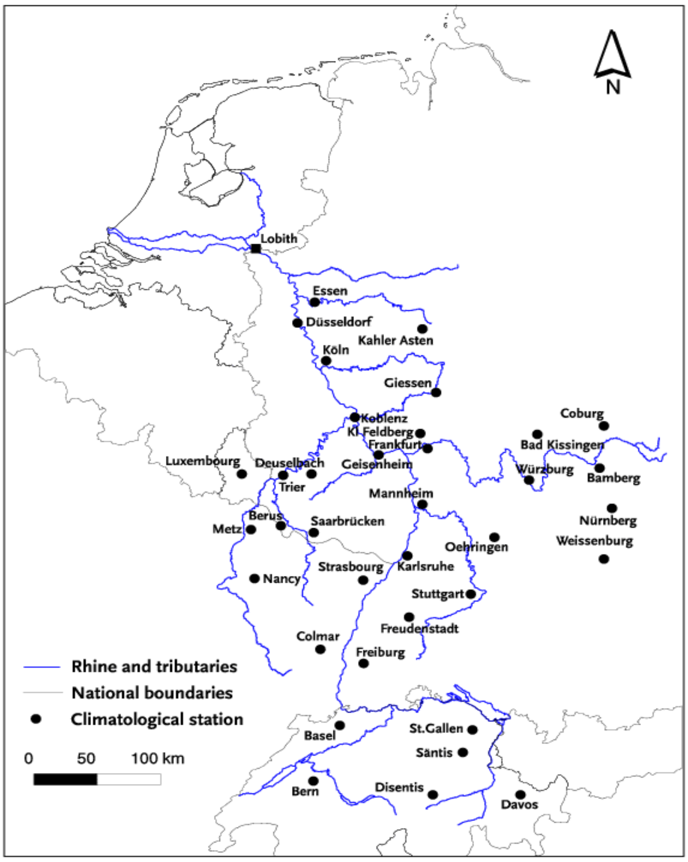

Figure 1: Location of Lobith in the Netherlands and the 36 stations in the drainage area of the river Rhine used in this study.

Most stations in Germany, Luxemburg and France are lowland stations with annual mean rainfall ranging from 500 to 900 mm. However, two stations in Germany, Kahler Asten and Freudenstadt, have a much larger annual mean rainfall (≈1500 mm). For the Swiss stations mean annual rainfall ranges from about 800 mm for Basel to almost 2400 mm for Säntis. The latter is an exceptional station lying at an altitude of 2500 m.

Because precipitation P and temperature T depend on the atmospheric flow, three daily circulation indices are also considered: (i) relative vorticity Z, (ii) strength of the westerly flow W and (iii) strength of the southerly flow S. These circulation indices were computed from daily mean sea-level pressure data on a regular 5° latitude and 10° longitude grid. The derivation of the circulation indices is similar to that in Jones *et al*. (1993), except that the grid was centered at the Rhine basin instead of the British Isles.

4 Standardization procedure

Before resampling the data were deseasonalized through standardization. The daily temperatures and circulation indices were standardized by subtracting an estimate m_d of the mean and dividing by an estimate s_d of the standard deviation for the calendar day d of interest:

$$\tilde{x}_t = (x_t - m_d)/s_d, \qquad t = 1,\ldots,365J; \quad d = (t-1) \bmod 365 + 1 \tag{3}$$

where x_t and $\tilde{x}_t$ are the original and standardized variables for day t, respectively, and J is the total number of years in the record. The estimates m_d and s_d were obtained by smoothing the sample mean and standard deviation of the successive calendar days.

Daily precipitation was standardized by dividing by a smooth estimate $m_{d,wet}$ of the mean wet-day precipitation amount:

$$\tilde{x}_t = x_t / m_{d,wet}, \qquad t = 1,\ldots,365J; \quad d = (t-1) \bmod 365 + 1 \tag{4}$$

A wet day was defined here as a day with $P \geq 0.1$ mm.

To reduce the effect of seasonal variation further, the search for nearest neighbours was restricted to days within a moving window, centered on the calendar day of interest. The width of this window was 61 days as in Brandsma and Buishand (1999).

