Extreme hydro-meteorological events and their probabilities

Jules Beersma

Promotor:

Prof. dr. A.A.M. Holtslag	Hoogleraar in de Meteorologie, Wageningen Universiteit

Co-promotor:

Dr. T.A. Buishand	Senior onderzoeker, Koninklijk Nederlands Meteorologisch Instituut (KNMI)

Samenstelling promotiecommissie:

Prof. dr. B.J.J.M. van den Hurk	Universiteit Utrecht; KNMI
Prof. dr. P. Kabat	Wageningen Universiteit
Prof. dr. H.H.G. Savenije	Technische Universiteit Delft
Prof. dr. F.W. Zwiers	Climate Research Divison, Environment Canada; University of Victoria (BC)

Dit onderzoek is uitgevoerd binnen de Buys Ballot Onderzoekschool (BBOS)

Extreme hydro-meteorological events and their probabilities

Extreme hydro-meteorologische gebeurtenissen en de kans daarop

Jules Beersma

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
Prof. dr. M.J. Kropff,
in het openbaar te verdedigen
op maandag 23 april 2007
des namiddags te 13:30 uur in de Aula.

ISBN 90-8504-615-7

> *The tool that is so dull that you cannot cut yourself on it is not likely to be sharp enough to be either useful or helpful.*
>
> John W. Tukey

Cited from *The Technical Tools of Statistics*, presented at the 125th Anniversary Meeting of the American Statistical Association, Boston, November 1964

Dankwoord

Voor ik iedereen ga noemen die van belang is geweest voor de totstandkoming van mijn proefschrift en voor mijn vorming als wetenschappelijk onderzoeker of als mens eerst een klein stukje persoonlijke geschiedenis. Het idee voor het schrijven van een proefschrift, op basis van gepubliceerde wetenschappelijke artikelen, stamt uit de tijd dat ik, in het kader van nauwere samenwerking tussen de werkgroep Klimaatscenario's (WKS) en de afdeling Voorspelbaarheidsonderzoek (VO), voor de helft van mijn tijd bij VO gedetacheerd was. De beide hoofden, Günther Können (WKS) en Theo Opteegh (VO), waren oprecht van mening dat een onderzoeker het aan zijn 'stand verplicht is' een proefschrift te schrijven. Zij waren bereid mij die ruimte te geven en vonden dat ik die ook moest benutten. Günther en Theo, ik ben jullie zeer dankbaar voor die ruimte en voor jullie inspiratie en aansporingen. Ook mijn latere baas Albert Klein Tank en huidige baas Arnout Feijt hebben mij alle ruimte gegeven en steeds gesteund.

Het oorspronkelijke idee was dat mijn proefschrift zowel zou gaan over het ontwikkelen van statistische methoden om verschillen of veranderingen in de variabiliteit te detecteren (zoals beschreven in hoofdstuk 2) als over onderzoek naar de oorzaken van verschillen en veranderingen in variabiliteit van het klimaat, met behulp van het binnen VO ontwikkelde 'intermediate complexity' klimaatmodel ECBilt. De eerder genoemde statistische methoden zouden daarbij in de praktijk worden gebracht om hun nut te demonstreren. Het liep echter anders. De detachering bij VO leverde wel een eenvoudige (diagnostische) wolkenparameterisatie, en, als gevolg daarvan, een noodzakelijke nieuwe kortgolvige stralingsparameterisatie voor ECBilt op maar geen publiceerbaar materiaal in het kader van onderzoek naar de variabiliteit van het klimaat. Geen papers voor mijn proefschrift dus, maar wel een nuttige en interessante ervaring op het gebied van klimaatmodellering. Uit die tijd bewaar ik goede herinneringen aan de samenwerking met Rein Haarsma, Frank Selten, Xueli Wang en Rob van Dorland. In dat rijtje hoort ook Michiel Schaeffer (RIVM) die op dat moment aan een nieuwe langgolvige stralingsparameterisatie voor

ECBilt werkte. Buiten de statistiek heb ik veel van jullie geleerd.

Dan het uiteindelijke onderwerp, of beter, de uiteindelijke onderwerpen van mijn proefschrift. Van de mensen van wie ik veel van statistiek in het algemeen en van resampling in het bijzonder heb opgestoken staat mijn co-promotor Adri Buishand met stip op één. Adri, het heeft misschien wat lang geduurd maar zonder jou was dit proefschrift er waarschijnlijk nooit gekomen. Daarnaast heb ik veel interessante gesprekken en discussies gehad met mijn kamergenoot Robert Leander, de man van ‘de Neerslaggenerator voor de Maas’, en diens voorganger Rafał Wójcik van wie ik nog heb geleerd om eetbare paddestoelen in het bos te zoeken. Memorabel zijn ook de bijeenkomsten met de RIZA partners. De gesprekken met Hendrik Buiteveld, Sacha de Goederen, Pieter Jacobs, Timo Kroon en Marcel de Wit, en vele anderen, hebben me duidelijk gemaakt wat de echte vragen zijn en wat voor de (hydrologische) praktijk van belang is. En dan natuurlijk Bert Holtslag, mijn promotor, die een heel herkenbare rol heeft gespeeld door, vooral in de slotfase, een vinger aan de pols te houden bij de planning en door mij te stimuleren om te proberen ‘voor een wat breder publiek’ te schrijven en daarmee de toegankelijkheid van het proefschrift te vergroten. Bert, bedankt voor je wijze en stimulerende invloed.

Naast de mensen die een inhoudelijke rol hebben gespeeld of van belang zijn geweest voor mijn wetenschappelijke vorming wil ik ook de (ex)KNMI-ers bedanken die op sportief en sociaal vlak voor onvergetelijke momenten hebben gezorgd; voor mij in betere tijden (vóór mijn knieongelukje) de leden van de hardloopclub en de voetbalclub, you know who you are, en last but not least de harde kern van ‘onze’ fietsclub: Cisco de Bruijn, Hans Cuijpers, Stephan de Roode, Wim de Rooy, Pier Siebesma, Job Verkaik en Rudolf van Westrhenen.

Tot slot wil ik hier stil staan bij mijn familie en dierbaren. Mijn ouders hebben, ondanks dat ze zelf nauwelijks een opleiding hebben genoten, omdat ze daarvoor de kans niet kregen, wel altijd ingezien dat je met een goede opleiding vaak verder komt dan zonder. Zij hebben mij die kans wel gegeven en mij daarin ook gestimuleerd en daar ben ik hen zeer dankbaar voor. Op het KNMI heb ik mijn grote liefde Janet Wijngaard leren kennen waarmee ik inmiddels al weer 15 jaar samen leef. In de periode dat het idee om te promoveren vaste vorm begon te krijgen werd Feija, onze eerste dochter, geboren. Ruim twee jaar later kwam ons tweede ‘meisie’ Hilde ons gezinnetje verblijden. Achterafgezien is het doen van promotieonderzoek, het schrijven van een proefschift en het krijgen en opvoeden van kinderen misschien niet helemaal optimaal getimed maar het is het wel helemaal waard geweest. Lieve Janet, jij hebt het mij nog heel gemakkelijk gemaakt door je altijd flexibel op te stellen en door een meer dan evenredig aandeel in de zorg voor ons voor je rekening te nemen. Zonder jou was ik wat dat betreft nergens geweest, maar ja, dan had ik waarschijnlijk ook geen twee bloedjes van meisjes gehad. Ik ben je daar bijzonder dankbaar voor. De komende jaren hoop ik het zowel voor jou als voor Feija en Hilde weer een beetje te kunnen compenseren. Meiden, na het serieuze werk is het nu tijd voor wat fun...

Abstract

Extreme hydro-meteorological events usually have a large impact on our society. For safety standards regarding life and property and for design purposes of large structures extreme events with return periods between 100 and 10 000 years are often required. A practical difficulty in determining such rare events is due to the fact that our instrumental meteorological records are typically not longer than about 100 years. We are thus interested in extreme events that may never have occurred in the instrumental history. Such extremes are therefore usually estimated by extrapolating a fitted probability distribution. The results obtained with statistical extrapolation methods, however, strongly depend on the assumed probability distribution. An attractive alternative to these classical methods is resampling of historical meteorological time series. Resampling is attractive since it is a nonparametric technique, which means that no assumptions about the underlying distributions of the data have to be made. In addition, resampling offers the opportunity to simulate different meteorological variables (multivariate) for different locations (multi-site) simultaneously, while the cross-correlations (between variables) and the spatial correlations (between locations) are automatically preserved. Resampling, finally, makes it possible to simulate much longer time series than the historical records from which is resampled. Such very long time series usually contain many unprecedented extreme events which can serve in a frequency analysis of the extremes. In short, resampling is a very suitable nonparametric technique to simulate multi-site multivariate meteorological time series that are much longer than those from the instrumental records. With specific hydrological applications in mind such very long resampled time series are used to determine the size and probabilities of occurrence of extremely wet periods in the Rhine basin (that may result in river flooding) and of extreme droughts in the Netherlands (leading to economic losses in agriculture and shipping). Resampling techniques are further used to determine the statistical uncertainty of extreme hydro-meteorological events and of other properties of (hydro-)meteorological data.

Contents

Dankwoord **vii**

Abstract **xi**

1 Introduction **1**
- 1.1 History of resampling procedures 4
- 1.2 Bootstrapping in a game of dice 6
- 1.3 Resampling in this thesis 8

2 A simple test for equality of variances in monthly climate data **13**

J.J. Beersma and T.A. Buishand, 1999,
Journal of Climate, **12**, 1770–1779

- 2.1 Introduction 14
- 2.2 Tests based on resampling 15
  - 2.2.1 Univariate tests 15
    - a. The jackknife 16
    - b. Permutation and bootstrap procedures 18
  - 2.2.2 Multivariate extensions 19
  - 2.2.3 A corrected jackknife test 20
  - 2.2.4 Power of tests for equality of variances 24
- 2.3 Examples 24
  - 2.3.1 Near-surface temperature 25
  - 2.3.2 Precipitation 27
  - 2.3.3 Comparison with other GCM simulations 29
- 2.4 Discussion 30

3 Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation **33**

J.J. Beersma and T.A. Buishand, 2003,
Climate Research, **25**, 121–133

3.1 Introduction 34
3.2 Methodology 35
  3.2.1 Nearest-neighbour resampling 35
  3.2.2 The analog method 37
  3.2.3 Data 38
  3.2.4 Standardization procedure 39
  3.2.5 Summary statistics 40
3.3 Model identification and simulation results 40
  3.3.1 Models used 40
  3.3.2 Model results 42
    a. Mean and second-order moments 43
    b. Selection effects 47
    c. Temporal dependence of spatial patterns 49
    d. Dry spell counts and dry spell lengths 50
  3.3.3 Long-duration simulations 52
3.4 Summary and conclusions 53

4 Joint probability of precipitation and discharge deficits in the Netherlands 57

J.J. Beersma and T.A. Buishand, 2004,
Water Resources Research, **40**, W12508, doi:10.1029/2004WR003265
4.1 Introduction 58
4.2 Drought characteristics 59
4.3 Probability distributions for the precipitation deficit 62
4.4 Probability distributions for the discharge deficit 63
4.5 Bivariate probability distributions 65
  4.5.1 Dependence structure 66
  4.5.2 Symmetry of dependence 69
  4.5.3 Goodness-of-fit 69
4.6 Failure regions 74
4.7 Discussion and conclusions 76

5 Drought in the Netherlands - Regional frequency analysis versus time series simulation 79

J.J. Beersma and T.A. Buishand, 2006,
Submitted to *Journal of Hydrology*
5.1 Introduction 80
5.2 Historic data and choice of districts 82
5.3 Regional frequency analysis 87
  5.3.1 Basic extreme-value distributions 87

5.3.2 Regional growth curve 89
5.4 Time series simulation 94
5.4.1 Nearest-neighbour resampling 94
5.4.2 Accuracy of quantile estimates 95
5.4.3 Reliability of jackknife and bootstrap standard error estimates 97
5.5 Discussion and conclusions 101

6 Summary and synthesis 105
6.1 Summary 105
6.2 Synthesis 108

A Estimation of correlation between pseudovalues 113

B Properties of kurtosis estimates 115

C Simulation of daily circulation indices 117
C.1 Introduction 117
C.2 Model construction 117
C.3 Model results 118
C.3.1 Autocorrelation of circulation indices 118
C.3.2 Run lengths of circulation types 119
C.4 Conclusions 120

D Nearest-neighbour resampling 121
D.1 Introduction 121
D.2 Model identification 122
D.3 Model results 123

E Maximum likelihood estimation of the dependence parameters ρ and α 125

F Significance of differences between districts 127

G AR(1) simulation model 129

Bibliography 131

Publications by the author related to this thesis 143

Samenvatting 149

Curriculum Vitae **155**

Acknowledgments **157**

Chapter 1

Introduction

Human societies are vulnerable to extreme hydro-meteorological events such as extreme drought and heavy precipitation. Typically societies are organized in such a way that events that occur, for instance, every year cause no or very little damage. For much rarer events, let's say events that occur less than once in hundred years, it is often not clear how large the damage in terms of life and property can be. Although effective protection may be possible, there is a price tag attached to the level of protection. Usually, decisions about protection levels are determined by cost-benefit and risk analyses as well as social and political choices. In the Netherlands, e.g., the level of protection against flooding along the embanked parts of the rivers Rhine and Meuse is based on the river discharge that is exceeded on average only once every 1250 years. This discharge is usually called the 1250-year discharge, or the design discharge.

In practice, it is often difficult to accurately determine the associated size or level of a T-year event[1]. This level is also denoted as the return level. To design a structure or system that meets a desired protection level, i.e., is able to withstand the T-year event, an accurate estimate of this event is needed.

The problem with very rare events is 'that they are so rare' that no statistics of these events exist. This is especially true when the extreme event is so rare that it has not occurred in the (documented) historical records at all. In that case the return level has to be 'predicted' from the less extreme events that have been recorded so far. The traditional way to obtain the return levels of very rare extreme events, which is common in (engineering) practice, is to

---

[1] By definition, the T-year event has a $1/T$ probability of being exceeded in any year, and therefore has a return period of T years. The T-year event corresponds with the $1 - 1/T$ quantile of the probability distribution of the variable of interest. Quantiles, T-year events and return levels therefore all refer to the same thing and are used interchangeably.

make use of a statistical method known as extreme value analysis. This involves fitting a probability distribution to the extreme values in the recorded data and extrapolating to the required level. Quite often, the extreme values are taken as the values exceeding a certain threshold or as the largest value in each year. As a result a large proportion of the data is unused. One argument for such a large thinning of the data is that fitting a probability distribution to the bulk of the data may lead to biased estimates of properties of extremes due to lack-of-fit in the tail of the distribution. However, the large thinning of the data in extreme value analysis introduces an undesired uncertainty to the resulting return levels.

Resampling is a computer-based technique that opens up new prospects to obtain reliable estimates of extreme events. The basic idea behind resampling is that new data samples are constructed in which the original data are 're-ordered' by sampling with replacement. Let us assume, for example, that daily rainfall data are resampled. Although resampling of daily rainfall amounts does not give new information about the probability of the 1-day rainfall amounts it does create new and unprecedented multi-day rainfall amounts as a result of the different temporal ordering of the daily data. Such unprecedented multi-day rainfall amounts could have occurred but did not occur, simply because of the limited length of the original data record. With resampling very long series of daily rainfall can be generated in which the (statistical) information of the original sample is used most efficiently. This enables a more accurate estimation of extreme multi-day rainfall events (Buishand, 2006). In short, while traditional extreme value analysis usually ignores a large part of the (statistical) information in the original data sample, resampling squeezes out as much (statistical) information as possible.

Is resampling then *the* method to go for? The answer, as always, depends on the type of problem that one wants to solve. In the example above it was already noted that resampling is *not* the solution if one is interested in the probability of extreme 1-day rainfall events and only daily rainfall data are available. But regarding the probability of extreme multi-day rainfall events resampling has added value. Similarly, resampling can improve the estimation of extreme 1-day rainfall events when sub-daily, e.g., hourly, rainfall data rather than daily data are available.

Severe impacts of hydro-meteorological events on society are often the result of extremes on the multi-day level. This applies for instance to accumulated (e.g., 10-day) extreme precipitation amounts which cause flooding in large river basins, or to extreme (summer) droughts which are largely due to persistent lack of precipitation during several months. For analyses of such hydro-meteorological extremes resampling of daily time series is a serious alternative.

And there is more in favour of time series resampling. Many hydro-meteorological applications need data that is both multivariate and multi-site in nature. For example, to model the (extreme) discharge of the river Rhine not only precipitation is important but also temperature since it largely determines evaporation (in summer) and it controls snow accumulation and snow melt, particularly in the mountainous areas (multivariate). In addition, precipitation and temperature are both needed, at a certain spatial resolution, for the whole river basin simultaneously (multi-site). In traditional parametric time series modelling it is generally necessary to make assumptions, which are often difficult to verify, about properties of the data before statistical methods can proceed. A big advantage of resampling procedures is that they are so-called *nonparametric* methods which means that they work without such assumptions about the underlying distribution of the data. In particular when multivariate and multi-site problems are considered, it is very convenient that no (correct or false) assumptions are needed regarding the distributional form and the temporal, spatial and mutual dependencies of the data since these are often seasonally or state dependent and therefore difficult to assess.

Besides as an alternative to traditional extreme value analysis resampling techniques are used in this thesis to assess statistical uncertainty. Statistical uncertainty is always around, although quite often we don't bother about it. This may be allowed for the mean of large samples, but in general, the uncertainty of characteristics of hydro-meteorological extremes is considerably larger than that of the mean and can therefore not be ignored. This is something we have to be aware of and have to live with. There is thus a serious need to quantify the statistical uncertainty of extreme events. Analytical expressions or approximations for the statistical uncertainty may strongly rely on assumptions about the underlying distribution. Resampling techniques do not require these assumptions and can even be used if no mathematical expression for the uncertainty is available.

Regarding statistical uncertainty, resampling in this thesis is used to determine the standard error (a measure of the statistical uncertainty) of the variance of (hydro-)meteorological variables, which makes it possible to test whether the variance of the data simulated with time series resampling equals that of the reference data, but also whether the variance of the data simulated with a climate model equals that of the observed data or whether the variance of the data simulated for the future climate differs from that for the current or past climate. Resampling is further used to determine the standard error of large quantiles (T-year events) and to construct confidence intervals for those quantiles.

It may be clear now that resampling procedures are the connecting thread in this thesis, but, although very important, they are only a means to an end. The ultimate goal is to identify and model relevant hydro-meteorological extremes and their probabilities of occurrence including uncertainty estimates.

This chapter continues with a brief overview of the history of resampling, followed by a simple example of playing dice to illustrate the potential of resampling. A description of the different topics and resampling procedures presented in the four papers (Chapters 2-5) that constitute the core of this thesis, including the interconnections between the topics and resampling procedures, concludes this chapter.

1.1 History of resampling procedures

The first resampling procedure that came into existence is the ‘jackknife’. The jackknife was introduced in the mid-1950s as a bias reduction technique. Quenouille (1956) laid the basis for the procedure that was called the jackknife for the first time by John W. Tukey in 1959 in an unpublished manuscript. In that manuscript Tukey (both a professor at Princeton University and a researcher at AT&T Bell Laboratories) compared the jackknife with a Boy Scout Jackknife, i.e., a large pocketknife with multipurpose blades (Brillinger, 1964) similar to a modern Swiss army knife, to underline that this tool can be used for more than bias correction only. In the 1960s and 70s the emphasis in resampling theory was on estimation of statistical uncertainty (i.e., standard errors).

To demonstrate the use of the jackknife, let us assume that we have a data sample of 100 values, and that we are interested in the statistical properties of a statistic (e.g., the mean, standard deviation, autocorrelation, skewness, kurtosis, etc.) of these data. The most used version of the jackknife consists of recomputing the statistic of interest a number of times with a different sample value (or group of sample values) deleted each time. Thus for the sample of 100 values, besides calculation of the statistic from the full sample (of 100 values), the statistic is recalculated from 100 jackknife samples in which one value is omitted (and 99 values are left in). The statistic from the full sample and those from the 100 jackknife samples together give the jackknife estimates of the bias and the standard error of the statistic.

In practice the jackknife works on almost any kind of statistic except for statistics directly related to sample quantiles, e.g., the median. It is clear that the jackknife procedure relies on the use of a computer and it is not surprising that in that era the development of resampling procedures and computers ran more or less parallel. Professor Tukey played a notable role in the development of computers (at AT&T) as well, and it is generally accepted that he has invented the names ‘bit’ and ‘software’.

About two decades later, in 1979, Bradley Efron, professor at Stanford University, introduced the ‘bootstrap’ to the world. He coined the term bootstrap since he wanted a word that sounds as good as ‘jackknife’ (Diaconis and Efron, 1983). The name bootstrap is derived from one of the surprising adventures of Baron Munchausen in which he ‘pulled himself up by his own bootstraps’ from the bottom of a deep lake (Efron and Tibshirani, 1993). Likewise, Efron’s bootstrap seems to accomplish the impossible (Johnson, 2001). Even more than the jackknife the bootstrap relies on the use of a computer. The bootstrap uses a computer to give a numerical value of the statistical uncertainty (e.g., the standard error) of a statistic without using a formula at all. The computer performs a (Monte Carlo) algorithm that consists of three steps: (i) a large number B, say 1000, of bootstrap samples is generated using a random number generator, where each bootstrap sample consists of a random sample of size n drawn with replacement from the original sample; (ii) for each bootstrap sample the statistic of interest is calculated (which results in B different values of the statistic); and (iii) the sample standard deviation of these values is calculated or a confidence interval is constructed from the empirical distribution of these B values.

In the above example of the sample of 100 values, each of the B bootstrap samples consists again of 100 values obtained from random sampling (with replacement) from the original sample. As a result certain values from the original sample will not be represented in a particular bootstrap sample while other values will be represented more than once. Note that, the number of possibly different bootstrap samples grows very rapidly with the size of the sample, e.g., for a sample of size 3 only 10 different bootstrap samples exist while for a sample of size 10 already 92 378 ($\sim$ hundred thousand) different bootstrap samples can be constructed. An analytical expression for the number of different bootstrap samples m from a sample of size n is provided by Davison and Hinkley (1997):

$$m = \frac{(2n-1)!}{n!(n-1)!} \tag{1.1}$$

A large part of the power of the bootstrap procedure comes from this rapidly growing number.

The bootstrap has been applied to a large number of problems including problems for which the correct answer is known. For the latter, it is shown that the algorithm provides a good uncertainty estimate, and it can be proved mathematically to work for similar, more complicated, problems (Diaconis and

Efron, 1983). And therefore, the bootstrap's good theoretical properties carry over into real statistical practice. In contrast to the jackknife, the bootstrap even works for statistics based on sample quantiles (although an accurate bootstrap estimate of the standard error of a sample quantile requires a large sample size).