5 Model identification

5.1 The feature vector

Daily P and T observations were available for the 36 climatological stations in Fig.1 Because of their rather extreme weather characteristics, the two Swiss mountain stations Davos and Säntis are not included in the feature vector. It is, however, still possible to simulate values for these stations passively (i.e. the simulated values for the passive stations have equal historical dates as those simulated for the stations used in the feature vector). To keep the dimension of the feature vector low, a small number of summary statistics was calculated for the remaining 34 stations. Both for P and T the arithmetic mean of the standardized daily values was used. In addition, the fraction F of stations with $P \geq 0.1$ mm was considered. F helps to distinguish between large-scale and convective precipitation. To keep the notation compact, the above components of the feature vector will be referred to as a sub-vector $\mathbf{V} = [\tilde{P}, F, \tilde{T}]^T$ where the tilde indicates standardized values. In some cases, the feature vector also contains the standardized circulation indices $\tilde{\mathbf{C}} = [\tilde{Z}, \tilde{W}, \tilde{S}]^T$.

5.2 Three simulation models

Basically two different kinds of simulations can be distinguished: *unconditional* simulations and *conditional* simulations on the atmospheric flow indices.

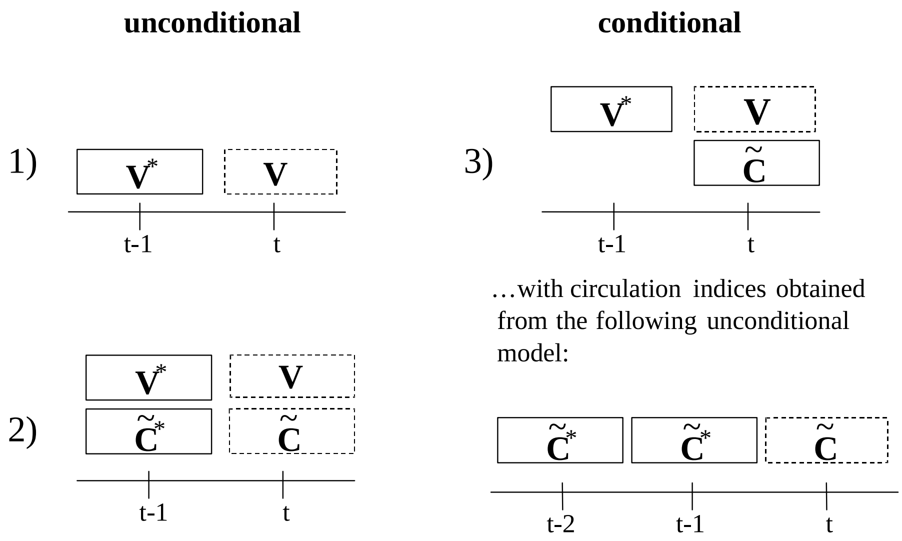

Figure 2: Components of the feature vector (solid boxes) for unconditional simulations 1), 2) and conditional simulation 3). The dashed boxes relate to variables to be resampled. The asterisks indicate that the corresponding variables are resampled values of the previous time steps.

In the unconditional simulations the feature $\mathbf{D}_t$ comprises generated variables for the previous day as shown in Fig. 2 (cases 1 and 2). Conditional simulation on the atmospheric flow requires that the circulation indices for day t are included in the feature vector as schematically represented in the upper panel of case 3 in Fig. 2. The conditional model presented here uses *simulated* circulation indices obtained with a second-order model (circ2.5) described in Beersma and Buishand (1999) (see Fig. 2 lower panel of case 3). Conditional nearest-neighbour resampling is closely related to the analogue method used in climate change studies (Zorita and von Storch, 1999). Details of the three models are given in Table 1.

Table 1: Definition of models for unconditional and conditional simulation. The weights for the circulation indices apply to all three components of $\tilde{\mathbf{C}}$. $\tilde{P}$ and $\tilde{T}$ denote respectively the standardized precipitation and temperature averaged over 34 stations, and F denotes the fraction of these stations with $P \geq 0.1$ mm. An asterisk indicates that a value was resampled in a previous time step.

Model	Elements of $\mathbf{D}_t$	Weights
unconditional		
UE	$\tilde{P}^*_{t-1}, F^*_{t-1}, \tilde{T}^*_{t-1}$	2,4,1
UEc	$\tilde{\mathbf{C}}^*_{t-1}, \tilde{P}^*_{t-1}, F^*_{t-1}, \tilde{T}^*_{t-1}$	**1**,3,5,2
conditional		
CE	$\tilde{\mathbf{C}}_t, \tilde{P}^*_{t-1}, F^*_{t-1}, \tilde{T}^*_{t-1}$	**1**,3,5,2

6 Reproduction of standard deviations and autocorrelation

Extreme river discharges in the lower part of the Rhine basin are mostly caused by prolonged heavy rainfall in winter. The reproduction of the standard deviations of daily temperature and precipitation, the standard deviations of the monthly average temperature and the monthly precipitation totals, and the autocorrelation coefficients is therefore only presented for the winter half-year (October - March). To reduce the influence of the annual cycle these statistics were first calculated for each calendar month separately. For each of the 34 stations the winter estimates were obtained by taking the arithmetic mean of the six winter months (October, ..., March).

Twenty-eight runs of 35 years were generated to investigate the performance of the resampling procedure. For each station *i*, the standard deviations and autocorrelation coefficients were first estimated for each simulation run separately and then averaged over the 28 runs. The average estimates $\bar{s}_{\mathrm{D}_i}^*$, $\bar{s}_{\mathrm{M}_i}^*$, $\bar{r}_i^*(l)$ for the daily and monthly standard deviations and the lag *l* autocorrelation coefficient respectively, were compared with the estimates $\bar{s}_{\mathrm{D}_i}$, $\bar{s}_{\mathrm{M}_i}$, $\bar{r}_i\,(l)$ for the historical data. The average relative difference $\langle \Delta \bar{s}_{\mathrm{D}} \rangle$ between the observed and simulated daily standard deviation is calculated using

$$\langle \Delta \bar{s}_{\mathrm{D}} \rangle = 1/34 \sum_{i=1}^{34} (\bar{s}_{\mathrm{D}_i}^* - \bar{s}_{\mathrm{D}_i}) / \bar{s}_{\mathrm{D}_i}\ 100\% \tag{5}$$

with a similar equation for the average relative difference $\langle \Delta \bar{s}_{\mathrm{M}} \rangle$ of the monthly standard deviation, and

$$\langle \Delta \bar{r}(l) \rangle = 1/34 \sum_{i=1}^{34} [\bar{r}_i^*(l) - \bar{r}_i(l)] \tag{6}$$

for the average difference $\langle \Delta \bar{r}(l) \rangle$ of the lag *l* autocorrelation coefficient.

In order to evaluate the statistical significance of $\langle \Delta \bar{s}_{\mathrm{D}} \rangle$, $\langle \Delta \bar{s}_{\mathrm{M}} \rangle$ and $\langle \Delta \bar{r}(l) \rangle$ standard errors *se* were calculated for the historical record. A criterion of 2 × *se* was used to indicate significant differences between the historical and simulated values. Table 2 presents $\langle \Delta \bar{s}_{\mathrm{D}} \rangle$, $\langle \Delta \bar{s}_{\mathrm{M}} \rangle$ and $\langle \Delta \bar{r}(l) \rangle$ for the models defined in Table 1.

Table 2: Percentage differences between the mean standard deviations of monthly and daily values, and absolute differences between the mean lag 1 and 2 autocorrelation coefficients in the simulated time series (twenty-eight runs of 35 years) and the historical records (1961-1995), averaged over 34 stations. Bottom lines: average historical estimates (standard deviations in mm for precipitation and in °C for temperature). Values in bold refer to differences more than 2 × *se* from the historical estimate.