Nearest-neighbour resampling was introduced to generate time series with a certain amount of 'persistence', i.e., dependence on previous values. It can be regarded as an extension of the ordinary bootstrap. The crucial difference between generating a bootstrap sample and nearest-neighbour resampling is that for the former the sampling of a new value (with replacement) is completely random while for the latter the sampling (with replacement) of a new value is – by nearest-neighbour selection – conditioned on the previously sampled value or on a number of previously sampled values. Another practical, but theoretically unimportant, difference is that bootstrap samples usually have the same size as the original sample since the bootstrap is usually used to determine statistical uncertainty which typically depends on the sample size, while nearest-neighbour resampling is predominantly used to generate time series (i.e., samples) which are much longer than the original one. It was however not until 1994 when Young (1994) introduced and applied such a type of time series model, which he called multivariate chain model, to simultaneously simulate minimum and maximum temperatures and daily precipitation. Independently, and apparently unaware of Young's work, Lall and Sharma (1996) developed resampling models similar in spirit to simulate hydrolic time series. They referred to their method as 'nearest-neighbor bootstrap' and 'nearest-neighbor resampling algorithm'. Lall, with various co-authors, pioneered and further developed the method. Rajagopalan and Lall (1999), e.g., used nearest-neighbour resampling to simulate daily precipitation and other weather variables simultaneously.

1.2 Bootstrapping in a game of dice

As an illustration of the power of bootstrapping, the bootstrap is applied to an experiment with playing dice. Let us assume that we throw a die 100 times and that we are interested in the chance that a run of 3 or more consecutive sixes occurs in those 100 throws. A run of at least 3 consecutive sixes in a series of 100 throws may serve as an analog of some (aggregated) extreme event in climate. If we assume that we use a perfect die (i.e., each of the six faces has a probability of exactly 1/6) than there is a rather complicated analytical solution for this probability which can be approximated very accurately with a formula[2] due to Feller (1968). This formula gives a probability of 0.316 for the required probability. Now let's roll the die. The outcome of the 100 throws is for example:

552654466454521315432324121111512655236156314656212
13264551446455454421654262242 $\boxed{666}$ 36551624221235535

In this sample of 100 throws a run of 3 sixes occurs in the second half of the series. Now assume that we are not in a situation to roll the die anymore and that we want to derive the probability of a run of at least 3 sixes from this particular sample (like in the real climate where we typically also have a single sample only). A wrong answer is obviously: we have *one* occurrence in *one* sample, so the probability is $1/1 = 1$. How can the bootstrap help us here? From our original sample we can construct new bootstrap samples by randomly selecting individual throws from the sample until they have a size of 100 again. If we construct a large number of bootstrap samples, we can approximate the probability of getting a run of at least 3 sixes in a series of 100 throws by dividing the number of bootstrap samples in which such a run occurs by the total number of bootstrap samples. The value that the bootstrap procedure (with 100 000 bootstrap samples) returns for this probability is 0.335 which is quite close to the theoretical value[3] of 0.316. We can even go a step further and use the bootstrap procedure for 'extrapolation'. Assume that instead of a run of at least 3 sixes we are interested in the probability of a run of at least 5 sixes, i.e., an event that is so rare that it doesn't even occur in our original sample (a situation that is very often encountered in real life problems). The bootstrap gives for this probability 0.011 while Feller's (1968) approximation gives 0.010. Thank you Mr. Efron for the bootstrap!

And that's still not everything. By applying an additional bootstrap procedure it is even possible to give an estimate of the statistical uncertainty of the (bootstrap) probability estimates. This procedure is also known as the double-bootstrap (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). First, N primary bootstrap samples (of size 100) are constructed from the original sample. Second, from each of the N primary bootstrap samples M secondary bootstrap samples (also of size 100) are constructed. For each of the N primary bootstrap samples, the M secondary bootstrap samples give

---

[2] The probability $P_{k,n}$ of at least k consecutive sixes in a series of n throws is approximated by: $P_{k,n} \approx 1-(1-px)/[q(k+1-kx)x^n]$ where $x = 1+qp^k+(k+1)q^2p^{2k}+(k+1)^2q^3p^{3k}+\dots$, $p = 1/6$ and $q = 1-p$.

[3] The reason why the bootstrap slightly overestimates this value is that in the sample of 100 throws the number of individual sixes, 17, is slightly larger than one would expect from a perfect die, i.e., 100 times 1/6 giving 16.667. As a consequence also the probability of runs of consecutive sixes will be somewhat overestimated

an estimate of the probability of getting a run of sixes in a series of 100 throws as in the single-bootstrap. The N primary bootstrap samples then provide a distribution of this probability. From this empirical distribution, which characterizes the statistical uncertainty, a standard error or confidence interval can be derived. Using our original sample, a double-bootstrap (with $N = 199$ and $M = 100\,000$) yields a 95% confidence interval[4] of [0.062, 0.738] for the probability of getting a run of at least 3 sixes in 100 throws, and similarly [0.001, 0.081] for a run of at least 5 sixes in 100 throws.

The example given above only serves to demonstrate the power of resampling techniques such as the bootstrap. It will be clear that the analysis of extreme events in hydro-meteorological data is incomparable with throwing a die. However, research in the past few decades has demonstrated that this is no restriction for resampling techniques to work in more complicated real life problems as encountered in the field of hydro-meteorology.

1.3 Resampling in this thesis

In this thesis resampling techniques are tailored and applied mainly in the field of hydro-meteorology. But this does not mean that these techniques could not be used in other areas. Neither does it mean that this thesis deals only with resampling. Although resampling definitely is a main theme, resampling methods are in several instances also accompanied and compared with more traditional (statistical) methods and the best performing method is used where possible. Another theme that recurs throughout this thesis is that of extreme events. The combination of extreme events and hydro-meteorology leads to: ‘extreme hydro-meteorological events’ as reflected in the title of the thesis. This involves modelling of extreme (unprecedented) hydro-meteorological events themselves (including their evolution in time), modelling of the return level (or the associated return period) of certain well-defined extreme (historical) events, as well as assessing their statistical uncertainty.

In Chapter 2 the jackknife is used to determine the statistical uncertainty of variances of climate data. Based on the jackknife uncertainty, tests for the equality of variances of monthly data are introduced. Note that, since the variance largely determines the width of the distribution, one can roughly state that the larger the variance, the larger the probability of extreme events. And thus a difference (or change) in the variance is a first (rough) indication of a difference (or change) in the frequency of extreme events.

---

[4] A 95% confidence interval denotes a random interval which contains the value of interest with 95% probability. In practice the interval is usually chosen such that there is equal probability (2.5%) that this value is outside the lower or upper boundary of the interval.

The motivation for developing these (non-parametric) variance tests is that the traditional F-test heavily relies on the assumption that the data come from a normal distribution and that it cannot be extended easily to the case of multiple correlated records on a grid as, e.g., provided by climate models. The jackknife provides a (distribution-free) standard error of the variance which is an essential ingredient for a test for equality of variances. The jackknife based test allows for a multi-site extension of the test by using data from several locations or grid points in a region. Such a combined test will be more powerful than that for a single location when the differences between the two climates have the same sign across the whole region because of the larger sample size.

As an example, let us assume that we have 30 locations and that at each location we have one sample of size 10 from a normal distribution with a certain variance and another independent sample of the same size from a normal distribution with a 4 times as large variance. The standard deviations differ thus by a factor of 2. If we consider only a single location (univariate test) the probability that this large difference in variance would be detected by a statistical test at the 5% level is less than 50%. Combining the data for the 30 locations (multivariate test) results in a detection probability of almost 100% if the samples at the different locations are not more than moderately correlated. A single combined test further avoids the difficult interpretation of multiple correlated (univariate) tests in a region which often leads to misinterpretation and is therefore notoriously known as the multiplicity problem (see, e.g., von Storch, 1982).

The multivariate test is applied in Chapter 2 to simulated time series of monthly mean near-surface temperature and precipitation from a climate change simulation (UKTR) with the UK Met Office Hadley Centre climate model. Besides this illustration, either the test statistic or the underlying jackknife procedure to estimate the standard error of the variance is applied for diagnostic purposes in all subsequent chapters.

Chapter 3 deals with modelling of extreme precipitation (and temperature) in the German part of the Rhine basin. With the 1250-year river discharge in mind, long-duration multi-site simulations of daily precipitation and temperature are performed by means of conditional nearest-neighbour resampling. In these long-duration simulations much larger 10-day area-average precipitation amounts are produced than observed in the 35-year reference period. Since, the weather observed on historical days is resampled as a whole, the dependence between daily precipitation at different sites and the dependence between daily precipitation and temperature is automatically preserved. As mentioned earlier, these dependencies often have a complicated structure, which may not be adequately described by parametric models or would otherwise require an unrealistic degree of modelling. For many hydrological applications, including the modelling of (extreme) river discharges in large river basins, the spatial and temporal dependencies are of crucial importance. This makes multivariate time series resampling models particularly suitable for hydrological purposes.

In conditional nearest-neighbour resampling the sampling of new values is conditioned on previously sampled values *and* on the large-scale atmospheric circulation. The rationale for simulating precipitation and temperature conditional on the large-scale atmospheric circulation is that the atmospheric circulation largely determines whether a day will be wet or dry and whether a day will be relatively warm or cold (for the time of the year). In these conditional resampling models the (change in) atmospheric circulation thus acts as a predictor for (the change) in precipitation and temperature. Recognizing that systematic changes in the atmospheric circulation are possible as a result of (anthropogenic) climate change, such conditional resampling models might be useful to assess (future) changes in precipitation statistics, in particular changes in extreme multi-day precipitation amounts.

In Chapter 3 conditional nearest-neighbour resampling models are also compared with the alternative, and closely related, analog method which is in use somewhat longer (Zorita et al., 1995; Zorita and von Storch, 1999) and which is also popular in the context of weather prediction.

While Chapter 3 deals with modelling of precipitation with a focus on extremely wet events, Chapter 4 deals with drought, i.e., extremely dry events. More explicitly, it deals with modelling the joint probability of precipitation deficits in the Netherlands and discharge deficits of the river Rhine. Large parts of the Netherlands can be supplied with water from this river in the case of precipitation deficits. For drought assessment it is therefore necessary to consider the joint probability distribution of precipitation and discharge deficits. In Chapter 4 nearest-neighbour resampling is used to estimate joint probabilities of precipitation and discharge deficits. The results are compared with those obtained from fitting different bivariate probability distributions. The (modelled) dependence structure between extreme precipitation and discharge deficits plays a crucial role in estimating joint (exceedance) probabilities. As can be expected nearest-neighbour resampling performs superior in this respect since the dependence structure is inherited from the original data without the need to make assumptions about it. In the framework of bivariate probability distributions the observed dependence structure can be reproduced by constructing a novel bivariate distribution: a bivariate normal distribution with the dependence structure taken from a bivariate Gumbel distribution.

In Chapter 5 spatial variation in the probability distribution of the precipitation deficit is the central theme. The Netherlands is therefore divided into six districts. In the probability distribution of the precipitation deficit of all six districts an extraordinary curvature shows up in the upper tail. Two alternatives are investigated to reproduce the spatial variation and the common extraordinary curvature in the tail: a regional frequency analysis and time series simulation by means of nearest-neighbour resampling. By introducing additional long-term persistence in the nearest-neighbour resampling procedure, the curvature in the upper tail of the distribution can be realistically reproduced.

The statistical uncertainty of quantile estimates based on nearest-neighbour resampling is assessed in Chapter 5 by combining nearest-neighbour resampling with the jackknife and the bootstrap. To achieve this, jackknife and bootstrap samples are constructed from the original data first. Subsequently, a simulation with the nearest-neighbour resampling model is performed based on each of the jackknife and bootstrap samples. The statistical uncertainty is finally determined in the same way as in the standard jackknife and bootstrap methods. In the case of the bootstrap, the procedure resembles very much that of the double-bootstrap in the dice example in Section 1.2. Apart from the standard errors of quantiles, nearest-neighbour resampling and the bootstrap are used to construct confidence intervals for the return periods of the largest observed precipitation deficit for each of the six districts.

Chapter 6, finally, gives a summary and synthesis of the research in this thesis.

Chapter 2

A simple test for equality of variances in monthly climate data

Jules J. Beersma and T. Adri Buishand, 1999

Published in *Journal of Climate*, **12**, 1770–1779

Abstract

Tests for equality of variances of monthly climate data using resampling techniques are discussed. The application of a jackknife test to spatially correlated time series is worked out in this chapter. Besides this spatial extension, it is also possible to combine the data for the individual calendar months into a single seasonal or annual test statistic. The derivation of the critical values of the test statistic from Student's t-distribution in such multivariate applications is investigated. A modification to improve the use of the t-distribution is given for the case that the distribution of the data is close to the normal distribution. The power of the simple jackknife test is compared with that of a permutation test.

The test is illustrated with a comparison of the variances of monthly temperatures and precipitation amounts in the anomaly simulation, with enhanced greenhouse gas concentrations, and in the control simulation of the high-resolution transient experiment (UKTR) with the Hadley Centre coupled ocean-atmosphere General Circulation Model. Three regions are considered: Central North America, Southern Europe and Northern Europe. For a num-

ber of regions and seasons the differences between the variances of the two simulations are significant at the 5% level. In particular, a significant increase in the variance of monthly precipitation over Northern Europe is found in the anomaly simulation for winter, summer and autumn. Limitations of the use of the test to monthly precipitation time series containing a large proportion of zeros are identified.

2.1 Introduction

The study of changes in the variance of meteorological variables is of recent interest. It is now well recognized that climate change may not be restricted to changes in the mean alone. Several authors have compared the variances of monthly and seasonal values of observed data or simulated data from General Circulation Models (GCMs). The determination of the statistical significance of observed differences meets, however, difficulties. Rind et al. (1989), Mearns et al. (1990), Cao et al. (1992) and Gordon and Hunt (1994) used the F-test for this purpose. The F-test assumes that the data are independent and normally distributed. Furthermore, the test often fails to discover meaningful differences in the variances due to lack of degrees of freedom. Zwiers and Thiébaux (1987) tried to overcome the low power of the F-test by deriving the interannual variability from the spectral density function of the daily values. Their test requires a careful elimination of the annual cycle in the mean. Moreover, the distribution of the test statistic has only been studied for (daily) samples from a normal distribution.

The above tests refer to data at a single location. GCM data consist, however, of a large number of correlated time series on a spatial grid. Wigley and Santer (1990) presented a number of tests to compare the variances of such multivariate data. Resampling techniques using computer-intensive Monte Carlo methods were proposed to decide whether a result is significant or not.

Buishand and Beersma (1996) discussed the use of the jackknife for the comparison of daily variability in observed and simulated climates. The jackknife method is a resampling technique which does not require Monte Carlo methods. The resulting tests are reasonably robust against non-normality of the data. The critical values can generally be based on Student's t-distribution both for univariate testing with data at a single location and for multivariate testing with data on a spatial grid.

This chapter focuses on the use of the jackknife for testing equality of variances of monthly values. Section 2.2 presents an overview of tests for equality of variances using resampling techniques. Particular attention is given to the jackknife method in the multivariate situation. The method is illustrated in

Section 2.3 with simulated monthly temperatures and precipitation amounts from the high-resolution transient experiment (UKTR) with the Hadley Centre coupled ocean-atmosphere GCM (Murphy, 1995; Murphy and Mitchell, 1995). Section 2.4 concludes the chapter with a discussion.

2.2 Tests based on resampling

For estimating standard errors resampling techniques are often good alternatives to analytic approximations. They also provide tests of significance in situations that the validity of the normal distribution is questionable. Several papers in the statistical literature have discussed the use of the jackknife and the bootstrap for testing equality of variances. An attractive property of these tests is that rather simple multivariate versions for samples on a spatial grid or samples of different seasons can be obtained. A correction of a standard jackknife test is proposed for such multivariate applications.

2.2.1 Univariate tests

In this section we confine ourselves to the monthly means (or totals) at a single location. A sample for J successive years (e.g., January average temperatures) is represented as $x_1, x_2, \ldots, x_J$. The sample mean is denoted as $\overline{x}$ and the unbiased sample variance s^2 is given by:

$$s^2 = \frac{1}{J-1} \sum_{j=1}^{J} (x_j - \overline{x})^2 \,. \tag{2.1}$$

The statistic s^2 is an unbiased estimate of the true variance σ^2 of the monthly values x_j if these data are independent, a quite common assumption for monthly data from different years. Tests for equality of variances are often based on $\hat{\theta} = \ln(s^2)$ rather than on s^2 itself, because the distribution of $\hat{\theta}$ is usually closer to the normal distribution than that of s^2. For independent data, $\mathrm{var}(\hat{\theta})$ can be approximated as (O'Brien, 1978):

$$\mathrm{var}(\hat{\theta}) \approx \frac{2 + \gamma_2}{J} \,, \tag{2.2}$$

where γ_2 is the kurtosis (a standardized fourth-order moment). For the normal distribution $\gamma_2 = 0$. An estimate of $\mathrm{var}(\hat{\theta})$ can be obtained by replacing γ_2 by the sample kurtosis:

$$\hat{\gamma}_2 = \frac{J \sum_{j=1}^{J} (x_j - \overline{x})^4}{\left[\sum_{j=1}^{J} (x_j - \overline{x})^2 \right]^2} - 3 \,. \tag{2.3}$$

Some caution is needed however, because $\hat{\gamma}_2$ can be seriously biased (see Appendix B). The jackknife provides a distribution-free alternative estimate of $\mathrm{var}(\hat{\theta})$.

Although s^2 is an unbiased estimate of σ^2 for independent and identically distributed data, $\hat{\theta} = \ln(s^2)$ is a biased estimate of $\theta = \ln(\sigma^2)$. The bias is of order $1/J$ (O'Brien, 1978):

$$\mathrm{bias}(\hat{\theta}) \approx \frac{-1 - 2\gamma_2}{J} \,. \tag{2.4}$$

a. The jackknife

In the jackknife method the statistic $\hat{\theta}$ is recomputed for each sub-sample of size $J-1$. Let $\hat{\theta}_{-j}$ be the value of the statistic after omitting x_j. From $\hat{\theta}$ and $\hat{\theta}_{-j}$ a pseudovalue can be formed as:

$$\theta_j^* = \hat{\theta} + (J-1)(\hat{\theta} - \hat{\theta}_{-j}) \,. \tag{2.5}$$

Although the pseudovalues θ_j^* can be seen as estimates of θ, they have a much larger variance than $\hat{\theta}$. However, their mean:

$$\hat{\theta}_{\mathrm{jack}} = \frac{1}{J} \sum_{j=1}^{J} \theta_j^* \,, \tag{2.6}$$

which is known as the jackknife estimate of θ (Miller, 1968), can be a good alternative to $\hat{\theta}$. The jackknife estimate reduces the bias in estimating $\ln(\sigma^2)$ to order $1/J^2$.

Unlike the $\hat{\theta}_{-j}$ values, the pseudovalues exhibit little correlation (Section 2.2.3). Jackknife tests treat the pseudovalues as independent normal variables. Tests for equality of variances are then similar to those for equality of means in normal populations using Student's t-distribution. These tests need $\hat{\theta}_{\mathrm{jack}}$ and its estimated variance:

$$\hat{V}_{\mathrm{jack}} = \frac{1}{J(J-1)} \sum_{j=1}^{J} \left(\theta_j^* - \hat{\theta}_{\mathrm{jack}} \right)^2 \,. \tag{2.7}$$

From the jackknife estimates a number of different statistics can be derived to test for equality of the variances $\sigma^2(\mathrm{I})$ and $\sigma^2(\mathrm{II})$ of two mutually independent time series of monthly climate data. Let $\hat{\theta}_{\mathrm{jack}}(\mathrm{I})$, $\hat{\theta}_{\mathrm{jack}}(\mathrm{II})$ be jackknife estimates of $\ln(\sigma^2)$ and $\hat{V}_{\mathrm{jack}}(\mathrm{I})$, $\hat{V}_{\mathrm{jack}}(\mathrm{II})$ their estimated variances, then the usual two-sample pooled t-statistic can be represented as:

$$T_{\mathrm{a}} = \left[\frac{JK(J+K-2)}{J+K}\right]^{1/2} \frac{\hat{\theta}_{\mathrm{jack}}(\mathrm{II}) - \hat{\theta}_{\mathrm{jack}}(\mathrm{I})}{\left[J(J-1)\hat{V}_{\mathrm{jack}}(\mathrm{I}) + K(K-1)\hat{V}_{\mathrm{jack}}(\mathrm{II})\right]^{1/2}}, \tag{2.8}$$

with J and K the number of years for climate I and climate II, respectively. Under the null hypothesis of equal variances, the distribution of T_{a} is approximated by Student's t-distribution with $J + K - 2$ degrees of freedom. For the case of equal sample sizes, Miller (1968) demonstrates that this approximation works well for sample sizes as small as $J = 10$. The test is also quite robust against non-normality. However, for $J \neq K$, Monte Carlo experiments show that the critical values of the test should be larger than those obtained from Student's t-distribution, especially for long-tailed distributions (Brown and Forsythe, 1974; Boos and Brownie, 1989). Besides the earlier mentioned assumptions about the normality and correlation of the pseudovalues, there are two additional complications in the case of unequal sample sizes that limit the use of Student's t-distribution with $J+K-2$ degrees of freedom (O'Brien, 1978). First, the pseudovalues have different variances in climate I and climate II if $J \neq K$. Second, the fact that $\mathrm{bias}(\hat{\theta}_{\mathrm{jack}})$ depends on J implies that the mean of the numerator of the test statistic slightly differs from zero under the null hypothesis. Furthermore, the two-sample Student test becomes less robust against non-normality if the sample sizes are unequal (Kendall and Stuart, 1973).

Keselman et al. (1979) suggested the use of Welch's t-statistic to cope with variance heterogeneity of the pseudovalues in case of unequal sample sizes. The test statistic reads:

$$T_{\mathrm{b}} = \frac{\hat{\theta}_{\mathrm{jack}}(\mathrm{II}) - \hat{\theta}_{\mathrm{jack}}(\mathrm{I})}{\left[\hat{V}_{\mathrm{jack}}(\mathrm{I}) + \hat{V}_{\mathrm{jack}}(\mathrm{II})\right]^{1/2}}. \tag{2.9}$$

Buishand and Beersma (1996) used a similar statistic to compare the daily variability in observed and simulated climates. The critical values of T_{b} are derived from Student's t-distribution with an effective number d^* of degrees of freedom:

$$d^* = \frac{\left[\hat{V}_{\mathrm{jack}}(\mathrm{I}) + \hat{V}_{\mathrm{jack}}(\mathrm{II})\right]^2}{\hat{V}^2_{\mathrm{jack}}(\mathrm{I})/(J-1) + \hat{V}^2_{\mathrm{jack}}(\mathrm{II})/(K-1)}. \tag{2.10}$$

For equal sample sizes, $T_\mathrm{b} = T_\mathrm{a}$, but d^* tends to be smaller than $J + K - 2$. Besides unequal sample sizes, differences in kurtosis γ_2 for the two climates also lead to variance heterogeneity. A correction of the test for correlation between pseudovalues is presented in Section 2.2.3.

b. Permutation and bootstrap procedures

Permutation procedures are computer-intensive techniques to determine the statistical significance of a result. The method is free of assumptions about the parametric form of the distribution of the data. A pooled permutation procedure can be used to test for equality of variances of two climate time series $\{x_1, \ldots, x_J\}$ and $\{y_1, \ldots, y_K\}$. The means $\overline{x}$ and $\overline{y}$ have to be subtracted first (Boos and Brownie, 1989). The method then assumes that under the null hypothesis each permutation of the combined sample $\{x_1 - \overline{x}, \ldots, x_J - \overline{x}, y_1 - \overline{y}, \ldots, y_K - \overline{y}\}$ is equally likely. A permutation sample is obtained by taking a sample of size J *without* replacement to represent the centred data for climate I; the remaining K values represent the centred data for climate II. For each permutation sample the ratio of the sample variances $s^2(\mathrm{I})$ and $s^2(\mathrm{II})$ is computed. Comparing the distribution of this ratio in the permutation samples with the observed ratio gives the achieved significance level.

In contrast to this permutation test, the pooled bootstrap procedure in Boos and Brownie (1989) resamples *with* replacement from the combined sample of centred data. The two techniques are further identical. Downton and Katz (1993) applied the pooled bootstrap technique to test for discontinuities in the variance in long-term records of seasonal mean maximum temperatures. They observed that a test at the 10% level can detect changes of 25–30% in the standard deviations of seasonal mean maximum temperatures in records of 10 years or more and that such a test is generally not sensitive enough to be able to detect changes less than 20%.

For the bootstrap it makes sense to consider studentized statistics like T_a and T_b instead of the ratio of the sample variances (Boos and Brownie, 1989). The actual rejection rate of the null hypothesis is then closer to the desired significance level. Boos and Brownie show that bootstrapping the jackknife statistic T_a results in an improvement compared with the use of Student's t-distribution with $J + K - 2$ degrees of freedom, in particular if $J \neq K$.

Pooled permutation and bootstrap procedures are not robust against unequal kurtoses. It is possible to achieve asymptotic correct significance levels in that case by bootstrapping the scaled samples $\{x_1/s(\mathrm{I}), \ldots, x_J/s(\mathrm{I})\}$, $\{y_1/s(\mathrm{II}), \ldots, y_K/s(\mathrm{II})\}$, separately (Boos et al., 1989). Because convergence is slow, the test cannot be applied to small and moderate samples (say $J, K \leq$

50).

2.2.2 Multivariate extensions

The jackknife procedure allows for a multivariate test for equality of variances in two different climates using data at several grid points in a region. Such a multivariate extension is presented in Buishand and Beersma (1996) First, the pseudovalues θ_1^*, θ_2^*, $\ldots$, θ_J^* are calculated for each grid point separately. These pseudovalues are then averaged over the various grid points, giving $\overline{\theta_1^*}$, $\overline{\theta_2^*}$, $\ldots$, $\overline{\theta_J^*}$. The jackknife statistic T_a or T_b is finally obtained by applying (2.6) and (2.7) to these average pseudovalues. This combined test will be more powerful than that for an individual grid point when the differences between climate I and climate II have the same sign across the whole region because of the larger sample size. The test is not suitable for very large regions (e.g., a hemisphere) where areas with negative differences may compensate those with positive differences.

The above multivariate extension compares spatial averages of the logarithms of the variances. This is equivalent with a comparison of the geometric means rather than the arithmetic means as in the SPRET1 statistic of Wigley and Santer (1990):

$$\mathrm{SPRET1} = \overline{s^2}(\mathrm{I})/\overline{s^2}(\mathrm{II})\,, \tag{2.11}$$

where $\overline{s^2}(\mathrm{I})$ and $\overline{s^2}(\mathrm{II})$ are the spatial averages of the sample variances[1] for climate I and climate II, respectively, for a particular calendar month. There is, however, not a simple approximation to the distribution of SPRET1 under the null hypothesis. Wigley and Santer (1990) used the pooled permutation procedure of Preisendorfer and Barnett (1983) to determine the statistical significance of the observed value of SPRET1. The method is usually unnecessarily restricted to equal sample sizes only. As for the univariate tests in Section 2.2.1, it is also necessary here to adjust the monthly values for differences in the means of the two climates (Santer and Wigley, 1990). Otherwise the kurtosis in each permutation would differ from that in the original series, resulting in an incorrect significance level.

The data for the individual (calendar) months can be combined into a single seasonal or annual test by averaging the monthly pseudovalues in a similar way. There is a gain in power when the sign of the differences in variance for the two climates is the same for the months under consideration.

---

[1] In contrast to the unbiased estimate in equation (2.1), Wigley and Santer divide the sum of the squared deviations about the mean by J rather than $J-1$. This choice does not influence the outcome of a permutation test and the value of the jackknife statistics T_a and T_b is affected only in case of unequal sample sizes.

On the other hand the combined test may fail when the sign of the differences varies over the year.

2.2.3 A corrected jackknife test

In Section 2.2.1 it was noted that in case of equal sample sizes the t-approximation of the null distribution did quite well in a jackknife test for sample sizes as small as 10. Correlation between the pseudovalues and the fact that their distribution deviates from the normal distribution, even if the data come from a normal distribution, limits the use of the t-distribution for smaller sample sizes. The situation is different in the multivariate extension of Section 2.2.2 because spatial averaging influences the distribution of the pseudovalues. The effect of spatial averaging on the validity of the t-approximation for the test based on the statistic T_{b} has been investigated in a Monte Carlo experiment. Table 2.1 considers both the situation of two single climate time series and that of averaging the pseudovalues of N independent sequences. This averaging does not affect the correlation between the pseudovalues, while the effect of non-normality of the pseudovalues decreases with increasing N. For N large

Table 2.1. Actual rejection rates of the null hypothesis of equal variances for two-sided tests based on the jackknife statistic T_{b} (5000 simulations). The pseudovalues in the test statistic are averaged over N independent sequences. The critical values of T_{b} are obtained from Student's t-distribution with d^* degrees of freedom.

				Significance level		
Distribution	J	K	N	0.100	0.050	0.010
Normal	5	5	1	0.074	0.035	0.009
			3	0.055	0.025	0.005
			5	0.051	0.021	0.003
			9	0.055	0.021	0.002
	10	10	1	0.099	0.050	0.012
			3	0.074	0.037	0.008
			5	0.077	0.032	0.005
			9	0.074	0.031	0.005
Exponential	10	10	1	0.169	0.107	0.037
			3	0.129	0.071	0.017
			5	0.123	0.066	0.016
			9	0.119	0.062	0.012

enough the distribution of T_{b} therefore no longer depends on N. The empirical significance levels for samples from the normal distribution are for large N much lower than the nominal values because of the negative correlation between the pseudovalues (O'Brien, 1978). The situation is in fact better if $N = 1$ because then the correlation effect is counteracted by the non-normality of the pseudovalues. For the case $J = K = 10$ the two effects just compensate. In the generated samples from the exponential distribution the correlation between the pseudovalues is positive (O'Brien, 1978). Because of this positive correlation and non-normality of the pseudovalues the test is progressive, i.e., the null hypothesis is rejected too frequently.

Like the F-test our jackknife statistic T_{b} has little power to detect differences in the variances of two short independent climate time series at a single location. Averaging over successive months or grid points is therefore necessary to obtain a meaningful test. Through the averaging procedure the effect of non-normality of the pseudovalues is small. Departures from the assumed t-distribution are then mainly due to correlation between the pseudovalues. These pseudovalues are equicorrelated, i.e.,

$$\mathrm{Corr}(\overline{\theta_i^*}, \overline{\theta_j^*}) = \rho \tag{2.12}$$

for all $i \neq j$. If ρ is known the test statistic can easily be corrected for this type of correlation (Walsh, 1947). The main point behind the correction is that $\hat{V}_{\mathrm{jack}}$ in equation (2.7) does not provide a purely unbiased estimate of $\mathrm{var}(\hat{\theta}_{\mathrm{jack}})$, but such an estimate is given by

$$\tilde{V}_{\mathrm{jack}} = \frac{1 + (J-1)\rho}{1-\rho} \hat{V}_{\mathrm{jack}} \, , \tag{2.13}$$

leading to the modified test statistic:

$$\tilde{T}_{\mathrm{b}} = \frac{\hat{\theta}_{\mathrm{jack}}(\mathrm{II}) - \hat{\theta}_{\mathrm{jack}}(\mathrm{I})}{\left[\tilde{V}_{\mathrm{jack}}(\mathrm{I}) + \tilde{V}_{\mathrm{jack}}(\mathrm{II})\right]^{1/2}} \, . \tag{2.14}$$

From the Satterthwaite procedure in Welch (1938) it follows that the variance estimate $\tilde{V}_{\mathrm{jack}}$ should also be used in equation (2.10) for the degrees of freedom. Table 2.2 shows that the null distribution of the corrected statistic $\tilde{T}_{\mathrm{b}}$ is generally much better approximated by Student's t-distribution than that of the jackknife statistic T_{b}. The corrected test even works well in case of unequal sample sizes despite the differences in the means of the jackknife estimates in the numerator of equation (2.14) under the null hypothesis.

Table 2.2 also presents estimates of ρ. Details about the derivation of these estimates are given in Appendix A. The table shows that the values

Table 2.2. Actual rejection rates of the null hypothesis of equal variances for two-sided tests based on the jackknife (2500 simulations for $J = 10$, $K = 30$; 5000 simulations in the other cases). The results in the first row refer to the jackknife statistic T_b and those in the second row to the corrected jackknife statistic $\tilde{T}_\text{b}$. The pseudovalues in the test statistics are averaged over $N = 9$ independent sequences. The estimated correlation coefficients between these pseudovalues in climate I and climate II are denoted as $\hat{\rho}(\text{I})$ and $\hat{\rho}(\text{II})$, respectively. The critical values of T_b and $\tilde{T}_\text{b}$ are obtained from Student's t-distribution with d^* degrees of freedom.

					Significance level		
Distribution	J	K	$\hat{\rho}(\text{I})$	$\hat{\rho}(\text{II})$	0.100	0.050	0.010
Normal	5	5	−0.063	−0.064	0.055	0.021	0.002
					0.095	0.048	0.008
	10	10	−0.017	−0.020	0.074	0.031	0.005
					0.098	0.050	0.010
	5	15	−0.066	−0.010	0.067	0.030	0.006
					0.111	0.054	0.015
	10	30	−0.015	−0.003	0.073	0.037	0.006
					0.100	0.049	0.013
Exponential	5	5	0.021	0.018	0.106	0.047	0.007
					0.090	0.037	0.006
	10	10	0.014	0.014	0.119	0.062	0.012
					0.098	0.050	0.008

of $\hat{\rho}$ are rather small. Nevertheless this correlation may have a considerable effect on the distribution of the test statistic, because it does not decrease with increasing separation in time. Unfortunately, the procedure on which the estimates of ρ in Table 2.2 are based does not apply to a single realization. Moreover, the amount of data is generally not sufficient to obtain a sensible estimate of ρ directly. The value of ρ is determined by the sample size and the underlying distribution. This dependence was examined in order to obtain a suitable modification of the jackknife statistic T_b.

Table 2.3 presents estimates of ρ for $J = 5$, 10 and 30 for a number of distributions. These values increase with increasing kurtosis γ_2 of the distribution. Both for the symmetric Laplace distribution and the skewed χ^2_4-distribution the effect of correlation on the distribution of the test statistic can be neglected. The kurtosis of these distributions is, however, as large as 3. The monthly means of climatic data generally have kurtosis close to zero. It is

Table 2.3. Estimated correlation coefficients between the pseudovalues of sequences of J independent observations from the normal and other distributions (10 000 simulations for $J = 5$ and $J = 10$; 2500 simulations for $J = 30$). As in Table 2.2 the correlation coefficients are derived from average pseudovalues taken over $N = 9$ independent sequences.

			$\hat{\rho}$		
Distribution	Skewness	Kurtosis	$J = 5$	$J = 10$	$J = 30$
Uniform	0	−1.2	−0.101	−0.040	−0.005
Normal	0	0	−0.064	−0.019	−0.003
Laplace	0	3	−0.013	0.002	0.002
χ_4^2	$\sqrt{2}$	3	−0.002	0.009	0.002
Exponential	2	6	0.020	0.014	0.004

therefore often sufficient to apply a correction to the test statistic valid for the normal distribution. The estimates of ρ for the normal distribution in Table 2.3 can be approximated as:

$$\tilde{\rho} = -J^{-1.7} \,. \tag{2.15}$$

Substitution of $\tilde{\rho}$ in equation (2.13) gives the desired correction. Unfortunately, it is difficult to verify the validity of this correction. The sample kurtosis in equation (2.3) has a very strong bias in small samples from distributions with positive kurtosis, the so-called leptokurtic distributions (see Appendix B). In the examples in Section 2.3, the kurtosis for a single grid point was estimated as

$$\hat{\gamma}_2 = \frac{n_s J \sum\limits_{i=1}^{n_s} \sum\limits_{j=1}^{J} (x_{i,j} - \overline{x}_{i\cdot})^4}{\left[\sum\limits_{i=1}^{n_s} \sum\limits_{j=1}^{J} (x_{i,j} - \overline{x}_{i\cdot})^2 \right]^2} - 3 \,, \tag{2.16}$$

where $x_{i,j}$ is the value of the ith calendar month for year J, $\overline{x}_{i\cdot}$ is the average of that calendar month, and n_s is the number of calendar months in the season of interest. The pooling over successive months reduces the bias because of the larger sample size. The estimate in equation (2.16) is, however, sensitive to a systematic variation of the variance within the season of interest.

2.2.4 Power of tests for equality of variances

A Monte Carlo experiment was performed to study the performance of the proposed jackknife test. The SPRET1 statistic of Wigley and Santer (1990) was also considered in that experiment. To demonstrate the effect of spatial averaging, one set of data was generated for univariate tests on the variances at a single location, and another set was generated for multivariate tests on the variances of $N = 30$ sequences. In the latter case, vectors of length 30 were generated from a multivariate normal distribution analogous to a Monte Carlo experiment of Zwiers (1987), where the correlation coefficient between the ith and jth sequence was set equal to the lag k autocorrelation coefficient of a second order autoregressive process:

$$\left.\begin{array}{rcl} \rho_0 & = & 1 \\ \rho_1 & = & \phi_1/(1-\phi_2) \\ \rho_k & = & \phi_1\rho_{k-1} + \phi_2\rho_{k-2}, \quad k \geq 2 \end{array}\right\} , \tag{2.17}$$

with $k = |i-j|$, $\phi_1 = 1.6$ and $\phi_2 = -0.8$. This correlation function represents a damped sine curve. It can be seen as the one-dimensional analog of the spatial correlation function of a climate variable exhibiting teleconnection patterns. From the Monte Carlo experiment it turns out that averaging the pseudovalues over $N = 30$ correlated sequences leads to a reduction in the standard error of $\hat{\theta}_{\text{jack}}$ of 66%, which is comparable to the effect of averaging over 9 independent sequences. The standard deviations, $\sigma(\text{I})$ in climate I and $\sigma(\text{II})$ in climate II, were taken to be the same for every sequence.

Table 2.4 presents the results for two-sided tests at the 5% level in samples of size 10. The power is low in the univariate case ($N = 1$), in agreement with the discussion on the power of the F-test in Zwiers and Thiébaux (1987). Even if $\sigma(\text{II})/\sigma(\text{I})$ is as large as 2 more than 50% of the cases passes the test. A large gain in power is achieved with the multivariate tests. About 80% of the cases is declared significant if $\sigma(\text{II})/\sigma(\text{I}) = 1.5$. It is further seen in Table 2.4 that the power of the simple jackknife test is comparable with that of a computer-intensive permutation test using the SPRET1 statistic.

2.3 Examples

The multivariate jackknife test in Section 2.2 was applied to simulated time series of monthly mean near-surface temperature and precipitation from the UKTR climate change experiment with the Hadley Centre coupled ocean-atmosphere GCM (Murphy, 1995; Murphy and Mitchell, 1995). Data of the

Table 2.4. Power of tests for equality of variances for various ratios of the standard deviations in climate I and climate II (1000 simulations, and 1000 permutations of each combined sample for the two climates to determine the statistical significance of the SPRET1 statistic). The values of the power refer to a two-sided test at the 5% level for samples of size 10 ($J = K = 10$) from a univariate ($N = 1$) or a multivariate ($N = 30$) normal distribution.

	$N = 1$		$N = 30$	
$\sigma(\mathrm{II})/\sigma(\mathrm{I})$	T_b	SPRET1	$\tilde{T}_\mathrm{b}$	SPRET1
1.2	0.077	0.076	0.273	0.262
1.5	0.176	0.175	0.808	0.842
2.0	0.434	0.399	0.996	0.994

last 10 years of the 75-year integration from the control simulation (with constant CO_2 concentration) are compared with those from the anomaly simulation for the same decade (with an increase in CO_2 of 1% per year, resulting in an effective CO_2 doubling after 70 years). The land areas of three regions are considered: central North America (CNA; 35°–50°N, 85°–105°W), southern Europe (SEU; 35°–50°N, 10°W–45°E) and northern Europe (NEU; 50°–70°N, 10°W–60°E). The first two regions were previously selected for analysis of regional climate change simulation by the Intergovernmental Panel on Climate Change (IPCC, 1990, 1996).

The latter region was introduced by Raisanen (1995) and later also considered by IPCC (1996). The monthly mean near-surface temperature is obtained by averaging the monthly mean maximum and minimum temperature. Results for monthly mean maximum and minimum temperature separately are similar to those for monthly mean temperature and are therefore not presented. The precipitation amounts considered are the sums of large scale and convective precipitation.

2.3.1 Near-surface temperature

Table 2.5 summarizes some relevant sample statistics. The values in this table are averages of monthly estimates over a season or year and over the land grid points in the region. The kurtosis estimates are generally close to zero. Exceptions occur in spring and autumn. Values of $\hat{\gamma}_2 > 1$ in those transition seasons are rather due to systematic differences between the temperature variances of successive calendar months than to leptokurtic distributions. Equation (2.15) was therefore used to correct for correlation between the pseudovalues in the

Table 2.5. Mean, variance and kurtosis of monthly near-surface temperature for central North America (CNA), southern Europe (SEU) and northern Europe (NEU). Here C refers to the control simulation and A to the anomaly simulation of the UKTR experiment: Dec–Feb (DJF), Mar–May (MAM), and etc.

Area	Data	DJF	MAM	JJA	SON	Year
			Mean (°C)			
CNA	C	−9.73	6.48	22.58	10.74	7.52
CNA	A	−4.96	8.95	27.00	15.55	11.63
SEU	C	−1.89	6.42	20.41	10.33	8.82
SEU	A	0.80	9.39	24.29	14.19	12.17
NEU	C	−22.04	−5.86	12.91	−2.70	−4.42
NEU	A	−17.34	−1.52	15.47	0.95	−0.61
			Variance (°C^2)			
CNA	C	13.39	5.65	4.28	4.98	7.07
CNA	A	11.58	4.70	7.30	4.84	7.10
SEU	C	11.74	6.19	4.38	4.24	6.64
SEU	A	11.90	3.94	5.33	3.63	6.20
NEU	C	17.49	11.01	2.92	8.09	9.88
NEU	A	23.16	6.53	3.73	5.37	9.70
			Kurtosis			
CNA	C	0.16	1.66	0.34	0.64	0.70
CNA	A	−0.55	−0.21	0.29	−0.23	−0.18
SEU	C	−0.08	1.15	0.07	1.90	0.76
SEU	A	−0.11	−0.24	0.53	−0.08	0.02
NEU	C	−0.36	0.00	0.17	1.20	0.25
NEU	A	−0.54	−0.01	0.35	0.55	0.09

jackknife test for equality of variances.

In the anomaly simulation the average temperature is about 4°C higher for most seasons. For the three regions the variances are larger in summer but smaller in spring and autumn. However, these changes in the variance resulting from enhanced greenhouse gas concentrations are statistically significant for only two cases; the 71% increase in summer for CNA and the 41% decrease in spring for NEU (Table 2.6). In winter the temperature variance changes have different signs.

Note that it is possible that the variance ratio indicates a decrease in variance whereas the test statistic indicates an increase in variance (see e.g., CNA

Table 2.6. Ratios of the sample variances of monthly near-surface temperature in the anomaly simulation to those in the control simulation and results of jackknife tests for equality of variances. Values in **bold** refer to statistically significant differences (two-sided test at the 5% level).

Area	Statistic	DJF	MAM	JJA	SON	Year
CNA	ratio	0.86	0.83	1.71	0.97	1.00
CNA	$\tilde{T}_{\mathrm{b}}$	−0.56	0.05	**2.54**	−0.26	1.01
SEU	ratio	1.01	0.64	1.22	0.86	0.94
SEU	$\tilde{T}_{\mathrm{b}}$	0.10	−1.41	0.96	−0.47	−0.43
NEU	ratio	1.32	0.59	1.28	0.66	0.98
NEU	$\tilde{T}_{\mathrm{b}}$	1.84	**−5.18**	0.74	−0.89	−0.75

in spring) or the other way around. However, this usually only happens when the test statistic is close to zero (and thus far from the critical value). It generally requires that the arithmetic mean is much different from the geometric mean which occurs if the variance shows large seasonal or spatial variation.

2.3.2 Precipitation

The distribution of monthly precipitation generally differs more from the normal distribution than that of monthly temperature. The largest departures from normality are found in areas or seasons where completely dry months frequently occur. If at a particular grid point the monthly mean precipitation is zero for the whole period considered, the sample variance is clearly also zero but the pseudovalues, since they involve $\ln(s^2)$, are undefined. Similarly, when only one of the monthly mean precipitation values in a time series is larger than zero, one of the pseudovalues is undefined. Both situations are found in a small number of the grid points in SEU in summer and autumn. Furthermore, time series containing many zeros have a strong effect on the spatial kurtosis estimate. To avoid problems related to such situations only those grid points are considered for which the monthly mean precipitation time series contains at least four values larger than zero.

It should further be noted that the precipitation in GCM simulations has often been regarded as being representative of the average over the grid box concerned (Reed, 1986; Gregory and Mitchell, 1995). The distribution of a spatial average of monthly precipitation is less skewed and has a lower kurtosis than that of monthly precipitation at a point.

Table 2.7 summarizes the sample statistics for precipitation in the same

Table 2.7. Mean, variance and kurtosis of monthly precipitation for CNA, SEU and NEU. Here C refers to the control simulation and A to the anomaly simulation of the UKTR experiment.