	$\langle \Delta \bar{s}_{\mathrm{M}} \rangle$		$\langle \Delta \bar{s}_{\mathrm{D}} \rangle$		$\langle \Delta \bar{r}(1) \rangle$		$\langle \Delta \bar{r}(2) \rangle$	
Model	*P*	*T*	*P*	*T*	*P*	*T*	*P*	*T*
UE	0.3	-1.1	0.2	0.2	**-0.019**	**-0.032**	-0.001	0.006
UEc	-1.7	-8.2	-1.2	-1.9	**-0.018**	**-0.036**	0.001	-0.020
CE	-6.4	**-18.8**	-2.3	**-7.0**	**-0.052**	**-0.087**	**-0.022**	**-0.050**
Historical	35.7	2.1	4.2	4.2	**0.283**	**0.826**	0.144	0.639
se	4.53	6.16	2.45	2.49	**0.008**	**0.007**	0.009	0.015

For the unconditional model which incorporates only the large-scale features of the *P* and *T* fields (model UE) the precipitation and temperature statistics are well reproduced. A slight, though statistically significant, bias is present in the lag 1 autocorrelation coefficients. Incorporation of the circulation indices into the feature vector (model UEc) generally worsens the reproduction of daily temperature statistics. The results for precipitation are, however, similar to those obtained in the unconditional model without circulation indices. The latter insensitivity is in line with the results in Buishand and Brandsma (2000).

Conditional resampling of *P* and *T* on simulated circulation indices (model CE) lags behind. In Beersma and Buishand (1999) it was also shown that this occurs for simulations conditional on historical circulation indices. All temperature statistics and the lag 1 and 2 autocorrelation coefficients for precipitation are significantly underestimated.

7 Reproduction of 10-day winter maximum precipitation

Three quantities are considered to verify the reproduction of the 10-day winter maximum precipitation amounts: (i) the maximum *MAX* of the 10-day winter maxima (highest 10-day precipitation amount in the record), (ii) the upper quintile mean *QM*5 of the 10-day winter maxima and (iii) the median *M* of the 10-day winter maxima. *QM*5 refers to the mean of the data beyond the highest quintile (upper 20%).

Analogous to equation (5), we calculated for each of the three quantities the percentage differences between the values for the simulated and historical data averaged over the 34 stations. Table 3 presents the results for the three models.

There is always an underestimation of the extreme-value properties, which is, however, not more than 2% for the unconditional models. Conditioning the resampling procedure on circulation indices (model CE), results in a bit larger underestimation of the extreme-value statistics than in the unconditional cases. This is in agreement with the poorer reproduction of second-order moment statistics for conditional simulations as observed in Table 2.

Table 3: Percentage differences between the maxima (*MAX*), upper quintile means (*QM*5) and medians (*M*) of the 10-day winter (October-March) precipitation maxima in the simulated data (twenty-eight runs of 35 years) and the historical records (1961-1995), averaged over 34 stations. Bottom line: average historical estimates (mm).

Model	MAX (%)	QM5 (%)	M (%)
UE	-1.4	-0.6	-0.2
UEc	-0.9	-2.0	-1.9
CE	-5.5	-5.3	-4.5
Historical	138.5	111.1	75.2

8 Reproduction of 10-day maximum snowmelt amounts

Snowmelt generally, contributes to extreme river discharges in the lower part of the Rhine basin. It is, however, only for the highest stations Kahler Asten, Freudenstadt, Kl. Feldberg , Disentis, Davos and Säntis that a considerable part of the winter precipitation falls in the form of snow. For these six stations the reproduction of extreme-value properties of 10-day snowmelt has been analysed.

Historical estimates and simulated values of snowmelt were derived from the historical and generated daily precipitation and temperature, respectively. It was assumed that precipitation accumulates if $T < 0$ °C (snow cover) and that snowmelt is proportional to T ($T > 0$ °C) with the constant of proportionality equal to 4 mm/°C. The 10-day winter maxima were taken from the calculated snowmelt amounts. As in the previous section, the statistics *MAX*, *QM*5 and *M* of these extremes were used to assess the reproduction of these maxima.

Table 4 presents the average percentage differences between the values of *MAX*, *QM*5 and *M* for the three models and the values of these statistics for the historical data for the six stations of interest. The extremes are satisfactorily reproduced by model UE. The largest discrepancies here are found for Kahler Asten and Davos (overestimation of the median of the 10-day maxima). A similar overestimation is found for model UEc. Conditional simulation (model CE) results, for most stations, in a relatively large underprediction of the extreme-value properties of 10-day snowmelt. This phenomenon can partly be explained by the considerable negative bias in the daily temperature autocorrelation coefficients, which reduces the likelihood that snow accumulates over long periods and thus the probability of extreme multi-day snowmelt.