Area	Data	DJF	MAM	JJA	SON	Year
			Mean (mm day^{-1})			
CNA	C	1.16	3.00	3.07	1.52	2.19
CNA	A	1.30	3.05	2.73	1.44	2.13
SEU	C	2.98	2.31	2.01	1.93	2.30
SEU	A	3.01	2.32	1.61	1.76	2.18
NEU	C	1.34	1.72	2.34	2.24	1.91
NEU	A	1.68	1.95	2.51	2.58	2.18
			Variance (mm^2 day^{-2})			
CNA	C	0.31	1.03	1.63	0.76	0.93
CNA	A	0.34	1.47	1.79	0.92	1.13
SEU	C	0.91	0.82	1.17	0.85	0.93
SEU	A	1.37	0.76	1.11	0.81	1.01
NEU	C	0.29	0.44	0.67	0.46	0.46
NEU	A	0.47	0.48	0.92	0.68	0.64
			Kurtosis			
CNA	C	0.62	−0.03	−0.11	0.47	0.24
CNA	A	1.19	0.20	0.12	0.33	0.46
SEU	C	0.29	−0.08	0.60	0.26	0.27
SEU	A	0.16	−0.06	1.49	0.79	0.59
NEU	C	−0.19	−0.07	−0.03	−0.22	−0.13
NEU	A	−0.06	0.08	0.06	0.14	0.06

way as for temperature in Section 2.3.1. For NEU the monthly mean precipitation in the anomaly simulation is for all seasons 5 to 25% higher than in the control simulation. This increase in the mean is accompanied by an increase in the variance. Table 2.8 shows that the changes in variance vary between 10% (spring) and 60% (winter), and are, except for spring, statistically significant. For CNA the anomaly simulation shows an increase in mean winter precipitation of about 10% and a decrease in mean summer precipitation of about 10%; for SEU there is a 20% decrease of mean precipitation in summer and an almost 10% decrease in autumn (Table 2.7). For these two regions the changes in the mean are not accompanied by similar changes in the variance. The largest changes in the variance are found in other seasons, namely a statistically significant increase of 44% in spring for CNA and an increase of 50%

Table 2.8. Ratios of the sample variances of monthly precipitation in the anomaly simulation to those in the control simulation and results of jackknife tests for equality of variances. Values in **bold** refer to statistically significant differences (two-sided test at the 5% level).

Area	Statistic	DJF	MAM	JJA	SON	Year
CNA	ratio	1.09	1.44	1.09	1.22	1.21
CNA	$\tilde{T}_\mathrm{b}$	1.10	**2.25**	0.50	1.15	**2.36**
SEU	ratio	1.50	0.93	0.95	0.95	1.09
SEU	$\tilde{T}_\mathrm{b}$	1.36	−1.85	−0.53	−0.48	0.04
NEU	ratio	1.61	1.09	1.37	1.49	1.37
NEU	$\tilde{T}_\mathrm{b}$	**3.72**	1.03	**2.61**	**3.01**	**4.77**

in winter for SEU (Table 2.8).

In the statistical tests above, a correction for correlation between pseudovalues was applied using equation (2.15) for the normal distribution. The kurtosis estimates in Table 2.7 support this correction for most cases. Exceptions are CNA in winter and SEU in summer and autumn. In particular for the SEU precipitation, the positive kurtosis cannot be attributed to within-season variations of the variance only. According to the simulation results in Appendix B, a spatial average of $\hat{\gamma}_2$ in the range of 1 to 1.5 indicates that $\gamma_2 \approx 3$, so that a correction for correlation between pseudovalues would not be needed. For the cases mentioned above the correction only had a small effect; the values of the test statistic T_b without correction are: 0.99 for CNA in winter, −0.47 for SEU in summer, and −0.43 for SEU in autumn.

2.3.3 Comparison with other GCM simulations

The results for the UKTR experiment only partly agree with those of Rind et al. (1989), Gordon and Hunt (1994) and Liang et al. (1995) for mixed-layer models. In contrast to a coupled model, as used in the UKTR experiment, a mixed-layer model cannot produce variability associated with dynamical ocean processes such as the Atlantic thermohaline circulation and the El Niño Southern Oscillation. Since such processes contribute to the interannual variability, they should be included in experiments that investigate the response of atmospheric variability to enhanced greenhouse gas concentrations, as is demonstrated by Meehl et al. (1994). They found that the changes of (interannual) temperature variability in a mixed-layer version of their model differed from those in a coupled version, particularly in the tropics.

However, particular responses, that can be understood from physical relationships, seem quite robust. Examples are: reduced temperature variability over areas where sea-ice retreats (Gordon and Hunt, 1994; Meehl et al., 1994; Liang et al., 1995), enhanced summer temperature variability in areas of reduced soil moisture (Meehl et al., 1994; Liang et al., 1995), and enhanced precipitation variability due the enhanced hydrological cycle and greater atmospheric moisture content in the extra-tropics (Rind et al., 1989; Liang et al., 1995).

Liang et al. (1995), for example, found increased summer temperature variability over CNA, which they ascribe to reduced soil moisture. In UKTR there is an increased temperature variability over CNA in summer, which is accompanied by a reduction in mean precipitation, and this generally leads to reduced soil moisture in a warmer climate. With respect to enhanced precipitation variability, all substantial changes in precipitation variance (larger than 10%), in the three areas considered, are increases. Increases in precipitation variability over CNA in spring and summer similar to those in UKTR were also reported by Liang et al. (1995).

2.4 Discussion

A test for equality of variances based on the jackknife has been described that is suitable for correlated time series of monthly climate data on a spatial grid (e.g., those produced by GCMs). In contrast to other resampling techniques the method does not require computer-intensive simulation to derive the statistical significance of observed differences in variances. The null distribution of the test statistic can be approximated by Student's t-distribution with an effective number of degrees of freedom. For a test on multivariate data this approximation can be improved by a correction for correlation between the pseudovalues in the jackknife procedure. The proposed correction does, however, not apply if there are strong departures from the normal distribution as is for instance the case for monthly precipitation data containing a considerable fraction of zeros.

Besides the reported non-normality of monthly precipitation during the dry season in SEU, more serious problems were encountered with the application of the jackknife procedure to monthly precipitation in South-east Asia (5°–40°N, 60°–101°E). Even for the wet monsoon the distributions of the monthly precipitation of several grid points in the area appeared to be very leptokurtic. The area-average kurtosis can be reduced by excluding the relatively dry grid points from the analysis. Disregarding grid points with mean monthly precipitation smaller than 0.5 mm day^{-1}, yields a 18% increase in monthly precipitation variance in summer (June–August) and a 25% increase in the monsoon season (June–September). Both increases are significant at the 5% level. These results are in line with the increase in interannual variability of the area-averaged south Asian or Indian monsoon precipitation reported by Meehl and Washington (1993), and Bhaskaran et al. (1995).

Like the traditional F-test, the jackknife test in this chapter assumes that the monthly values from different years are independent. If there is a positive correlation between the values in successive years, then the jackknife variance tends to underestimate the true variance, which results in a progressive test.

Tests for equality of variances are known to have little power for typical sample sizes encountered in climate change experiments. In a jackknife test the low power is due to variability of $\hat{\theta}_{\text{jack}}$. The averaging of the pseudovalues over calendar months and/or grid points in a region leads to a considerable reduction in the standard error of $\hat{\theta}_{\text{jack}}$. Because monthly data generally exhibit no or only weak autocorrelation, averaging over three successive calendar months reduces the standard error of $\hat{\theta}_{\text{jack}}$ by about 40%. In the application to the monthly values in the UKTR experiment, spatial averaging over the grid points in each of the three regions yields a reduction in standard error of about 50% for temperature and 65% for precipitation. For temperature, the total reduction in standard error is comparable with that in the Monte Carlo experiment in Section 2.2.4. Despite these reductions in standard error quite substantial differences in variances can pass the test. For instance, for the monthly temperatures of NEU the changes in variance for the four seasons are 32%, −41%, 28% and −34%, respectively. Only the largest of these changes (corresponding to a change in standard deviation of about 20%) is significant at the 5% level. Furthermore, for precipitation the observed changes in the variance of 37% (NEU, summer), 44% (CNA, spring) and 49% (NEU, autumn) are statistically significant at the 5% level, but this is not the case for the observed increase of 50% in the variance of monthly precipitation in SEU during winter.

Despite the focus on monthly values, the presented jackknife procedure can, of course, also be used to compare the variances of seasonal values. However, for nearly normally distributed data, a test on the seasonal values (e.g., winter temperatures) has only about the same power as a test on the values for a particular calendar month (e.g., January temperatures). This is because $\text{var}(\hat{\theta}) \approx 2/J$ for both the monthly and seasonal values. For leptokurtic data, a seasonal mean or total will have much smaller kurtosis than the individual monthly values. It is therefore possible that the proposed correction for correlation between pseudovalues can be applied to the variances of seasonal values but not to the variances of monthly values. Furthermore, $\text{var}(\hat{\theta})$ will be smaller for the seasonal values due to their reduced kurtosis. This is advantageous for the power of a jackknife test on the seasonal variances.

Although there is strong evidence of an increase in the variance of monthly precipitation over NEU in the anomaly simulation, the relative variability or coefficient of variation (standard deviation divided by the mean) shows much less change. In principle, a test for equality of variation coefficients can be developed along the same lines as that for the variance in this paper. In case of absence of zero values, a test on the relative variability can also be obtained by applying the jackknife procedure to the variance of the logarithms of the monthly precipitation amounts.

Acknowledgments. The GCM data were kindly provided by D. Viner through the Climate Impacts LINK Project (Climatic Research Unit, University of East Anglia, Norwich, UK). The authors would like to thank the reviewers for their useful comments.

Chapter 3

Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation

Jules J. Beersma and T. Adri Buishand, 2003

Published in *Climate Research*, **25**, 121–133

Abstract

Nearest-neighbour resampling is used to generate multi-site sequences of daily precipitation and temperature in the Rhine basin. The simulation is conditional on the values of three continuous indices of the atmospheric circulation. An advantage of nearest-neighbour resampling is that the spatial correlations of the daily precipitation and temperature data are automatically preserved in the simulated data. Comparison of different resampling models shows that the simulation of the precipitation and temperature for a new day should not only be conditioned on the circulation characteristics of that day but also on the simulated precipitation and temperature for the previous day, in order to achieve the appropriate level of persistence and variability in the generated data.

With a hydrological application in mind, 980-year multi-site simulations of daily precipitation and temperature were performed conditional on a simulated time series of circulation indices that was obtained with a second resam-

pling model. The distribution of the extreme 10-day area-average precipitation amounts in these long-duration simulations was compared with the distribution of the historical 10-day area-averages. Again, the models in which the precipitation and temperature of the previously simulated day were taken into account performed best, but even these models somewhat underestimate the quantiles of the distribution of the 10-day area-average precipitation. The long-duration simulations demonstrate that nearest-neighbour resampling is capable to produce much larger 10-day area-average precipitation amounts than the historical maximum.

3.1 Introduction

A wide range of stochastic models is in use to simulate synthetic daily time series of precipitation. Some models are also able to simulate daily precipitation simultaneously with other weather variables. A limited number of techniques is available to generate weather variables simultaneously at multiple locations, which is of particular interest for hydrological applications. For assessments of the effects of anthropogenic climate change, there has been considerable interest to condition stochastic daily precipitation models on the large-scale atmospheric circulation. Quite often a classification of observed pressure fields into weather classes has been used for this purpose. The parameters of the precipitation model are then determined for each weather class separately (e.g., Bárdossy and Plate, 1992; Wilson et al., 1992; Schubert, 1994; Corte-Real et al., 1999; Fowler et al., 2000; Qian et al., 2002; Stehlík and Bárdossy, 2002). An alternative is to resample from the observed precipitation in the appropriate weather class (e.g., Hughes et al., 1993; Conway et al., 1996; Palutikof et al., 2002). A somewhat different approach is to describe daily precipitation by non-homogeneous hidden Markov models (e.g., Charles et al., 1999; Bellone et al., 2000). The parameters of such a stochastic precipitation model also depend on a discrete set of weather states, but these states are unobserved. The sequence of weather states is modelled as a first-order Markov chain of which the transition probabilities are determined by atmospheric predictor variables. Wilby et al. (1998) used regression techniques to link wet-dry transition probabilities and the means of a suite of weather variables to atmospheric circulation characteristics.

Zorita et al. (1995) and Zorita and von Storch (1999) used the analog method for the conditional simulation of multi-site daily precipitation. The word analog refers to the historical day that is closest to the target day in terms of atmospheric circulation characteristics. In this method the analog is sampled rather than days in a specific weather class. An advantage of resampling methods is that no conceptual extensions are required to generate multivariate and/or multi-site daily sequences. Neither do they require assumptions about the underlying distributions and the spatial correlations. The analog method can be seen as a special case of nearest-neighbour resampling. The earliest applications of nearest-neighbour resampling to weather data were on multivariate single-site simulation, without conditioning on the atmospheric circulation (Young, 1994; Rajagopalan and Lall, 1999). Using nearest-neighbour resampling Brandsma and Buishand (1998) compared unconditional simulations of daily temperature and precipitation with several simulations conditional on the atmospheric circulation. For the conditional simulations, it was found that the autocorrelation coefficients and extreme-value distributions of precipitation were better reproduced if, apart from circulation characteristics of the target day, the simulated precipitation and temperature of the previous day were also taken into account. A multi-site extension of unconditional nearest-neighbour simulation of daily precipitation and temperature was presented in Buishand and Brandsma (2001).

This chapter compares a stochastic version of the analog method with more general nearest-neighbour resampling techniques for conditional multi-site simulation of daily precipitation and temperature in the German part of the Rhine basin. This area was chosen with a specific application to rainfall-runoff modelling in mind. Since in the downstream area of the river Rhine the largest discharges occur in winter, the reproduction of precipitation statistics is studied for the winter half-year (October–March). Temperature is generated because rainfall-runoff models often use temperature to determine evapotranspiration, snow accumulation and snow melt.

The methodology and the data are presented in Section 3.2. Section 3.3 gives a description of the resampling models used and compares statistical properties of simulated data with those of observed data. In Section 3.4, finally, the results are summarized and conclusions are drawn.

3.2 Methodology

3.2.1 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 was used by Buishand and Brandsma (2001) for multi-site generation of daily precipitation and temperature.

In the nearest-neighbour method weather variables like precipitation and temperature are sampled simultaneously with replacement from the historical data. To incorporate autocorrelation, resampling is conditioned on 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. $\mathbf{D}_t$ was based on the standardized weather variables generated for day $t-1$ in Rajagopalan and Lall (1999) and of summary statistics of precipitation and temperature in Buishand and Brandsma (2001). Summary statistics are particularly needed for multi-site simulations to avoid problems with the high dimensional data space. As in earlier papers 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:

$$\delta(\mathbf{D}_t, \mathbf{D}_u) = \left(\sum_{j=1}^{q} w_j (v_{tj} - v_{uj})^2 \right)^{\frac{1}{2}} , \tag{3.1}$$

where v_{tj} and v_{uj} are the jth components of $\mathbf{D}_t$ and $\mathbf{D}_u$ respectively and w_j 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_j that the jth closest neighbour is resampled is given by:

$$p_j = \frac{1/j}{\sum\limits_{i=1}^{k} 1/i}, \quad j = 1, ..., k\,. \tag{3.2}$$

This probability kernel was also adopted in earlier applications of nearest-neighbour resampling for the Rhine basin (Buishand and Brandsma, 2001).

For the simulation of weather variables conditional on the atmospheric circulation (or CNNR: conditional nearest-neighbour resampling) the procedure is slightly different. In that type of simulation one searches for days in the historical record that have similar atmospheric circulation characteristics as those of the conditioning day. Again one of these nearest neighbours is randomly selected and the observed values of that nearest neighbour are adopted as the simulated values for the conditioning day t. The feature vector $\mathbf{D}_t$ should therefore at least consist of circulation characteristics of the conditioning day

t. In addition, simulated weather variables and/or circulation characteristics of day $t-1$ and earlier days could be included in the feature vector.

Apart from creating a feature vector, the number k of nearest neighbours and the weights w_j have to be specified. The choice of k depends on the type of probability kernel $\{p_j\}$, the number n of daily values from which the nearest neighbours are selected, and the dimension q of the feature vector. Lall and Sharma (1996) recommended for the decreasing kernel (equation 3.2) $k = n^{1/2}$ provided that $1 \leq q \leq 6$ and $n \geq 100$. Young (1994) recommended $k = 3$ using a uniform kernel, while q was 3 and $n \approx 1200$. A sensitivity analysis in Buishand and Brandsma (2001), with the decreasing kernel and similar values for n and q as in our application, gave best results for $k = 2$ and $k = 5$. In this study the decreasing kernel with $k = 5$ was adopted. To obtain an equal contribution of all feature vector elements to the Euclidean distance, the weights w_j should be inversely proportional to the variance of the feature vector elements. This is usually a good starting point and Buishand and Brandsma (2001) showed that variation of the weights generally has little effect on the statistical properties of the simulated data. In Wójcik and Buishand (2003) an alternative approach is introduced that avoids specification of the weights by using the Mahalanobis distance instead of the Euclidean distance.

3.2.2 The analog method

The analog method (e.g., Zorita et al., 1995; Zorita and von Storch, 1999) is basically a special case of CNNR. In nearest-neighbour resampling, one of the k nearest neighbours is randomly selected from the historical record, whereas in the analog method, the closest one is always selected. The analog method is therefore identical to CNNR with $k = 1$.

Zorita et al. (1995) and Zorita and von Storch (1999) based the search for analog days on characteristics of a single conditioning day or a sequence of conditioning days. In those papers the conditioning characteristics referred to the atmospheric circulation only. Since in the analog method no randomness is involved, this method is in essence deterministic. There is thus only one realization of the simulated time series for each conditioning time series. Consequently, for simulation conditional on the historical time series of circulation indices, the conditioning day itself must be excluded, because otherwise the historical time series of weather data would be generated.

In this thesis the five best analog days were extracted from the historical record and one of these analogs was randomly selected using the decreasing kernel (equation 3.2) with $k = 5$. This stochastic version is better comparable with the CNNR models than the originally deterministic analog method.

3.2.3 Data

As in Buishand and Brandsma (2001) daily precipitation and temperature data from 25 German stations in the Rhine basin for the period 1961–1995 were used (see Figure 3.1). For the 22 stations that lie below 500 m, the mean annual precipitation ranges from 542 mm (Geisenheim) to 944 mm (Freiburg) and the mean annual temperature lies between 8.2°C (Coburg) and 10.9°C (Freiburg). The three remaining stations, at an altitude of about 800 m, have relatively lower mean annual temperatures and higher mean annual precipitation, the

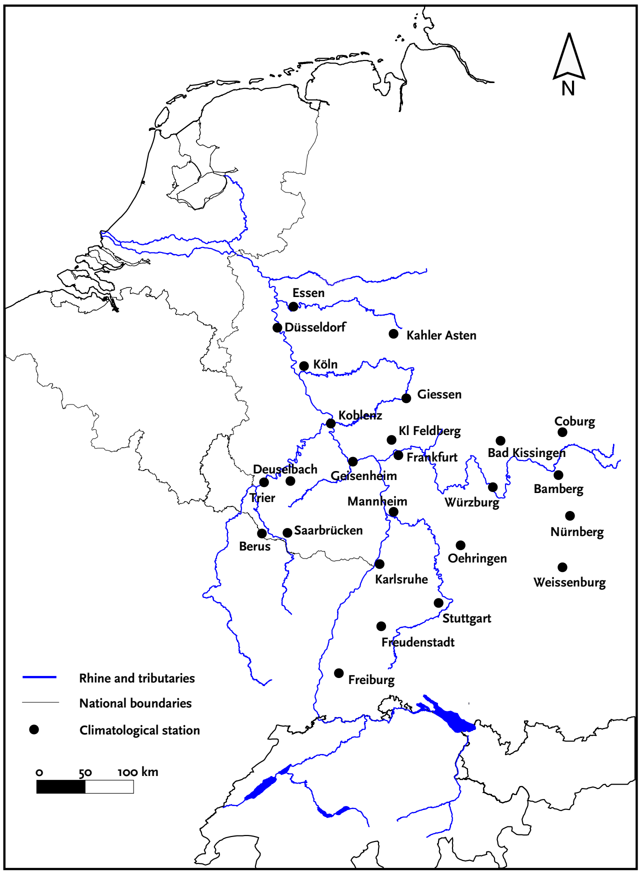

Figure 3.1. Locations of the 25 German stations in the drainage basin of the river Rhine used in this study.

latter due to orographic enhancement. The lowest mean annual temperature (5.0°C) is observed for Kahler Asten and the largest mean annual precipitation (1691 mm) for Freudenstadt.

For the same 35-year period three daily indices of the atmospheric circulation were used: (i) strength of the westerly flow W; (ii) strength of the southerly flow S; and (iii) relative vorticity Z. As in Jones et al. (1993) these circulation indices were derived from daily mean sea level pressure data from the UK Meteorological Office on a 5° latitude by 10° longitude grid, except that the grid was centered at the Rhine basin instead of the British Isles. In a number of studies the same circulation indices were used to obtain an objective version of the Lamb classification (e.g., Jenkinson and Collinson, 1977; Jones et al., 1993; Goodess and Palutikof, 1998; Linderson, 2001).

3.2.4 Standardization procedure

To reduce the seasonal variation in the feature vector elements, precipitation, temperature and circulation indices were standardized. 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:

$$\tilde{x}_t = (x_t - m_d)/s_d, \quad t = 1, ..., 365J \quad \text{and} \quad d = (t-1) \bmod 365 + 1\,, \tag{3.3}$$

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 in a similar way as in Brandsma and Buishand (1998) and Wójcik and Buishand (2003).

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

$$\tilde{x}_t = x_t/m_{d,\text{wet}}, \quad t = 1, ..., 365J \quad \text{and} \quad d = (t-1) \bmod 365 + 1\,, \tag{3.4}$$

with wet days defined as days with 0.1 mm precipitation or more.

To facilitate the reproduction of seasonally varying weather characteristics 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 (1998) and Buishand and Brandsma (2001). Thus for the 35-year historical record the nearest neighbours are selected from $n = 2135$ days. At the end of the resampling procedure the simulated standardized variables are re-transformed to their original scale using the inverse of equations (3.3) and (3.4).

3.2.5 Summary statistics

For the 25 stations in Figure 3.1 precipitation P and temperature T observations were available for each day. Parsimony of feature vector elements requires that the P and T fields are described by a small number of summary statistics, like the three circulation indices were used to characterize the mean sea level pressure field (e.g., the atmospheric circulation). Otherwise, considerable differences between the k nearest neighbours may occur because of the large dimension of $\mathbf{D}_t$. Computer time also increases with the dimension of $\mathbf{D}_t$.

Two important summary statistics are the arithmetic means of the standardized values of the P and T fields:

$$\tilde{P} = \frac{1}{25} \sum_{i=1}^{25} \tilde{P}_i \tag{3.5}$$

$$\tilde{T} = \frac{1}{25} \sum_{i=1}^{25} \tilde{T}_i \,, \tag{3.6}$$

where $\tilde{P}_i$ and $\tilde{T}_i$ are the standardized P and T values, respectively, for the ith station. Because of the relatively large spatial variation of the P field, there is some need for a more complete summary of this field than just $\tilde{P}$. An additional statistic to summarize the P field is the fraction F of stations with precipitation above some threshold as suggested in Buishand and Brandsma (2001). Here F was used with a threshold of 0.1 mm. The statistic F helps to distinguish between large-scale and convective precipitation. Buishand and Brandsma (2001) also considered two alternatives to the $\tilde{P}$ and F combination: a vector consisting of the daily averages of the standardized values over five different sub-regions and a vector consisting of the five leading principal components obtained from the sample covariance matrix of the $\tilde{P}_i$. With respect to the reproduction of the standard deviation and the autocorrelation coefficients of both precipitation and temperature these two alternatives did not give better results than the simulations with the combination of $\tilde{P}$ and F (Buishand and Brandsma, 2001).