The historical winter snowmelt maxima at the Swiss stations are not higher than those at Kahler Asten and Freudenstadt. In particular for Säntis there is, however, a lot of snowmelt outside the winter period. For example, the maximum 10-day snowmelt amount (*MAX*) calculated for the whole year at this station is as high as 444.8 mm, while for the winter period it is only 198.0 mm.

Table 4: Percentage differences between the maxima (*MAX*), upper quintile means (*QM*5) and medians (*M*) of the 10-day snowmelt extremes for the simulated data (twenty-eight runs of 35 years) and the historical records (1961-1995) for six stations in the Rhine basin. The columns denoted with Hist. give the historical values (mm).

	MAX (%)				*QM5 (%)*				*M (%)*			
Station	UE	UEc	CE	Hist.	UE	UEc	CE	Hist.	UE	UEc	CE	Hist.
Kahler	-12.1	-12.8	-20.9	287.2	2.6	2.8	-7.3	184.6	20.7	20.3	9.2	86.6
Freudenstadt	-10.0	-6.4	-31.6	234.7	-11.2	-10.5	-30.2	180.4	-6.7	-8.8	-22.5	93.7
Kl. Feldberg	5.6	2.2	-18.4	151.6	-1.7	-6.0	-22.3	121.7	-6.5	-11.0	-23.0	71.3
Disentis	-13.6	-3.4	-34.2	171.8	-7.0	-3.3	-35.2	119.1	-7.6	-8.6	-23.0	67.8
Davos	-6.7	-4.9	-15.1	176.4	-4.8	-3.7	-12.8	137.8	11.0	14.5	8.7	63.6
Säntis	1.1	-5.8	-22.0	198.0	-13.6	-16.2	-29.2	157.0	-0.5	-1.9	-11.7	63.9

9 Long-duration simulations

For three 1000-year simulations Fig. 3 shows Gumbel plots of the 10-day winter precipitation maxima for the average of the 34 stations used in the feature vector.

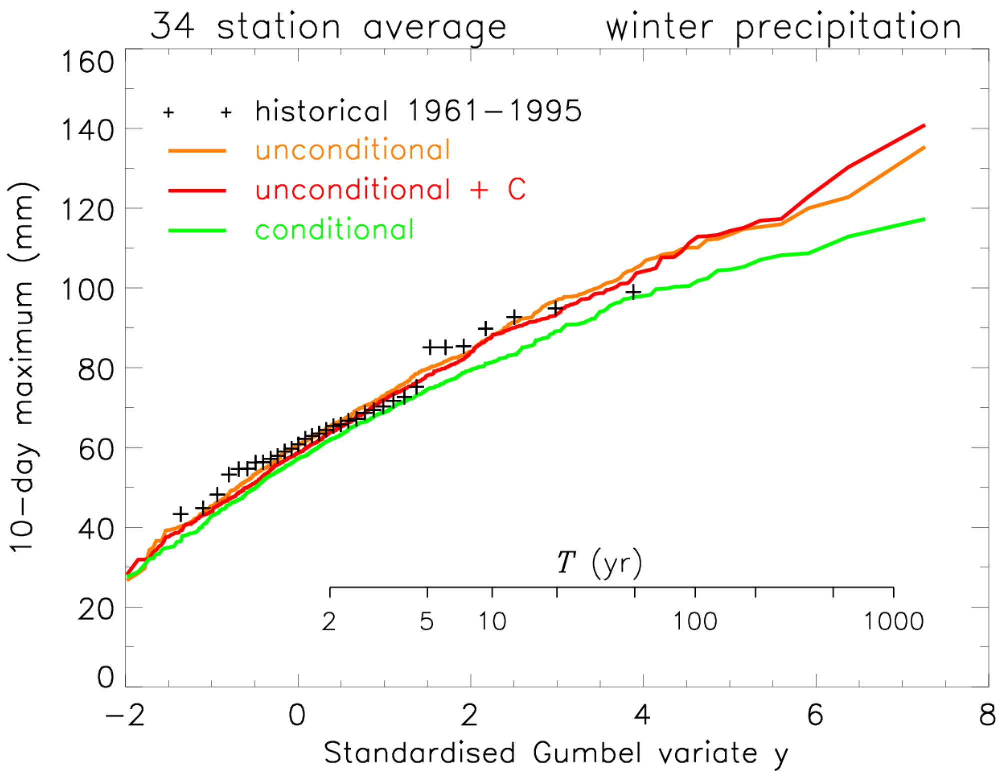

Figure 3: Gumbel plots of 10-day winter precipitation maxima for observed and simulated data (runs of 1000 years)