3.3 Model identification and simulation results

3.3.1 Models used

Six resampling models were considered. Three analog-type models were distinguished: a first-order, a third-order and a fifth-order model. In the first-order analog model (analog 1) the circulation indices of the conditioning day

(day t) were used to find the analog days. In the third and fifth-order analog models (analog 3 and analog 5) the search for analog days was based on the circulation indices of respectively 3 and 5 consecutive conditioning days (days $t-2$, $t-1$, t and $t-4$, $\ldots$, t respectively). In the higher-order analog models thus also a part of the evolution of the atmospheric circulation is taken into account. Zorita et al. (1995) refer to the fifth-order model as a 'five-day-segment' model. The remaining three resampling models are CNNR models. The first of these models (CNNR 1) contains the circulation indices of day t and simulated precipitation and temperature characteristics of day $t-1$ as feature vector elements. The second model (CNNR 2**C**) additionally contains the circulation indices of day $t-1$, yielding a second-order model in terms of the atmospheric circulation. Model CNNR 2F, finally, uses in addition to CNNR 1 the fraction F of stations with precipitation of day $t-2$ to determine the nearest neighbours, resulting in a second-order model in terms of precipitation. Figure 3.2 schematically presents the feature vectors of these six models.

For a fair comparison of both types of models a stochastic version of the analog method was used (see Section 3.2.2). Further, the selection of the conditioning day was excluded in the CNNR models as in the analog models. Allowing the selection of the conditioning day is considered to generate 'artificial skill' (Zorita et al., 1995). Consequently, the only difference between the analog models and the CNNR models examined here is the composition of the feature vector (see Figure 3.2).

The weights w_j in equation (3.1) are in the CNNR models approximately equal to the reciprocal of the variance of the feature vector elements. The weights for $\tilde{P}$, F and $\tilde{T}$ were rounded to 2, 5 and 1 respectively, and for $\tilde{Z}$, $\tilde{W}$ and $\tilde{S}$ the weights are 1 as a result of the standardization. For the analog models all weights equal 1 since the feature vector elements involve only standardized circulation indices.

In Section 3.3.2, two types of conditional simulations are investigated: simulations conditional on the 1961–1995 time series of circulation indices and simulations conditional on simulated time series of circulation indices. The simulated time series of circulation indices were obtained with an unconditional nearest-neighbour resampling model. A description of that model is given in Appendix C. Time series of simulated circulation indices are needed to generate longer time series of P and T than the historical time series of circulation indices (see Section 3.3.3). In the simulations conditional on *simulated* time series of circulation indices the conditioning day itself was not excluded from being selected since no artificial skill can be inherited from a simulated time series of circulation indices.

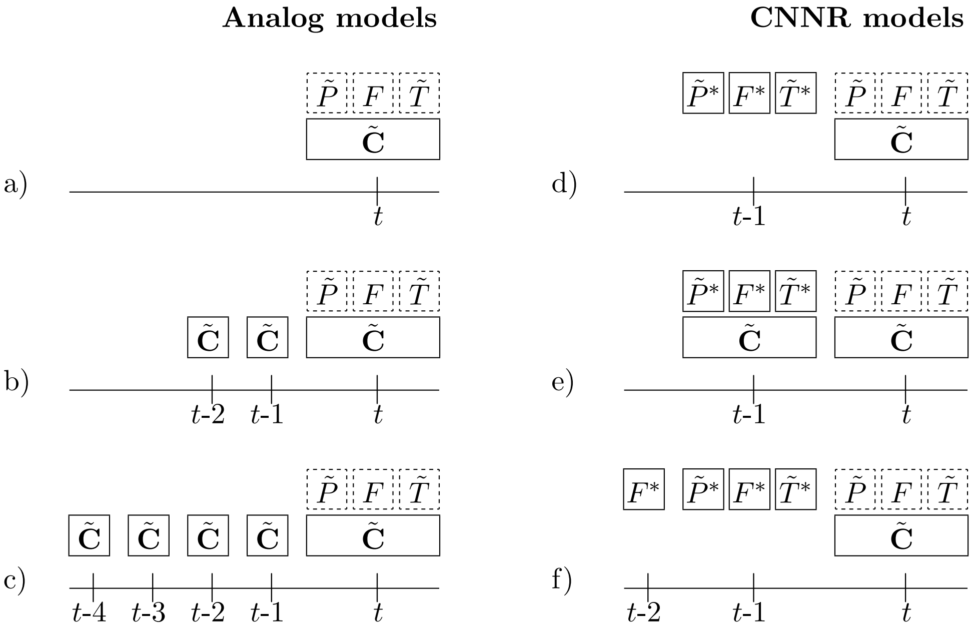

Figure 3.2. Elements of the feature vector (solid boxes) for conditional simulations of new variables (dashed boxes): (a) analog 1; (b) analog 3; (c) analog 5; (d) CNNR 1; (e) CNNR 2**C** and (f) CNNR 2F. The vector **C** contains the three circulation indices Z, W and S; asterisk indicates that the corresponding variable was simulated in a previous time step; tilde refers to a standardized value.

3.3.2 Model results

With all six models two 980-year simulations were performed, a simulation consisting of 28 runs of 35 years conditional on the same 35-year record of circulation indices, and a single 980-year simulation run conditional on 980 years of simulated circulation indices. For comparisons with the historical data the latter was split into 28 independent 35-year records. Given the application of rainfall-runoff modelling for the river Rhine, the presented statistics refer to the winter half-year (October–March). Second-order moments (such as standard deviations and autocorrelation coefficients) were first calculated for each calendar month separately as in Buishand and Brandsma (2001) and then averaged over the six calendar months, the 25 stations and the 28 periods of 35 years in order to reduce the influence of the seasonal cycle in the mean on these statistics.

a. Mean and second-order moments

Table 3.1 gives an overview of the reproduction of the means, the standard deviations of the monthly and daily values, s_{M} and s_{D} respectively, and the lag 1 and lag 2 autocorrelation coefficients of the daily values, $r(1)$ and $r(2)$ respectively. The first part of the table refers to simulations conditional on historical circulation indices, and the second part to simulations conditional on simulated circulation indices. The table also gives the historical estimates and

Table 3.1. Differences in means and second-order moments between the 980-year simulations and the historical data for the winter (October–March), averaged over the 25 stations and the 28 (35-year) runs. For the mean precipitation (monthly totals), the mean temperature and the mean lag 1 and lag 2 autocorrelation coefficients, $r(1)$ and $r(2)$, the absolute differences are given, and for the mean standard deviations of monthly and daily values (s_{M} and s_{D}) the percentage differences are given. Bottom lines: average historical (1961–1995) estimates and their standard error, *se*. Mean and standard deviations are in mm for precipitation and in °C for temperature. *se* are in mm for precipitation, in °C for temperature, in % for standard deviations and dimensionless for the autocorrelation coefficients. Values in **bold** refer to statistically significant differences.

	mean		s_{M}		s_{D}		$r(1)$		$r(2)$	
Model	*P*	*T*	*P*	*T*	*P*	*T*	*P*	*T*	*P*	*T*
Historical circulation indices (1961–1995)										
CNNR 1	1.0	0.27	−5.6	**−19.9**	−0.1	**−6.8**	**−0.047**	**−0.096**	**−0.028**	**−0.069**
CNNR 2**C**	−5.3	0.31	−8.8	**−24.6**	−4.6	**−7.8**	**−0.048**	**−0.123**	**−0.026**	**−0.120**
CNNR 2*F*	0.2	0.29	−2.1	**−22.0**	0.0	**−7.5**	**−0.042**	**−0.104**	**−0.012**	**−0.087**
Analog 1	−1.7	0.02	**−15.4**	**−42.4**	−0.9	−1.4	**−0.181**	**−0.493**	**−0.078**	**−0.397**
Analog 3	**−7.9**	0.20	**−14.8**	**−34.4**	**−5.4**	−2.5	**−0.156**	**−0.352**	**−0.057**	**−0.288**
Analog 5	**−12.1**	0.21	**−18.5**	**−28.5**	**−9.4**	−2.2	**−0.161**	**−0.326**	**−0.063**	**−0.248**
Simulated circulation indices										
CNNR 1	−0.8	0.26	−8.7	**−23.5**	−1.2	**−7.3**	**−0.049**	**−0.099**	**−0.027**	**−0.077**
CNNR 2**C**	−6.1	0.24	**−13.6**	**−27.6**	−5.0	**−7.2**	**−0.054**	**−0.118**	**−0.030**	**−0.118**
CNNR 2*F*	−1.8	0.28	−7.6	**−26.5**	−2.0	**−7.8**	**−0.048**	**−0.108**	**−0.016**	**−0.096**
Analog 1	−2.9	0.04	**−18.6**	**−42.2**	−2.1	−1.3	**−0.168**	**−0.453**	**−0.073**	**−0.364**
Analog 3	−6.8	0.16	**−17.7**	**−36.2**	−4.4	−1.7	**−0.147**	**−0.336**	**−0.062**	**−0.282**
Analog 5	**−9.3**	0.20	**−20.2**	**−33.9**	**−6.1**	−1.9	**−0.152**	**−0.325**	**−0.067**	**−0.256**
Historical	65.0	3.54	35.9	2.2	4.2	4.2	0.287	0.825	0.148	0.639
se	3.8	0.17	4.8	6.2	2.6	2.5	0.009	0.007	0.010	0.015

their standard errors. The standard errors of the autocorrelation coefficients were obtained with the jackknife method of Buishand and Beersma (1993). The standard errors of the standard deviations of the daily and the monthly values were calculated in a similar way following Buishand and Beersma (1996) and Beersma and Buishand (1999b; Chapter 2 of this thesis) respectively. Differences larger than twice the standard error of the historical data are referred to as statistically significant (this corresponds approximately to a two-sided test at the 5% level).

For the three analog models the average winter precipitation is underestimated due to the selection effects discussed in Section 3.3.2.*b*. The underestimation increases with the order of the model and becomes significant for the simulation with the analog 3 model based on historical circulation indices and for both simulations with the analog 5 model. The largest underestimation is 12.1%, whereas for the CNNR models the differences in monthly mean precipitation are not more than 6.1%. The standard deviations s_{M} and s_{D} are generally underestimated. The underestimation of the monthly standard deviations for precipitation and temperature in the analog models is about twice as large as in the CNNR models. In the analog models the underestimation of the daily standard deviation for precipitation is also somewhat larger. But for temperature, the underestimation of the daily standard deviations is at least two times smaller in the analog models. The biases in the lag 1 and lag 2 autocorrelation coefficients for precipitation and temperature are in the analog models about three times as large as in the CNNR models. But even for the simulations with the latter models the autocorrelation coefficients are significantly underestimated. The bias in the autocorrelation coefficients is the main cause for the underestimation of the standard deviations of the monthly values.

For the CNNR models the biases are generally somewhat larger for the simulations based on simulated circulation indices than for those based on historical indices. For the statistics in Table 3.1, the best performing analog model on the whole is analog 3. Its performance is, however, still below that of the weakest CNNR model (CNNR 2**C**).

Time series of the area-average winter precipitation totals and the area-average winter temperatures for the simulations conditional on the historical circulation indices are compared with the observed 1961–1995 winter area-average precipitation and temperature in Figures 3.3 and 3.4 respectively. For each simulation the coloured symbols represent the averages of the 28 runs in each winter. For the simulations with the CNNR 1 and analog 1 models the whiskers represent the range of the 28 runs. The skill score S in the figures is

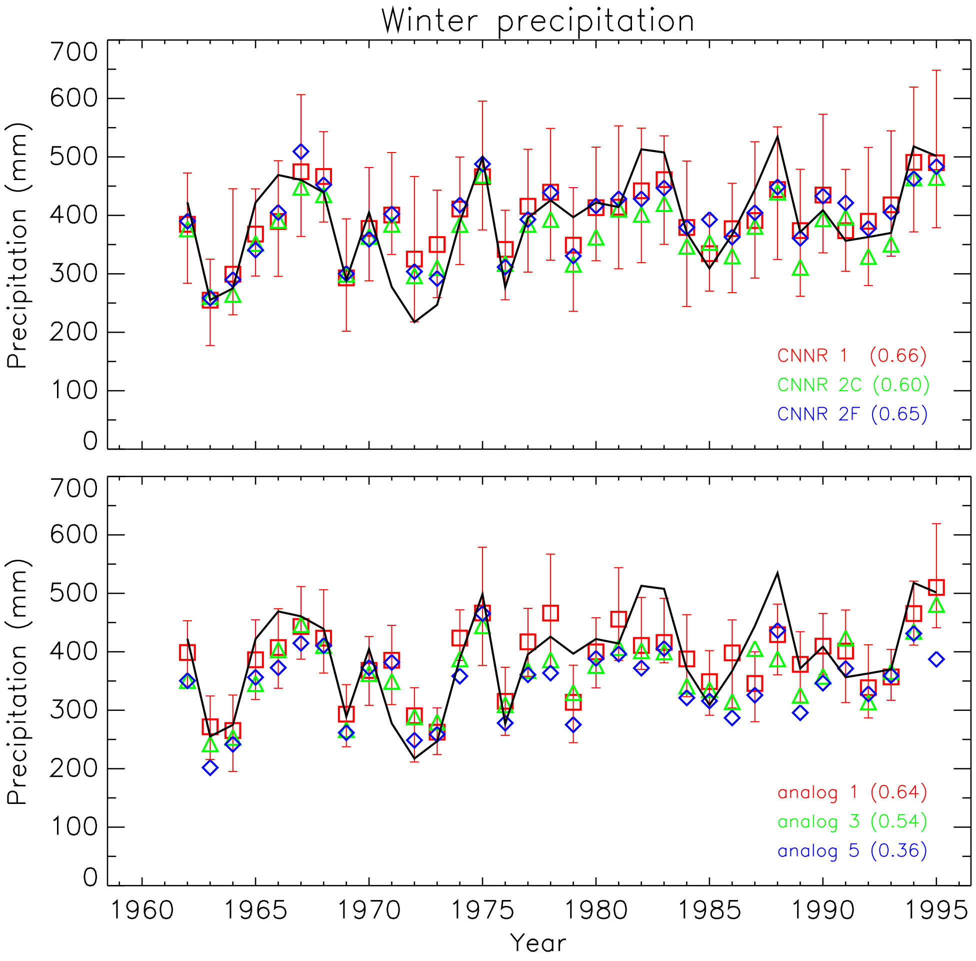

Figure 3.3. Observed and simulated area-average winter precipitation totals. (top) CNNR models; (bottom) analog models. Black line: historical values; coloured symbols: simulated values averaged over 28 runs; red whiskers: range of values in 28 runs (CNNR 1 and analog 1 models only). Values in parentheses are skill scores (see text).

defined as:

$$S = 1 - \sum_j (y_j - \hat{y}_j)^2 / \sum_j (y_j - \bar{y})^2 \,, \tag{3.7}$$

where the y_j are the historical winter precipitation totals or temperature averages, $\bar{y}$ is the overall historical average and the $\hat{y}_j$ are the simulated values for each winter, averaged over the 28 runs. Note that $S = 1$ for a perfect predictor, and $S = 0$ if $\bar{y}$ is taken as predictor.

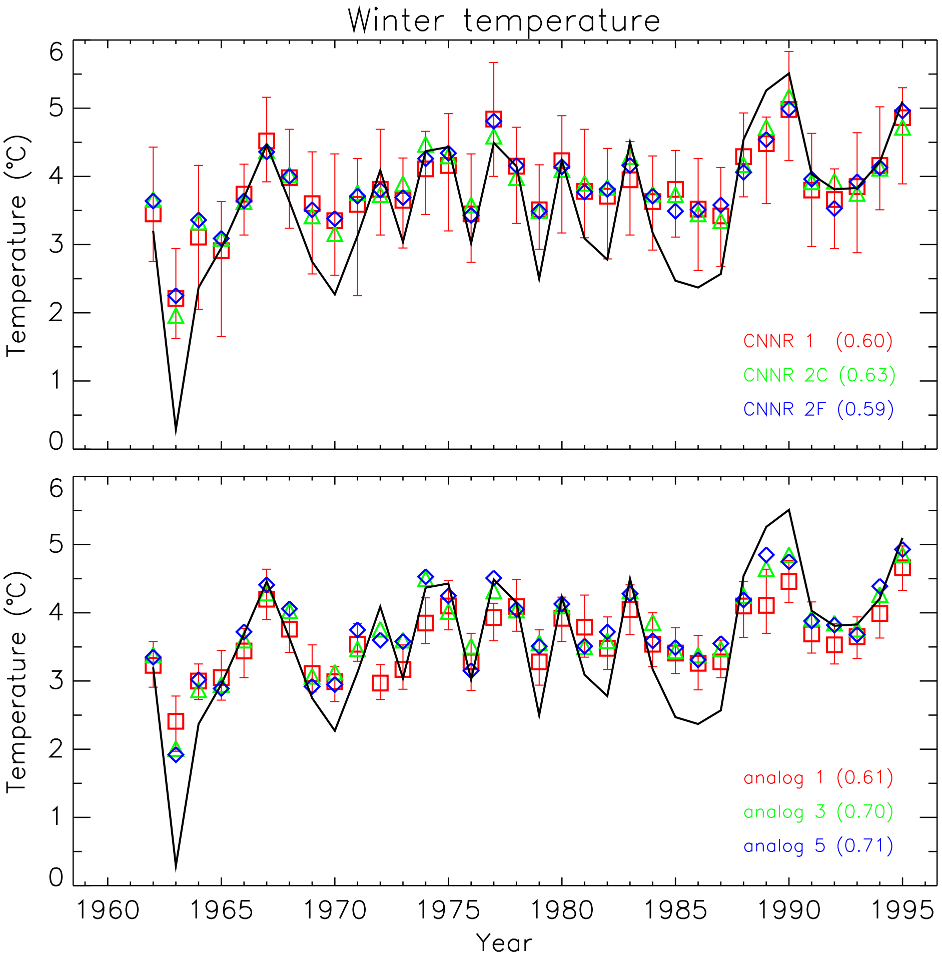

Figure 3.4. Observed and simulated area-average winter temperatures. Details as in Figure 3.3.

The temporal variation of the area-average winter precipitation is well described by the CNNR and analog 1 models, with S ranging between 0.60 and 0.66. The higher-order analog models exhibit less skill due to the significant underestimation of the mean precipitation amounts (see Table 3.1). This underestimation is also visible in Figure 3.3. For most winters the whiskers of the simulated precipitation amounts are considerably wider for the CNNR models than for the analog models, as shown for the CNNR 1 and analog 1 models. The difference in whisker width between the CNNR and analog models is even larger for the area-average winter temperature (Figure 3.4). The larger width of the whiskers in the CNNR models is likely due to a larger variation in the potential analogs (see Section 3.3.2.*b*).

The simulations overestimate the average temperature of the coldest winters (1963 and during the mid 1980s) and underestimate the temperature of the warm winters (around 1990). This conditional bias is weakest for the analog 5 model. As a result, this model has the highest skill score ($S = 0.71$). So for the analog models, the highest order is favourable for the predictive skill of the winter temperature, while the lowest order is favourable for the predictive skill of the winter precipitation due to the large underestimation of the mean precipitation in the higher-order models.

b. Selection effects

As a result of random sampling with replacement, some historical days will appear more frequently in a simulation run than other days. In the standard bootstrap such differences are purely random. Nearest-neighbour resampling may, however, also lead to a systematic underselection of certain days and an overselection of other days (Young, 1994). This explains for instance that the mean and the daily standard deviation s_{D} of the historical data are not necessarily reproduced in the simulations.

The selection effects of the simulations in this chapter are studied in the same way as in Buishand and Brandsma (2001). Let K_t be the number of times that day t ($t = 1, \ldots, 365J$) appears in a simulation run of J^* years. In the case of random sampling, K_t has a binomial distribution which can be can be approximated by a Poisson distribution with parameter $\nu = J^*/J$:

$$\Pr(K_t = r) \approx \frac{\nu^r e^{-\nu}}{r!}\,. \tag{3.8}$$

Note that the distribution of K_t does not depend on the use of a moving window. For nearest-neighbour resampling the number of historical days that is drawn r times can be compared with the number expected from the Poisson distribution with parameter ν. The latter equals $365J \times \Pr(K_t = r)$.

In Table 3.2 the frequency distributions of r for the 980-year simulations are compared with the frequencies for the Poisson distribution with $\nu = 28$. The frequency distributions are wider than the theoretical frequency distribution for random sampling as a result of selection effects. In the CNNR models the selection effect is slightly larger for the simulations based on historical circulation indices than for those based on simulated indices. Of the CNNR models CNNR 2**C** (second-order in the atmospheric circulation) has the largest selection effect. In the analog models the selection effect increases considerably with the order of the model. Thus, the more information of the evolution of the atmospheric circulation is used, the larger the selection ef-

Table 3.2. Number of historical days drawn r times in 980-year simulations compared with the number expected for the standard bootstrap. The largest number of times that a historical day is drawn is given in the last column.

	r							
Model	0	1-10	11-20	21-30	31-40	41-50	>50	$r_{\max}$
Historical circulation indices (1961–1995)								
CNNR 1	44	1096	2995	3857	2683	1220	880	152
CNNR 2**C**	155	1814	2830	2913	2301	1443	1319	131
CNNR 2F	37	1128	2949	3795	2671	1324	871	131
Analog 1	107	903	2787	3820	3004	1511	643	92
Analog 3	716	1993	2576	2413	1922	1378	1777	133
Analog 5	1078	2330	2498	2041	1528	1187	2113	179
Simulated circulation indices								
CNNR 1	13	1021	2999	4002	2636	1270	834	144
CNNR 2**C**	17	1432	3127	3272	2429	1394	1104	120
CNNR 2F	12	1135	2952	3809	2698	1314	855	127
Analog 1	0	224	2187	5620	3792	843	109	65
Analog 3	6	1335	3162	3254	2464	1520	1034	94
Analog 5	20	2305	3256	2523	1771	1218	1682	154
Bootstrap	0	1	928	7890	3797	158	1	-

fects tend to grow. Further, the differences in the selection effects between the simulations based on the historical circulation indices and those based on simulated indices are much larger than in the CNNR models. This is mainly due to the relatively large selection effect in the analog simulations conditional on historical circulation indices. Recall that in that case each simulation consists of 28 runs based on the same 35 years. In the analog models the nearest neighbours (or analogs) are determined by the circulation characteristics of the historical record only. A particular conditioning day therefore has in each of the 28 runs the same nearest neighbourhood, i.e., the same potential analogs. In the CNNR models the nearest neighbourhood of a particular conditioning day varies among the 28 runs, since the precipitation and temperature characteristics of the previously simulated day determine the potential analogs as well. This larger variation in potential analogs is probably responsible for the smaller selection effect.

Surprisingly, large differences in the selection effects do not necessarily lead to large differences in the standard deviations and autocorrelation coefficients. In particular for the analog 1 model the selection effect for the simulation based on historical indices is much larger than for the simulation based on simulated indices, but the differences between the second-order moments (Table 3.1) are relatively small. In the simulations with the largest selection effects, however, the average monthly precipitation amount is significantly underestimated, indicating an underselection of days with large rainfall and an overselection of relatively dry days.

c. *Temporal dependence of spatial patterns*

A resampling technique automatically preserves the spatial patterns of the daily precipitation and temperature fields, but it does not necessarily reproduce the dependence between the patterns of successive days. Two measures of the temporal dependence between the spatial patterns were considered: the pattern correlation of two days that lie l days apart (for precipitation, days with no rainfall were excluded), and the length of the difference vector of two days that are l days apart. The pattern correlation correlates the spatial fields relative to their respective spatial means (centered statistic). The reproduction of the pattern correlation and the difference vector of the precipitation and the temperature fields is presented in Table 3.3.

There is always an underestimation of the pattern correlation. The biases in the lag 1 precipitation pattern correlation are similar in both types of models. For temperature the biases are somewhat larger, in particular for the analog models. The biases in the lag 2 pattern correlation coefficients are about half of those in the lag 1 coefficients. The simulations based on simulated circulation indices have similar biases (not shown).

Most lagged difference vectors are overestimated. Overall, the overestimation is worse for temperature than for precipitation, and worse for the analog models than for the CNNR models. This overestimation is mainly due to the underestimation of the autocorrelation coefficients (Table 3.1). The underestimation of the latter also contributes to the underestimation of the pattern correlation. However, unlike the difference vector, the pattern correlation also depends on lagged cross-correlations between the daily temperatures (or daily precipitation) at different locations. From a first-order approximation of the expected value of the pattern correlation, it can be shown that the effect of the underestimation of the autocorrelation is partly compensated by biases in the lagged cross-correlations. As a result the pattern correlation looks less sensitive to biases in the temporal dependence than the difference vector.

Table 3.3. Differences in the lag 1 and lag 2 pattern correlations, $r_{\mathrm{p}}(1)$ and $r_{\mathrm{p}}(2)$ respectively, and the lengths of the lag 1 and lag 2 difference vectors, $d(1)$ and $d(2)$, between the 980-year simulations and the historical data for the winter (October–March), averaged over the 28 (35-year) runs. Absolute differences are given for $r_{\mathrm{p}}(1)$ and $r_{\mathrm{p}}(2)$, and percentage differences for $d(1)$ and $d(2)$. Bottom line: average historical (1961–1995) estimates. $r_{\mathrm{p}}(1)$ and $r_{\mathrm{p}}(2)$ are dimensionless, and $d(1)$ and $d(2)$ are given in mm for precipitation and in °C for temperature.

	$r_{\mathrm{p}}(1)$		$r_{\mathrm{p}}(2)$		$d(1)$		$d(2)$	
Model	P	T	P	T	P	T	P	T
Historical circulation indices (1961–1995)								
CNNR 1	−0.089	−0.128	−0.044	−0.065	6.2	18.4	3.2	3.4
CNNR 2**C**	−0.083	−0.121	−0.041	−0.070	0.5	22.0	−2.3	8.1
CNNR 2F	−0.087	−0.122	−0.040	−0.060	5.2	19.3	1.2	4.7
Analog 1	−0.100	−0.167	−0.040	−0.087	16.0	90.1	5.6	42.2
Analog 3	−0.089	−0.141	−0.039	−0.082	7.6	66.6	−1.2	30.6
Analog 5	−0.088	−0.145	−0.040	−0.082	3.7	61.9	−4.9	26.9
Historical	0.271	0.774	0.196	0.672	86.3	53.0	98.3	75.2

d. Dry spell counts and dry spell lengths

Table 3.4 presents the relative biases of the average number of dry days ($P = 0$ mm); the average number of dry spells (i.e., series of consecutive dry days); the average dry spell length DSL; the longest dry spell in a 35-year period $\mathrm{DSL}_{\max 35}$; and the bias of the lag 1 wet-dry autocorrelation coefficient $r_{\mathrm{wd}}(1)$. These statistics were again calculated for the winter half-year (October–March). The number of dry days is nearly correct in the CNNR and first-order analog models, but the third and fifth-order analog models overestimate this by 4 to 6%. This overestimation partly explains the underestimation of the mean precipitation (see Table 3.1) and is due to the earlier discussed selection effects. In both types of models the number of dry spells is overestimated: in the CNNR models somewhat more than 10%, and in the analog models more than 30%. As a result the average spell length is under-

Table 3.4. Differences in the number of dry days, the number of dry spells, the average dry spell length, DSL, the maximum dry spell length in 35 years, $\mathrm{DSL}_{\mathrm{max}\,35}$, and the lag 1 wet-dry autocorrelation coefficient, $r_{\mathrm{wd}}(1)$, between the 980-year simulations and the historical data in winter (October–March). Absolute differences are given for $r_{\mathrm{wd}}(1)$, and relative differences (%) for the other statistics. Absolute and relative differences are averaged over the 25 stations and the 28 runs of 35 years. Bottom line: historical (1961–1995) estimates; counts are per winter, lengths are in days and $r_{\mathrm{wd}}(1)$ is dimensionless.

Model	No. of dry days	No. of dry spells	DSL	$\mathrm{DSL}_{\mathrm{max}\,35}$	$r_{\mathrm{wd}}(1)$
Historical circulation indices (1961–1995)					
CNNR 1	−1.6	13.3	−13.1	−8.8	−0.082
CNNR 2**C**	1.1	12.5	−10.1	−4.7	−0.074
CNNR 2F	−1.0	13.2	−12.4	−5.3	−0.081
Analog 1	1.1	38.1	−26.8	−28.1	−0.227
Analog 3	3.8	33.7	−22.3	−16.0	−0.198
Analog 5	5.8	35.4	−21.8	−15.6	−0.208
Historical	83.8	26.6	3.1	24.4	0.411

estimated: in the CNNR models slightly more than 10%, and in the analog models slightly more than 20%. Apart from the analog 1 model, the relative underestimation of $\mathrm{DSL}_{\mathrm{max}\,35}$ is considerably smaller than that of the average dry spell length.

Like the lag 1 autocorrelation coefficient of the daily precipitation amounts in Table 3.1, the lag 1 wet-dry autocorrelation coefficient is significantly underestimated in all simulations (the jackknife *se* of the lag 1 wet-dry autocorrelation for the historical data is 0.010). The underestimation of both $r_{\mathrm{wd}}(1)$ and DSL can be understood from the relation (Buishand, 1978):

$$r_{\mathrm{wd}}(1) = 1 - 1/\mathrm{DSL} - 1/\mathrm{WSL}\,, \tag{3.9}$$

with WSL the average wet spell length. If, as in Table 3.4, the number of dry spells is overestimated then the number of wet spells is also overestimated (since by definition a wet spell follows a dry spell and vice versa). As a result both DSL and WSL are underestimated and consequently $r_{\mathrm{wd}}(1)$ is underestimated.

In terms of spell counts and spell lengths there is very little difference between the simulations based on historical circulation indices and those on simulated circulation indices (not shown).

3.3.3 Long-duration simulations

Monte Carlo techniques enable us to produce synthetic time series of precipitation and temperature that are much longer than the observed records. Using such long-duration simulations as input into a rainfall-runoff model offers the opportunity to get more reliable estimates of the probabilities of extreme river discharges.

The 980-year simulations conditional on simulated circulation indices that were split into 28 independent 35-year records earlier are now used as single long-duration simulations. The distribution of the extreme 10-day area-average precipitation amounts in these simulations is examined in this section. An interval of 10 days was chosen because flooding of the river Rhine is often caused by large rainfall in winter over periods of about 10 days. An analysis of the January 1995 Rhine flood in Germany (Fink et al., 1996) demonstrated that in parts of the basin the monthly totals were more than three times as large as the climatological averages and that about 70 to 80% of these high monthly totals fell within a period of only 10 days.

The largest 10-day area-average precipitation amounts (average of all 25 stations) in each winter (October–March) were extracted from the 980-year simulations and the 35-year historical data. Figure 3.5 presents Gumbel plots of these winter maxima (the horizontal scale in these plots is such that the ordered maxima follow a straight line in the case of a Gumbel distribution). The Gumbel plots show that much larger 10-day area-average amounts (up to 35%) are simulated than the historical (1961–1995) maximum. Such unprecedented rainfall events can be very useful for hydrological design. Figure 3.5 further shows that all models underestimate the 10-day area-average precipitation amounts for return periods between 5 and 20 years. The underestimation is largest for the analog models and the CNNR **2C** model. The analog models systematically underestimate the 10-day area-average precipitation amounts for all return periods, indicating that these models are not very suitable for applications where the extreme multi-day precipitation amounts are of interest.

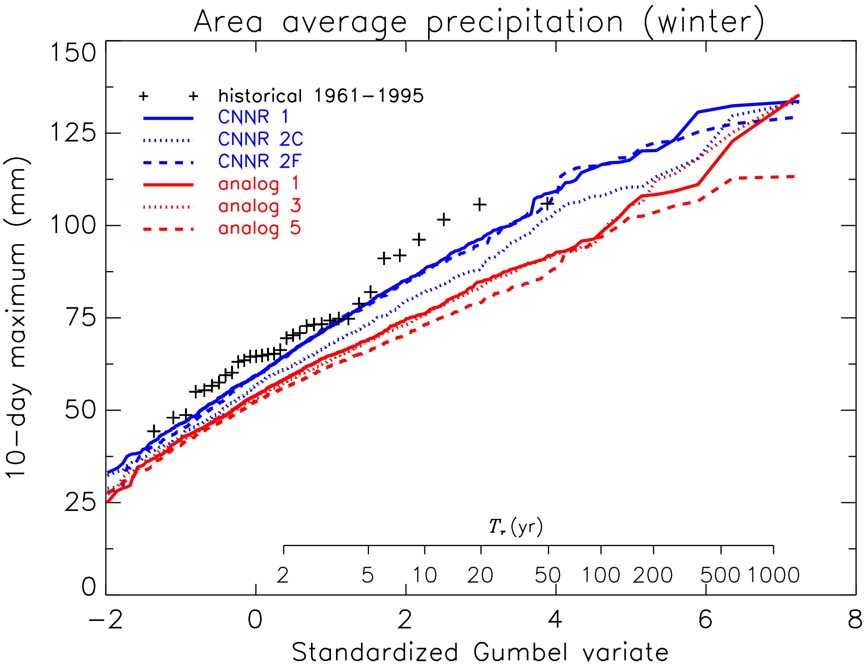

Figure 3.5. Gumbel plots of the 10-day winter precipitation maxima for 980-year simulations conditional on simulated circulation indices and for the historical 1961–1995 data. T_r represents the return period in years.

3.4 Summary and conclusions

Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation has been studied for 25 stations in the German part of the Rhine basin using CNNR and a stochastic version of the analog method. To fully explore the differences between the CNNR and the analog models the simulations were divided into simulations conditional on historical time series of circulation indices and simulations conditional on simulated time series of circulation indices. A second resampling model was used to generate long-duration time series of circulation indices.

All conditional simulation models have a tendency to underestimate the standard deviations and autocorrelation coefficients of daily precipitation and temperature and the standard deviations of the monthly precipitation totals and the monthly mean temperatures. In general, the underestimation is larger for the analog models than for the CNNR models, with the exception of the underestimation of the standard deviation of the daily temperatures, which is much smaller in the analog models. The number of days that is never or almost never selected is relatively large for the analog models conditional on historical circulation indices. The mean precipitation amounts are significantly underestimated in the simulations with the largest selection effects. All simulations conditional on the historical circulation indices underestimate the average temperature in the warmest winters and overestimate the average temperature in cold winters. The reproduction of the temporal dependence of the spatial patterns of precipitation and temperature by both types of models turned out favourably for the CNNR models although the differences were not as large as for the univariate autocorrelations. The biases in the dry spell counts and the dry spell lengths are for the analog models often more than twice as large as for the CNNR models. The CNNR models also reproduced the extreme-value distribution of the 10-day area-average winter precipitation amounts better than the analog models. Despite an underestimation of the quantiles of this distribution, the highest 10-day area-averages were in most 980-year simulations much larger than the highest observed 10-day area-average.

Since, the observed weather of historical days is resampled, the dependence between daily precipitation at different sites and the dependence between daily precipitation and temperature is automatically preserved. These dependencies often have a complicated structure, which may not be adequately described by parametric models. For many hydrological applications the spatial dependency is of crucial importance. This makes multivariate resampling models particularly suitable for hydrological purposes. The comparison between the analog models and the CNNR models, however, demonstrates that besides the circulation characteristics of the target day, also the precipitation and temperature characteristics of the previously simulated day should be taken into account.

A few potential limitations of the methodology can also be identified. The method is rather data intensive and, resampling of multivariate data may become problematic if data are missing (which is quite common in observational records). The method does not produce new daily precipitation and temperature fields but merely reshuffles the historical days to form realistic new sequences of those fields. As a result daily rainfall amounts cannot be larger than those observed. Similarly, daily temperatures for a particular location cannot be lower or higher than the observed minimum or maximum value for that location. The latter limitation may seriously bias the results of CNNR in climate change applications. To overcome this limitation, Lall and Sharma (1996) suggested to evaluate the means of the required variables first from the selected nearest neighbours and then to perturb these values by a residual using nearest-neighbour resampling. For conditional simulation on atmospheric predictors this strategy may need extension to allow for predictor values outside their range in the historical data.

It should further be noted that several studies of climate change simulations with General Circulation Models (GCMs) have revealed that changes in precipitation usually cannot be explained by changes in the atmospheric circulation alone. Consequently it becomes recognized that the simulation of precipitation should also be conditioned on (large-scale) predictors directly related to the atmospheric moisture and the temperature (for an overview see Giorgi et al., 2001). For similar reasons the simulation of temperature should include predictors like the large-scale (2 m) temperature, geopotential height or the thickness of an atmospheric layer (e.g., Huth et al., 2001; Benestad, 2002).

Acknowledgments. The authors would like to thank two anonymous reviewers for their useful comments. The UK Meteorological Office gridded mean sea level pressure 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 stations were made available by the Deutscher Wetterdienst via the International Commission for the Hydrology of the Rhine Basin (CHR/KHR). The work was performed in cooperation with the Institute for Inland Water Management and Waste Water Treatment (RIZA).

Chapter 4

Joint probability of precipitation and discharge deficits in the Netherlands

Jules J. Beersma and T. Adri Buishand, 2004

Published in *Water Resources Research*, **40**, W12508,
doi:10.1029/2004WR003265

Abstract

The Netherlands are situated at the downstream end of the river Rhine. A large part of the country can be supplied with water from the river in the case of precipitation deficits. For drought assessment it is therefore necessary to consider the joint distribution of precipitation and discharge deficits. A transformed bivariate normal distribution as well as a bivariate Gumbel distribution are fitted to this data. In addition, nearest-neighbour resampling is used to estimate joint probabilities of precipitation and discharge deficits. Both the reproduction of the marginal distributions and the dependence structure are explored. It is found that the transformed bivariate normal distribution underestimates the probability that both the precipitation and discharge deficit are extreme due to its asymptotic independence. Nearest-neighbour resampling also underestimates this probability, mainly because the upper tails of the marginal distributions are not properly reproduced by the simulations. From the two fitted bivariate distributions a novel bivariate distribution is constructed with transformed normal marginals and a logistic Gumbel depen-

dence structure, which gives the best description of the upper tail of the joint distribution. The use of a failure region based on economic damage rather than on joint exceedances considerably reduces the differences between the probabilities of drought from the various bivariate models.

4.1 Introduction

A large part of the Netherlands is situated in the delta of the river Rhine, the largest river in northwestern Europe (drainage area 185 000 km^2). The Rhine rises in the Swiss Alps and flows via France and Germany to the Netherlands, where it divides a number of times. The Rhine plays a major role in the overall water balance of the Netherlands; the amount of Rhine water that flows through the Netherlands is on average twice as large as the amount that the country receives as precipitation (Middelkoop and van Haselen, 1999). As a result, large parts of the country can be supplied with water from the river in the case of precipitation deficits.

This chapter addresses the probability of drought in the Netherlands. For droughts with a large economic impact, it is important to consider the joint distribution of the precipitation deficit in the Netherlands and the discharge deficit of the river Rhine. Two approaches are compared to estimate joint probabilities: (i) fitting bivariate distributions to the historical data, and (ii) time series simulation.

In the first approach a transformed bivariate normal distribution and a limiting bivariate Gumbel distribution are used. Both bivariate distributions have been applied in the water resources literature (Leytham, 1987; Kroll and Stedinger, 1998; Yue et al., 1999; Yue, 2001; Shiau, 2003), but a thorough comparison is lacking. Apart from differences between the marginal distributions, the dependence structure of a limiting bivariate Gumbel distribution is quite different from that of the classical bivariate normal distribution, in particular regarding the joint occurrence of large values. To assess the suitability of the dependence structure of these bivariate distributions new diagnostics from the statistical literature on multivariate extremes (Ledford and Tawn, 1996; Coles et al., 1999) are used. The adequacy of the fit of the bivariate distributions is further explored by comparing theoretical and empirical joint exceedance probabilities.

In the second approach nearest-neighbour resampling is used to generate a long sequence (10^5 years) of precipitation and discharge deficits. This resampling technique has successfully been applied to simulate time series of river flows (Lall and Sharma, 1996) and weather variables (Young, 1994; Rajagopalan and Lall, 1999; Buishand and Brandsma, 2001). In the nearest-neighbour resampling procedure the variables of interest (which in our case include precipitation, evaporation and river discharge) are sampled simultaneously with replacement from the historical data. A convenient characteristic of resampling is that no assumptions have to be made about the underlying distributions of each of the variables and of the dependencies between those variables.

The sensitivity of joint probabilities to the form of the marginal distributions and the dependence structure is discussed. Besides the probability of joint exceedances, attention is given to the probability associated with a failure region based on economic damage.

In Section 4.2 the historical data are described, and the precipitation deficit in the Netherlands and the discharge deficit of the Rhine are defined. Different probability distributions for the precipitation deficit are compared in Section 4.3. Section 4.4 presents probability distributions for the discharge deficit. The joint distribution of the precipitation and discharge deficits is discussed in Section 4.5. Return periods for joint exceedances estimated from the fitted bivariate distributions and from nearest-neighbour resampling are given for a number of extreme years in the historical record. In Section 4.6 the concept of a failure region, based on economic damage, as an alternative for the classic joint exceedance is discussed and illustrated with the same historical years. Section 4.7 concludes with a summary and a discussion of the results.

4.2 Drought characteristics

Two variables which describe drought in the Netherlands are considered here; the country-average precipitation deficit and the discharge deficit of the river Rhine. The precipitation deficit is defined as the cumulative difference between precipitation and grass reference evaporation, from April, 1 onward. When the precipitation deficit becomes negative it is reset to zero. The annual maximum precipitation deficit is the largest precipitation deficit that occurs during the summer half-year (April–September). Both for precipitation and evaporation daily values were available for the period 1906–2000, giving 95 independent annual maximum precipitation deficits. For practical reasons and for efficiency of the resampling procedure all daily data were converted into decades of days prior to the analysis. Decades of days were obtained by dividing each calendar month into three decades; the first two decades in a month always represent 10 days and the third decade represents the remaining days. Each year thus contains 36 decades.

Average precipitation for the Netherlands was obtained by averaging the precipitation sums from 13 stations spread over the country. The grass refer-

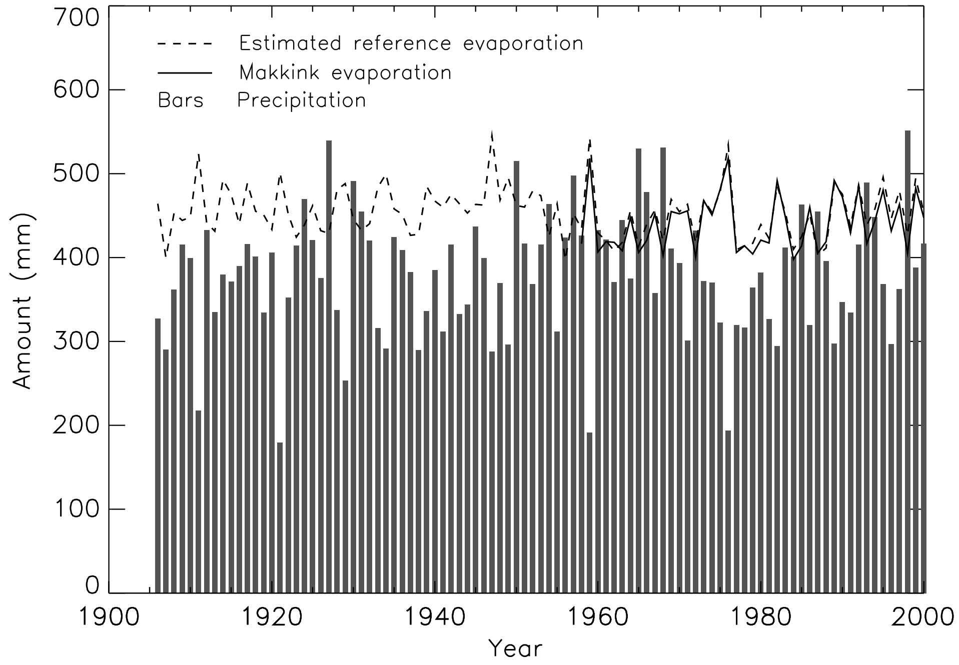

Figure 4.1. Evaporation and precipitation in the Netherlands in the summer half-year (April–September) for the period 1906–2000. Estimated reference and Makkink evaporation are explained in the text.

ence evaporation was derived from temperature and sunshine duration at De Bilt using the Makkink formula (e.g., de Bruin and Stricker, 2000). The global radiation in that formula was estimated from an empirical relation between global radiation and sunshine duration due to Frantzen and Raaff (1982). Figure 4.1 presents time series of precipitation and estimated reference evaporation for the summer half-years of the period 1906–2000. From 1958 onward the values of the original Makkink evaporation are also given. It can be seen that these values are close to the estimates used in this study. For most years the reference evaporation is larger than precipitation, giving rise to a precipitation deficit. A precipitation deficit also builds up during dry periods in wet summer half-years. Figure 4.1 further shows that the driest years (1911, 1921, 1959 and 1976) have above normal reference evaporation.