There is a reasonable correspondence between the historical and simulated distributions. The figure clearly shows the underestimation of the extreme-value properties for the conditional model CE, discussed in Section 7. Furthermore, in Fig. 3, a large part of the curve for the conditional 1000-year simulation lies below the curves for the unconditional 1000-year simulations. Realistic multi-day precipitation amounts much larger than the largest historical precipitation amounts are generated in all simulation experiments shown in Fig. 3.

10 Conclusions

The unconditional simulations preserved the second-order moment statistics of daily and monthly precipitation and 10-day maximum precipitation well. The lag 1 autocorrelation coefficients for daily precipitation and temperature were, however, significantly underestimated. The reproduction of the second order moments of temperature became worse in simulations where atmospheric circulation indices were added to the feature vector. Despite this deficiency the reproduction of 10-day maximum snowmelt was satisfactory.

Multi-site simulation of *P* and *T* conditional on simulated atmospheric circulation indices performed somewhat poorer than the unconditional simulations. Especially for temperature the reproduction of second-order moment statistics became worse. As a result a significant underestimation (up to 20-30%) of the median and the upper quintile mean of 10-day snowmelt amounts was observed for four high-elevation stations (Freudenstadt, Kl. Feldberg, Disentis, Säntis).

The ability of both unconditional and conditional models to generate realistic unprecedented multi-day rainfall events was demonstrated with simulation runs of 1000 years. Especially those extreme events may cause large peak discharges of the river Rhine in the Netherlands. A single simulation run of 1000 years does not provide, however, an accurate estimate of a 1000-year event. More simulations are needed for that purpose and even then a considerable uncertainty remains due to the use of a relatively short 35-year historical record for resampling.

Acknowledgements

The UK Meteorological Office gridded MSLP data were kindly provided by P.D. Jones (Climatic Research Unit, University of East Anglia, Norwich). The daily precipitation and temperature data for the German, Luxemburgian, French and Swiss stations were made available by the following institutions: Deutscher Wetterdienst, Service de la météorologie et de l 'hydrologie de Luxembourg, Météo France and the Swiss Meteorological Institute via the International Commission for the Hydrology of the Rhine Basin (CHR/KHR). The work was performed in co-operation with the Institute for Inland Water Management and Waste Water Treatment (RIZA).

References

J.J. Beersma and T.A. Buishand. Rainfall generator for the Rhine basin: Nearest-neighbour resampling of daily circulation indices and conditional generation of weather variables. KNMI-publication 186-III, KNMI, De Bilt, 1999.

T. Brandsma and T.A. Buishand. Rainfall generator for the Rhine basin: Multi-site generation of weather variables by nearest-neighbour resampling. KNMI-publication 186-II, KNMI, De Bilt, 1999.

T.A. Buishand and T. Brandsma. Multi-site simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbour resampling. Preprint 2000-14, KNMI, De Bilt (Submitted to Water Resour. Res.), 2000.

P.D. Jones, M. Hulme, and K.R. Briffa. A comparison of Lamb circulation types with an objective classification scheme. *Int. J. Climatol.*, 13: 655-663, 1993.

U. Lall and A. Sharma. A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.*, 32: 679-693, 1996.

B. Rajagopalan and U. Lall. A k-nearest-neighbor simulator for daily precipitation and other variables. *Water Resour. Res.*, 35: 3089-3101, 1999.

R. Wójcik, J.J. Beersma, and T.A. Buishand. Rainfall generator for the rhine basin: Multi-site generation of weather variables for the entire drainage area. KNMI-publication 186-IV, KNMI, De Bilt, 2000.

K. C. Young. A multivariate chain model for simulating climatic parameters from daily data. *J. Appl. Meteorol.*, 33: 661-671, 1994.

E. Zorita and H. von Storch. The analog method as a simple statistical downscaling technique: Comparison with more complicated models. *J. Climate*, 12: 2474-2488, 1999.