The discharge deficit of the river Rhine was based on discharge measurements at the German-Netherlands border (gauging station Lobith). Only decades for which the discharge was below a threshold of 1800 m^3 s^{-1} contribute to the discharge deficit. For those decades the (nonnegative) difference between the threshold and the discharge is added to the discharge deficit. The

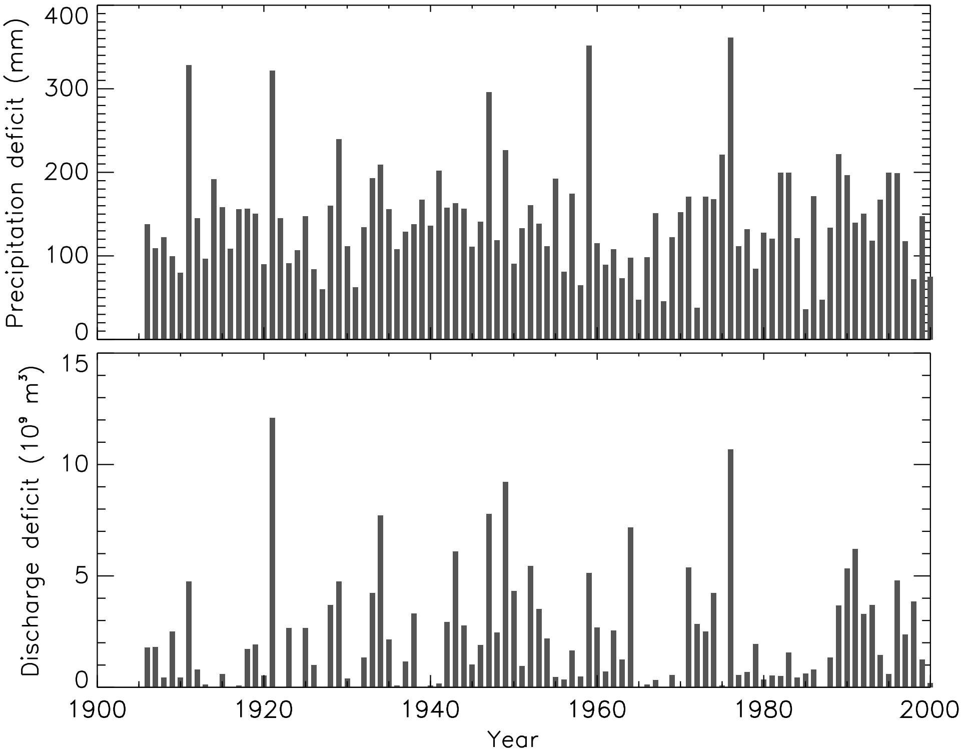

Figure 4.2. (top) Maximum precipitation deficit and (bottom) discharge deficit in the summer half-year (April–September) for the period 1906–2000.

discharge deficit was also calculated for the summer half-year and was available for the same period (1906–2000) as the precipitation deficit. A threshold of 1800 m^3 s^{-1} roughly corresponds to the 20% quantile of the decade average discharge during the summer half-year. In 8 years the discharge is never below this threshold giving a zero discharge deficit. A lower threshold would result in many more years with a zero discharge deficit. Figure 4.2 presents time series of the maximum precipitation and discharge deficits for the period 1906–2000. No visible trends are found in these deficits during the past century. As expected, there is a positive correlation between the precipitation and discharge deficits.

4.3 Probability distributions for the precipitation deficit

Two distributions were fitted to the largest precipitation deficit in each year; the Gumbel distribution and the lognormal distribution, where the Gumbel distribution is given by

$$F(x) = \Pr(X \leq x) = \exp\left[-e^{-(x-\mu)/\sigma}\right] . \tag{4.1}$$

The parameters of the distributions were estimated by the maximum likelihood (ML) method. In this section the fitted distributions are compared with a simulated distribution based on nearest-neighbour resampling of historical precipitation (P), evaporation (E) and discharge (Q) data. Conditions are imposed on the resampling process to reproduce the temporal dependence and the annual cycle of these variables as well as possible. A detailed description of the resampling model is given in Appendix D. With the resampling model, 10^5 years (i.e., 36×10^5 decades) were simulated to empirically estimate the probability distribution of the (annual maxima of the) precipitation deficit and the discharge deficit. The simulated deficits are occasionally larger than those in the historical record as a result of reshuffling of the historical decade data (Appendix D).

Figure 4.3 presents a Gumbel probability plot of the maximum precipitation deficits. The horizontal axis in this plot is chosen such that the Gumbel distribution corresponds to a straight line. It can be seen that the fitted lognormal distribution has a heavier upper tail than the fitted Gumbel distribution. If one is interested in the exceedance probabilities of the largest historical values, one might argue to use the lognormal distribution since this distribution performs best in this range of the data. The fitted distributions were subjected to the Anderson-Darling (A-D) test (as by Stephens, 1986a) and the 'probability plot correlation coefficient' (ppcc) test (Vogel, 1986). These tests were selected because they are known to be sensitive to deviations in the upper tail. Both tests give for the lognormal distribution a significant result at the 5% level, but not at the 1% level, while the Gumbel distribution passes both tests at the 5% level. Thus, although the lognormal distribution describes the tail of the distribution better, these tests indicate that over the whole domain the lognormal distribution does not properly fit the data while the Gumbel distribution does.

The curvature in the plot for the simulated data from the resampling model is more or less in agreement with that for the historical data, only the upper tail of the simulated distribution seems to be somewhat too light. The simulated distribution suggests that the precipitation deficit is limited which is

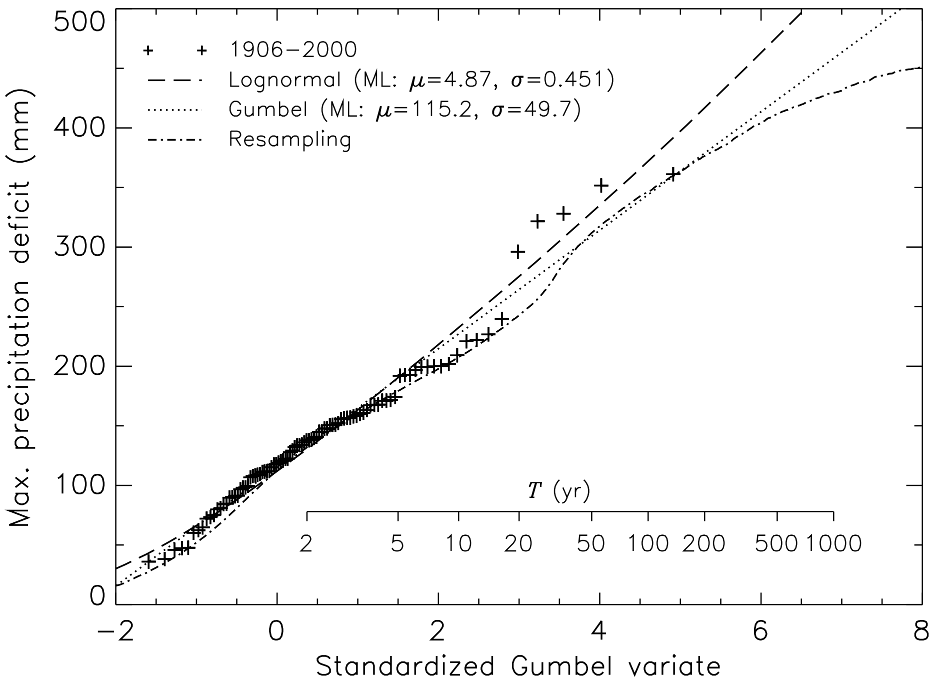

Figure 4.3. Ordered historical annual maximum precipitation deficits, the fitted Gumbel and lognormal distributions, and the simulated distribution from a resampling model. The parameters μ and σ for the lognormal distribution refer to the mean and standard deviation of the underlying normal distribution.

true in fact. When there would be no precipitation at all during the summer half-year, the precipitation deficit is completely determined by the reference evaporation. Under present day climate conditions (in particular with respect to temperature and global radiation), the largest precipitation deficit is estimated to be 600 mm. For the fitted lognormal distribution (heaviest tail) a precipitation deficit of 600 mm is exceeded on average once in 2800 years.

4.4 Probability distributions for the discharge deficit

Probability distributions for the discharge deficit were obtained in a similar way as for the precipitation deficit. A Gumbel distribution was fitted to the annual discharge deficits. A sqrt-normal distribution (which assumes that the square root of the data are normally distributed) was also fitted. The choice of this distribution was based on the ML estimate of the optimal power

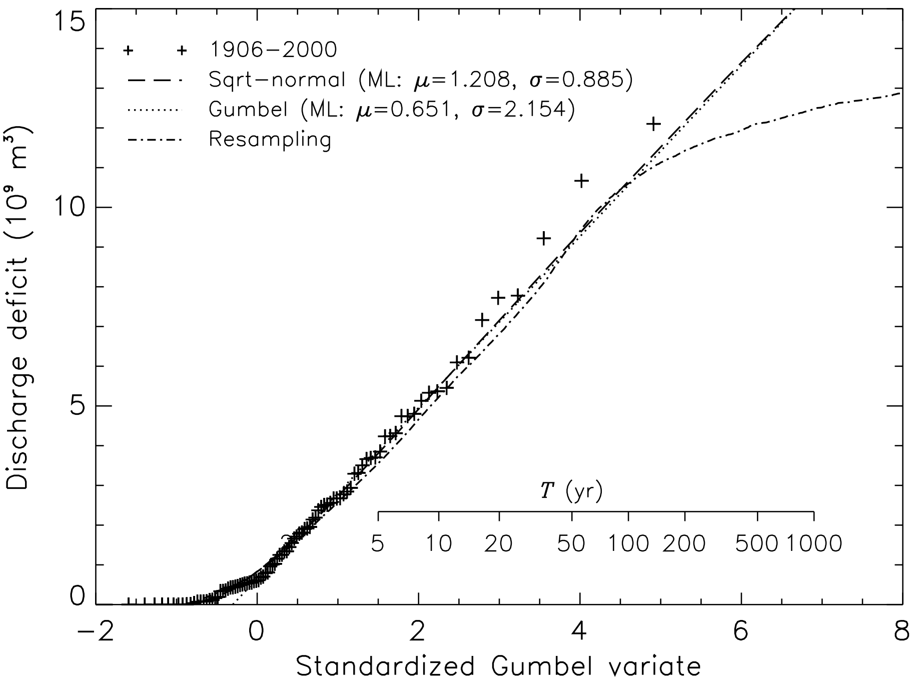

Figure 4.4. Ordered historical annual discharge deficits, the fitted Gumbel and sqrt-normal distributions and the simulated distribution from a resampling model. The parameters μ and σ for the sqrt-normal distribution refer to the mean and standard deviation of the underlying normal distribution.

transformation in the Box-Cox family (Shumway et al., 1989).

To avoid a large influence of small values of the discharge deficit on the estimated parameters the sample was censored at a low threshold of 0.03×10^9 m^3 in the fit of the sqrt-normal distribution and at 0.6×10^9 m^3 in the case of the Gumbel distribution. For data below the threshold only the information that they are smaller than the threshold is then used rather than their actual values. The parameters were estimated by the ML method, see e.g., Shumway et al. (1989) for a transformed normal distribution and Leese (1973) for the Gumbel distribution.

Figure 4.4 presents a Gumbel probability plot of the discharge deficits. For values larger than 1.0×10^9 m^3, the fitted Gumbel and sqrt-normal distributions are nearly indistinguishable. The simulated distribution from the resampling model agrees well with the two fitted distributions for return periods up to about 100 years. For longer return periods they start to deviate. The discharge deficit is also bounded. Applying the lowest observed discharge

(725 m^3 s^{-1}) to the whole summer half-year leads to a discharge deficit of 17×10^9 m^3. The largest historical discharge deficit (of 1921) amounts to 71% of this practical upper limit, and the largest simulated discharge deficits in the resampling model are about 85% of this limit.

Because of the censoring the goodness-of-fit tests used in the previous section can not be applied. Both the Gumbel and the sqrt-normal distribution pass the adapted ppcc test for censored data in Stephens (1986b) at the 5% level.

4.5 Bivariate probability distributions

So far univariate probabilities have been considered. In the introduction it was noted that from a drought impacts point of view it is more interesting to look at joint exceedance probabilities. Drought events that have the largest economic impact are those events that have both a large precipitation deficit and a large discharge deficit. The latter makes compensation of the local water shortage by water from elsewhere in the Rhine basin very difficult.

A logical way to proceed is to combine the univariate (marginal) probability distributions into a bivariate probability distribution. In the case that the maximum precipitation deficit is described by a lognormal distribution and the discharge deficit by a sqrt-normal distribution it would be natural to consider the bivariate normal distribution. The joint density of the standardized transformed precipitation and discharge deficits is then given by:

$$\phi_2(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right] , \qquad (4.2)$$

where ρ is the correlation coefficient of the transformed values.

A family of bivariate extensions of the Gumbel distribution is provided by the theory of multivariate extremes (e.g., Tawn, 1988; Coles, 2001). This family can be represented as

$$F(x,y) = \Pr(X \le x, Y \le y) = \exp\left[-(e^{-x} + e^{-y}) A\left(\frac{e^{-x}}{e^{-x} + e^{-y}}\right)\right] , \qquad (4.3)$$

where $A(\cdot)$ is the dependence function. $A(w) = 1$ implies that X and Y are independent, whereas $A(w) < 1$ implies that X and Y are positively associated ($0 < w < 1$). Perfect dependence, e.g., $\Pr(X = Y) = 1$, corresponds with $A(w) = \max[w, (1-w)]$. Note that $A(0) = A(1) = 1$ both for independent and dependent Gumbel variables.

Several parametric models for $A(w)$ have been proposed in the literature (e.g., Kotz and Nadarajah, 2000). A popular one is the (symmetric) logistic dependence model:

$$A(w) = \left[w^{1/\alpha} + (1-w)^{1/\alpha}\right]^{\alpha}, \quad 0 \leq w \leq 1\,;\; 0 \leq \alpha \leq 1 \tag{4.4}$$

which leads to

$$F(x, y) = \exp\left[-(e^{-x/\alpha} + e^{-y/\alpha})^{\alpha}\right], \tag{4.5}$$

where α characterizes the strength of the dependence between X and Y; $\alpha = 1$ corresponds with independence and $\alpha = 0$ with perfect dependence. The correlation between X and Y equals $1 - \alpha^2$ (Tiago de Oliveira, 1980). Yue (2001) used the logistic Gumbel model to describe the joint distribution of storm peaks and amounts and Shiau (2003) applied this model to extreme flood events (peaks and volumes).

The dependence structure of the bivariate normal distribution differs from that of bivariate Gumbel distributions. A classical result for the bivariate normal distribution with $\rho < 1$ is that its components are asymptotically independent (Sibuya, 1960). For the standard bivariate normal distribution in equation (4.2) asymptotic independence implies that:

$$\lim_{u \to \infty} \Pr(Y > u \mid X > u) = 0. \tag{4.6}$$

A loose interpretation of this is that the probability that Y is extreme given that X is extreme tends to zero, or in other words, extreme values of X and Y do not occur simultaneously. For the bivariate logistic Gumbel distribution in equation (4.5), however, $\Pr(Y > u \mid X > u)$ tends to $2 - 2^{\alpha}$, and this distribution is therefore asymptotically dependent if $\alpha < 1$. Note that asymptotic dependence holds for all limiting bivariate extreme value distributions (including the logistic Gumbel distribution).

In this section the dependence structure of the data is investigated first. Then the observed joint exceedance probabilities are compared with the theoretical ones from the bivariate models, and with those from the data simulated by nearest-neighbour resampling.

4.5.1 Dependence structure

Dependence measures for bivariate extremes have been discussed by Coles et al. (1999). To remove the influence of the marginal distributions the variables X and Y are transformed to standard uniform variables, via $U = F_X(X)$ and $V = F_Y(Y)$. The joint distribution of U and V is called a copula. It contains all information about the association between X and Y. Copulas have been applied recently in bivariate hydrological frequency analysis by Favre et al. (2004). For the data $(x_i,\ y_i),\ i = 1, \ldots,\ N$ the influence of the marginal distributions can be removed in a similar way using the empirical distribution functions:

$$\begin{aligned} u_i &= \hat{F}_X(x_i) = \frac{\#\ x_j\text{'s} \leq x_i}{N+1} \\ v_i &= \hat{F}_Y(y_i) = \frac{\#\ y_j\text{'s} \leq y_i}{N+1}\,. \end{aligned} \qquad (4.7)$$

Buishand (1984) introduced a measure of dependence to estimate the interstation dependence in the extremes of daily precipitation. A slight modification of this dependence measure is the quantity $\chi(u)$ suggested by Coles et al. (1999):

$$\chi(u) = 2 - \frac{\ln \Pr(U < u,\ V < u)}{\ln \Pr(U < u)} \quad \text{for } 0 < u \leq 1\,. \qquad (4.8)$$

Independence corresponds with $\chi(u) = 0$ and perfect dependence with $\chi(u) = 1$. For the bivariate Gumbel distributions $\chi(u) = 2 - 2A(1/2)$, which reduces to $\chi(u) = 2 - 2^\alpha$ for the logistic dependence model. Further, for sufficiently large u,

$$\chi(u) \sim \Pr(V > u \mid U > u). \qquad (4.9)$$

For asymptotically independent distributions like the bivariate normal distribution $\chi(u) \to 0$ as $u \to 1$. The measure $\chi(u)$ is not influenced by a monotonic increasing transformation of the data such as the log and sqrt transformation applied to the precipitation and discharge deficits to achieve normality.

An empirical estimate of $\chi(u)$ can be constructed by substituting empirical estimates of the probabilities in the right-hand side of equation (4.8). Figure 4.5 presents such estimates of $\chi(u)$ for the historical and simulated data and the theoretical values for the fitted bivariate distributions. The parameters ρ and α in these distributions were estimated by the ML method, taking into account the censoring of low discharge deficits (Appendix E). The figure shows that $\chi(u)$ is almost constant for the historical precipitation and discharge deficits. For the resampled data, the average level of $\chi(u)$ is slightly lower, with a weak minimum near $u = 0.5$. For large u, the estimates of $\chi(u)$ for the historical and the simulated data are more in line with the theoretical values for the bivariate Gumbel distribution than those for the bivariate normal distribution. For the latter $\chi(u)$ gradually decreases, but for u near 1 it abruptly drops to zero. From a physical point of view, this behavior is not very realistic since a severe drought typically extends over a large area and will thus affect the precipitation in the Netherlands as well as in the

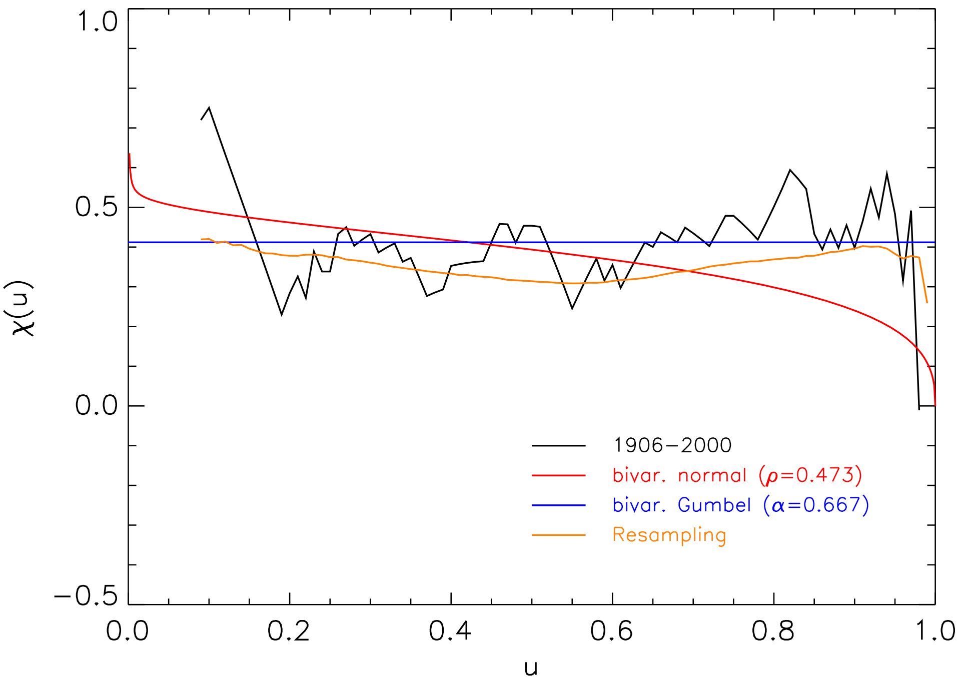

Figure 4.5. Dependence measure $\chi(u)$ for the historical and simulated data and for the fitted bivariate distributions. The value 0.667 for the parameter α in the bivariate Gumbel model corresponds to a correlation coefficient of 0.555.

upstream Rhine catchment. The use of the multivariate normal distribution to describe droughts over large geographic areas was already questioned by Leytham (1987). He observed that this distribution underestimated the frequency of simultaneous low precipitation amounts or low river flows at widely separated sites.

The question whether the data are asymptotically dependent or not can be investigated further by calculating for each year $T_i = \min(-1/\ln u_i,\ -1/\ln v_i)$. For large z, the probability that $T_i > z$ can be approximated by the Pareto distribution (Ledford and Tawn, 1996):

$$\Pr(T_i > z) \approx cz^{-1/\eta}\,, \tag{4.10}$$

where c and η are the scale and shape parameters. For the bivariate Gumbel distribution $\eta = 1$, whereas for asymptotically independent data $\eta < 1$; $\eta = (\rho + 1)/2 = 0.74$ for the bivariate normal distribution. The parameter η can be estimated from the k largest values of T_i using the ML method (Hill, 1975).

A quantile plot suggests that k can be taken as large as 70. This results in $\hat{\eta} = 1.12$ with a standard error of 0.13, which supports the bivariate Gumbel distribution.

4.5.2 Symmetry of dependence

Both the bivariate normal distribution and the bivariate logistic Gumbel distribution have a symmetric dependence structure. Here symmetry implies that the dependence structure is such that the joint probabilities are unchanged when X and Y are interchanged. For a limiting bivariate Gumbel distribution this holds only if $A(w)$ is symmetrical about $w = 1/2$. This can be explored by estimating $A(w)$ with a nonparametric method. Pickands (1981) observed that $Z(w) = \min\left[e^{-X}/(1-w),\ e^{-Y}/w\right]$ has an exponential distribution with mean $1/A(w)$, for each $w \in (0,1)$. Transforming again the original variables to standard uniform variables, the following nonparametric estimate of $A(w)$ is obtained (Hall and Tajvidi, 2000):

$$\hat{A}(w) = n\left[\sum_{i=1}^{n} Z_i(w)\right]^{-1}, \tag{4.11}$$

where

$$Z_i(w) = \min\left[\frac{\ln u_i}{(1-w)\overline{\ln u}},\ \frac{\ln v_i}{w\overline{\ln v}}\right] \quad (0 \le w \le 1)\,,$$

with $(u_i,\ v_i)$ defined in equation (4.7) and $\overline{\ln u}$, $\overline{\ln v}$ the arithmetic means of $\{\ln u_i\}$, $\{\ln v_i\}$ respectively. For discharge deficits equal zero ($y_i = 0$), the numerator of v_i in equation (4.7) is based here on their average rank, i.e., $[(\#\ y_j\text{'s} = 0) + 1]/2$. Note that $\hat{A}(0) = \hat{A}(1) = 1$ (in agreement with $A(0) = A(1) = 1$).

Figure 4.6 compares $\hat{A}(w)$ for the historical and simulated data with $A(w)$ for the fitted logistic dependence model. Apart from the bump around $w = 0.75$, which is partly due to the zero discharge deficits, $\hat{A}(w)$ is nearly symmetrical. The figure shows that the overall level of $\hat{A}(w)$ agrees with $A(w)$ for the logistic dependence model with $\alpha = 0.667$. The minimum of $\hat{A}(w)$ for the resampled data is somewhat larger than that of $A(w)$ but this is consistent with the lower average values of $\chi(u)$ for the resampled data in Figure 4.5.

4.5.3 Goodness-of-fit

In the previous subsections criteria were presented to discriminate between different models for the dependence between two random variables. To test the overall adequacy of a bivariate model, both the dependence structure and

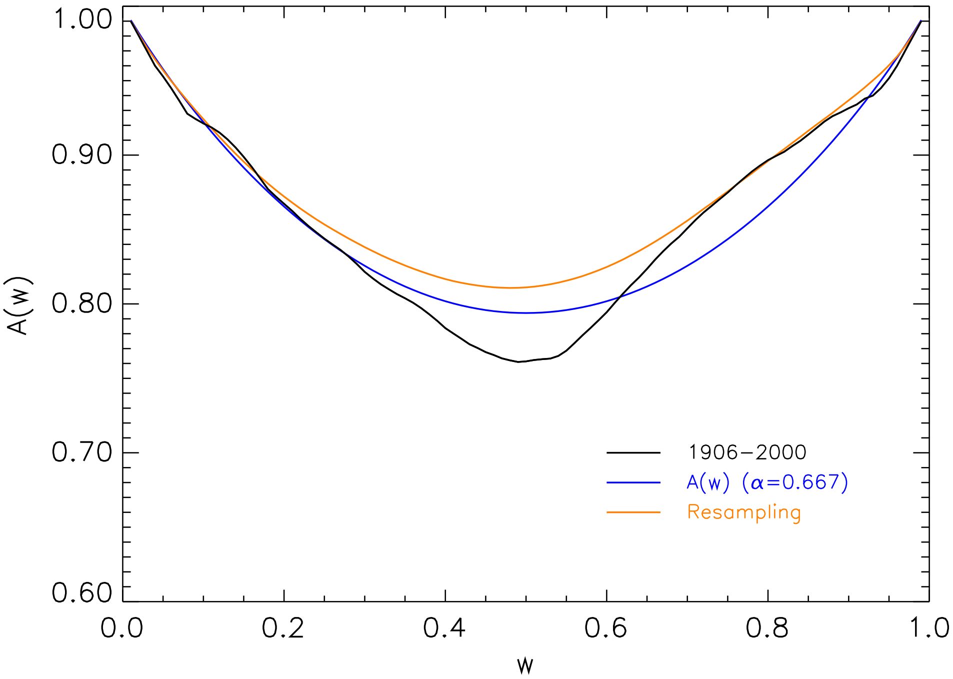

Figure 4.6. Nonparametric estimates of $A(w)$ from the historical and simulated data and $A(w)$ for the fitted logistic dependence model.

the fits of the individual marginal distributions should be taken into account. Here the goodness-of-fit of a bivariate model is assessed with joint exceedance probabilities. This is similar to Yue et al. (1999) and Yue (2001) who tested the validity of a bivariate model with empirical nonexceedance probabilities. Exceedance probabilities are preferred here because of the interest in discrepancies in the upper tail of the joint distribution.

For each data pair (x_i, y_i), a joint exceedance probability can be estimated as:

$$\hat{p}(x_i,\ y_i) = \frac{\#\ \text{pairs}\ (x_j,\ y_j)\ \text{with}\ x_j \geq x_i\ \text{and}\ y_j \geq y_i}{N+1}\,, \tag{4.12}$$

and this can be compared with the theoretical value of $\Pr(X \geq x_i,\ Y \geq y_i)$ for the fitted bivariate model or a similar empirical estimate for the resampled data.

Besides the bivariate normal distribution and the bivariate Gumbel distribution a novel bivariate distribution is considered, namely a bivariate normal distribution with a logistic Gumbel dependence structure. The latter is a logical combination of the other two bivariate distributions and it is constructed from the bivariate Gumbel model, using the transformations:

$$\tilde{X} = \hat{H}_X^{-1}\left[\hat{G}_X(X)\right] \qquad \tilde{Y} = \hat{H}_Y^{-1}\left[\hat{G}_Y(Y)\right] , \tag{4.13}$$

where $\hat{G}_X$ and $\hat{G}_Y$ are the fitted Gumbel distributions, and $\hat{H}_X$, $\hat{H}_Y$ the fitted lognormal and sqrt-normal distributions, respectively. Since these transformations are monotonic increasing, $(\tilde{X},\ \tilde{Y})$ has the same logistic dependence structure as $(X,\ Y)$. The transformations in equation (4.13) are similar to the normal quantile transformation in Kelly and Krzysztofowicz (1997). The inverse of the normal quantile transform has been used to obtain variables having marginal extreme value distributions and a multivariate normal dependence structure (Hosking and Wallis, 1988; Bortot et al., 2000). Equation (4.13) is, however, needed if a logistic Gumbel dependence structure is required.

Figure 4.7 shows joint probability plots for the three bivariate models and for the simulated data from the resampling model. To emphasize the upper tail, the exceedance probabilities are plotted on a logarithmic scale. In this

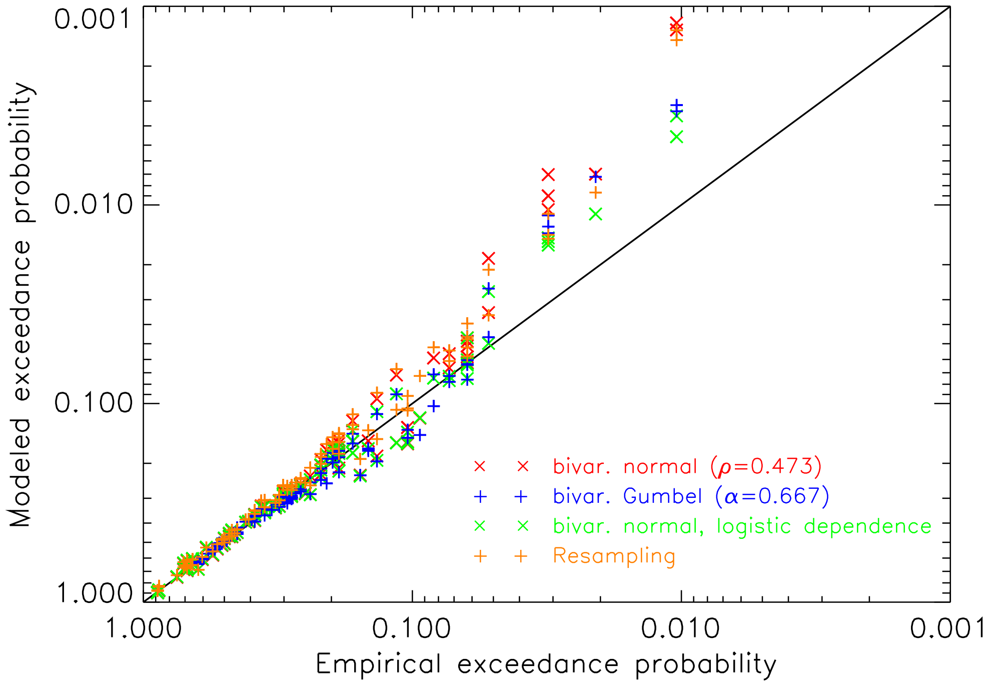

Figure 4.7. Joint probability plots for the fitted bivariate normal and Gumbel distributions, for the bivariate normal distribution with logistic Gumbel dependence structure, and for the data simulated with the resampling model.

tail region, the modeled probabilities tend to deviate systematically from the empirical probabilities, partly because these empirical probabilities are biased. The bias of $\hat{p}(x_i,\ y_i)$ depends on the degree of association of large values. From each bivariate distribution 10^4 samples of 95 years were generated to explore this bias. Figure 4.8 shows the bias for the three bivariate distributions. The bias is identical for the bivariate Gumbel distribution and the bivariate normal distribution with logistic Gumbel dependence and somewhat larger for the bivariate normal distribution. By comparing Figures 4.7 and 4.8 it is clear that the observed differences between the modeled and empirical joint exceedance probabilities in the upper tail region (in Figure 4.7) are larger than the simulated bias (in Figure 4.8), in particular for the bivariate normal distribution and the resampling model. This lack of fit in the upper tail for the resampling

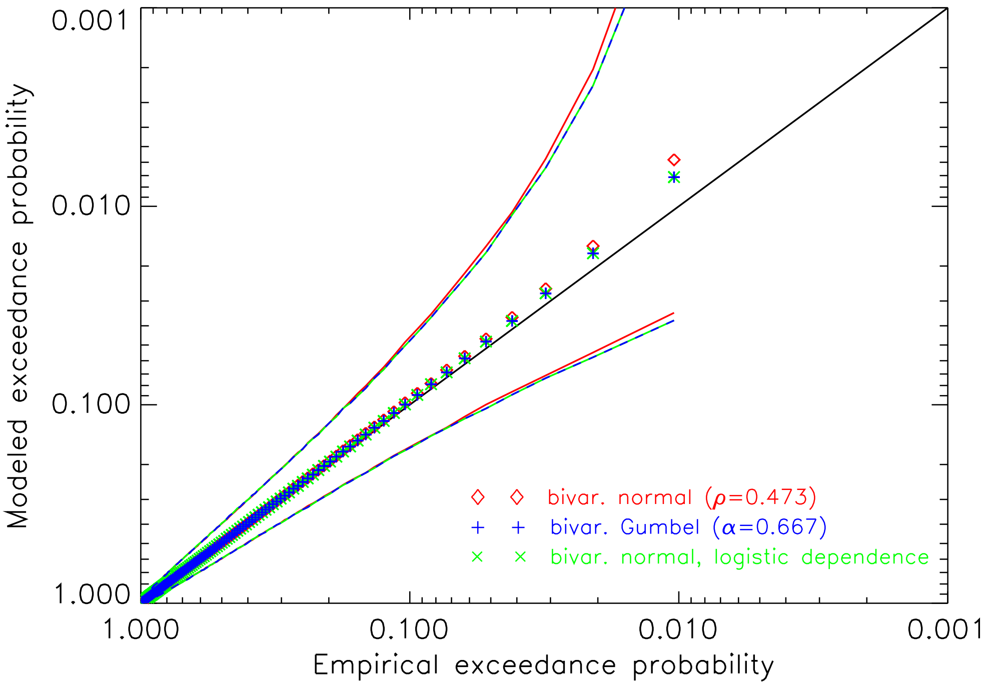

Figure 4.8. Bias of the empirical joint exceedance probabilities for the fitted bivariate normal and Gumbel distributions and for the bivariate normal distribution with logistic Gumbel dependence from a Monte Carlo experiment (10^4 simulations of 95 years). For each of the 95 empirical exceedance probabilities the symbols refer to the median value of the simulated theoretical exceedance probabilities and the lines denote a pointwise 95% interval for the theoretical exceedance probabilities.

Table 4.1. Mean return periods (yr) of joint exceedances of the observed precipitation and discharge deficits in given years for different bivariate distributions and the resampling model.

Year	Precipitation deficit (mm)	Discharge deficit (10^9 m^3)	Normal	Gumbel	Normal, logistic dependence	Resampling
1921	321.6	12.1	824	318	281	757
1976	361.1	10.7	760	296	221	676
1959	351.7	5.1	143	139	90	116
1947	296.1	7.8	142	78	65	90
1949	226.7	9.2	111	72	68	68

model is mainly the result of light tails of the simulated marginal distributions (see Figures 4.3 and 4.4), and for the bivariate normal distribution it is due to its asymptotic independence (Figure 4.5). Although all four models have a tendency to underestimate the joint exceedance probabilities in the upper tail region, the bivariate normal distribution with logistic Gumbel dependence performs best.

For 5 extreme years in the historical record the return periods of joint exceedances of the observed precipitation and discharge deficit, i.e., $T = 1/\Pr(X > x_i,\ Y > y_i)$ were determined. Table 4.1 compares the estimates of T from the different bivariate models. The return periods for the most extreme years (1921 en 1976) are more than 600 years for the bivariate normal distribution and the resampling model. These return periods reduce to less than 300 years if a bivariate normal distribution with logistic Gumbel dependence structure is assumed. Apart from a large sensitivity to model choice, the return periods are very uncertain due to sampling variability (see, for the univariate case, Stedinger et al., 1993). Yet this does not entirely explain why the estimates in Table 4.1 considerably exceed the length of the historical records from which they were derived. An important point is that the probability that two different variables exceed some high level simultaneously is smaller than the marginal exceedance probabilities for each of the two variables. The magnitude of this effect can be estimated from the same Monte Carlo experiment that was used to determine the bias of $\hat{p}(x_i, y_i)$. For each 95-year sample from the normal distribution with logistic dependence, the return periods of the joint exceedances of the simulated precipitation and discharge deficits were determined. The median of the longest return period in the 10^4 simulations of 95 years is 320 years which is quite large compared to the size of the sample. As a result of this effect all 5 years considered in Table 4.1 have return periods longer than 60 years.

4.6 Failure regions

In practical applications, the joint probability that X and Y lie in a 'failure region' different from the rectangle defined by $(X > x,\ Y > y)$ might be of interest. For example, structures often fail if a combination of the constituent variables becomes extreme. This combination then marks the boundary of the failure region. For the assessment of droughts in the Netherlands it is useful to base the failure region on the economic damage D_{E}.

The economic damage from 7 historical years (1949, 1959, 1967, 1976, 1985, 1995 and 1996 (T. Kroon, personal communication, 2004)) reveals that D_{E} can be approximated as:

$$D_{\mathrm{E}} = ax + by + c\,, \tag{4.14}$$

with x the maximum precipitation deficit and y the discharge deficit. The regression coefficients a, b and c were estimated by a least squares fit. Let x_i and y_i be the observed precipitation and discharge deficits for the year of interest. Events with a precipitation and discharge deficit in the region above the line through $(x_i,\ y_i)$ and with slope $\Delta = -a/b$ should then be considered as more extreme in terms of economic damage. For the years 1976, 1959 and 1949, Figure 4.9 compares the boundary of this failure region with the rectangle $(X > x_i,\ Y > y_i)$. The slope of the bounding line indicates that the economic damage is relatively more sensitive to the precipitation deficit. Table 4.2 presents, for each of the historical years in Table 4.1, the return periods for the failure region based on equation (4.14). These return periods were obtained empirically from 10^6 simulated pairs $(x_i,\ y_i)$ from the corresponding bivariate distribution and from the 10^5 simulated years in the case of nearest-neighbour resampling. The estimated return periods in Table 4.2 are much shorter than those in Table 4.1, in particular for 1921 and 1976. Using a failure region related to the economic damage gives the longest return period for 1976 while in Table 4.1 the longest return period is found for 1921. This is a result of the relatively smaller contribution of the discharge deficit to the economic damage (see Figure 4.9). In Table 4.1 the return periods are longest for the bivariate normal distribution while in Table 4.2 the longest return periods are found for the bivariate Gumbel distribution and the resampling model. The shortest return periods are found in both tables for the bivariate

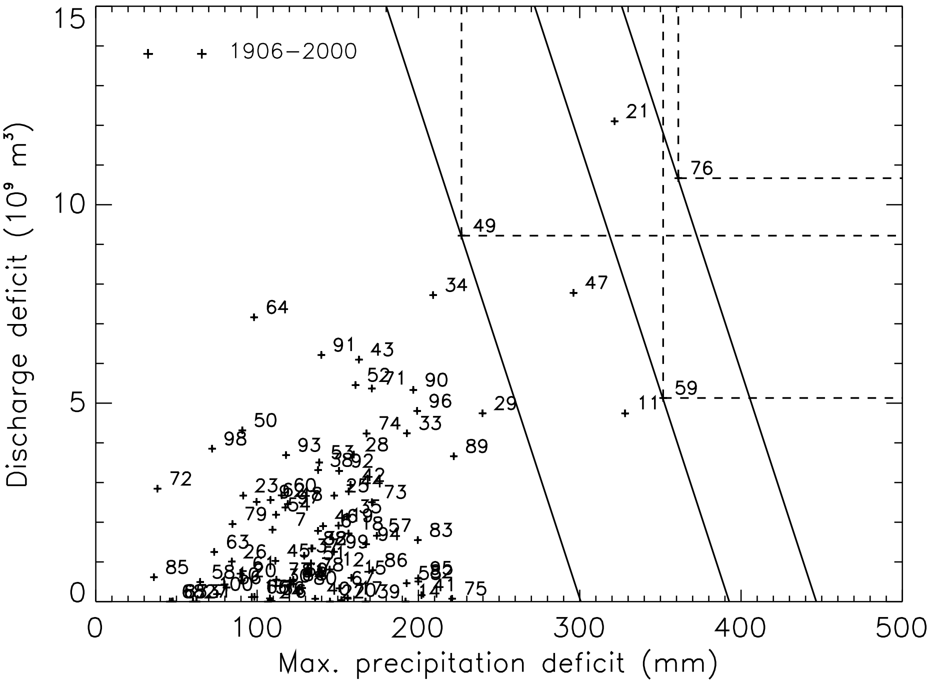

Figure 4.9. Failure regions related to the economic damage (equation 4.14) and rectangles ($X > x_i$, $Y > y_i$) for the historical years 1976, 1959 and 1949 (indicated as 76, 59 and 49).

normal distribution with logistic Gumbel dependence, but the difference from the standard bivariate normal distribution is much smaller in Table 4.2. This is in line with results of Tawn (1988) and Coles and Tawn (1994) that the sensitivity of joint probabilities to assumptions about the dependence structure varies considerably with the type of failure region. For the best fitting model (bivariate normal distribution with logistic dependence), the estimated return period of 110 years for the most extreme year in terms of economic damage, 1976, is close to the length of the historical record. In contrast to the return periods in Table 4.1, the estimates in Table 4.2 can be considered as a univariate exceedance probability, namely that for the economic damage D_{E}.

Although the regression coefficients in equation (4.14) differ significantly from zero at the 10% level, the slope Δ is quite uncertain. To determine the effect of this uncertainty on the estimated return periods, the latter were recalculated with the 5th percentile Δ_{L} and the 95th percentile Δ_{U} of the empirical distribution of the estimated slope in 10^4 bootstrap samples of size 7. The resulting spread in the return periods is presented in Figure 4.10. For

Table 4.2. Mean return periods (yr) of situations where the precipitation and discharge deficits are more extreme than the observed deficits in the given years in terms of economic damage (equation 4.14) for the different bivariate distributions and the resampling model.

Year	Precipitation deficit (mm)	Discharge deficit (10^9 m^3)	Normal	Gumbel	Normal, logistic dependence	Resampling
1921	321.6	12.1	99	113	79	98
1976	361.1	10.7	147	172	110	178
1959	351.7	5.1	66	75	55	67
1947	296.1	7.8	41	46	36	46
1949	226.7	9.2	17	19	17	24

1959, a year with a relatively small discharge deficit, a failure region with slope Δ_{L} leads to a longer return period and a region with slope Δ_{U} shortens the return period, while for the other years in Table 4.2 the return periods change the other way round. Within the uncertainty of Δ, 1976 always has the largest economic damage and thus the longest return period. However, 1959 becomes more extreme than 1921 if the failure region has slope Δ_{L} and it becomes less extreme than 1947 if the failure region has slope Δ_{U}. So the ranking of the drought events also depends on the slope of the failure region.

4.7 Discussion and conclusions

Different probability distributions have been fitted to the annual maximum precipitation deficit in the Netherlands and the annual discharge deficit of the river Rhine. The fitted distributions have been compared with an empirical bivariate distribution obtained with a resampling model. It is found that the degree of association between large values is too weak if the dependence structure of a bivariate normal distribution is assumed. This results in a strong underestimation of the probabilities of joint exceedances of extreme values. The joint occurrence of large values is better described by the dependence structure of a limiting Gumbel distribution. Its symmetric nature is also in agreement with the data. This dependence function has therefore not only been studied with Gumbel marginals but also with transformed normal marginals. The latter describes the upper tail of the precipitation deficit dis-

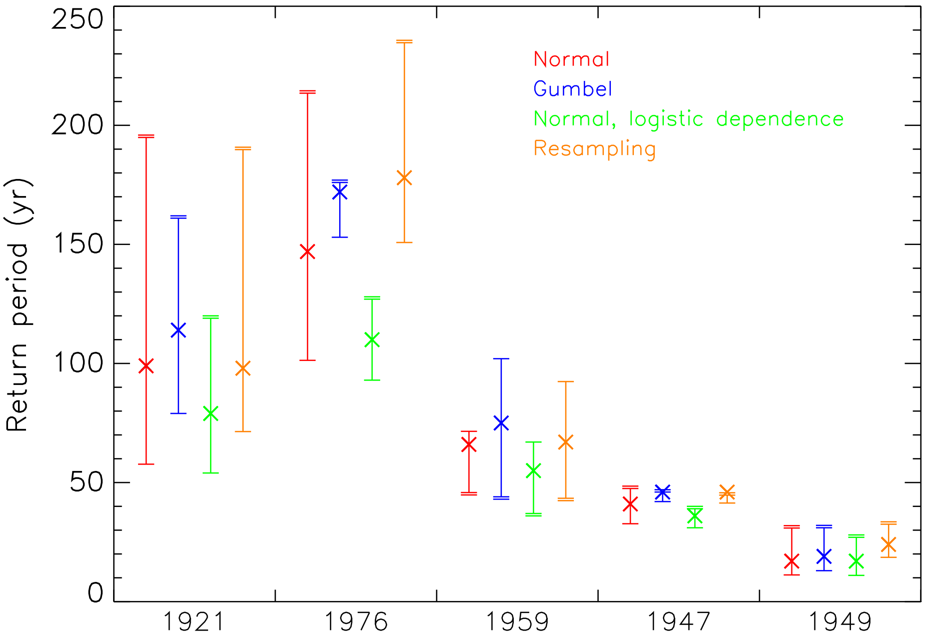

Figure 4.10. Spread in return periods due to the uncertainty of the failure region slope Δ. The single horizontal bars correspond with Δ_L, the double horizontal bars with Δ_U and the crosses ($\times$) with the return periods in Table 4.2.

tribution better, leading to shorter return periods between extreme bivariate events than the Gumbel distribution. The assumption of Gumbel marginals is, however, not rejected by the Anderson-Darling and the ppcc tests. For the resampling model the dependence structure and its symmetry agree well with the data. The resampling model is the only model which can (to some extent) reproduce the curvature in the tail of the historical distribution of the maximum precipitation deficit, although it underestimates the most extreme quantiles of this distribution. The tail of the simulated distribution of the discharge deficit seems too light as well, in particular near the most extreme event (1921). This discrepancy seems to be related with differences in the strength of the autocorrelation between the variables $E-P$ and Q. Decade values of $E-P$ exhibit only weak autocorrelation whereas for discharge Q there is still considerable autocorrelation at a lag of 10 decades (see Appendix D). A much better simulation of the upper tails of the marginal distributions can be achieved when $E-P$ and Q are resampled individually rather than simultaneously (see Section 5.4). This is, however, not of interest for the estimation of the drought probabilities considered in this chapter.

The use of a failure region based on economic damage has been studied as an alternative to ordinary joint exceedances. This failure region not only shortens the estimated return periods of historical drought events, it also reduces the differences between the various bivariate models. For the most extreme year in terms of economic damage, 1976, the return period is 172 years for the bivariate Gumbel distribution, 110 years for the transformed normal distribution with logistic Gumbel dependence and 178 years for nearest-neighbour resampling. A detailed study of the uncertainty of these return periods was beyond the scope of this chapter, but the uncertainty is dependent on lack of fit in the upper tail of the joint distribution, the limited sample size, and the uncertainty in the slope of the failure region. The size of the latter uncertainty depends on the type of distribution and may vary considerably from year to year (see Figure 4.10).

Acknowledgments. The authors thank Prof. Bert Holtslag at Wageningen University, two anonymous reviewers and an Associate Editor of Water Resources Research for their comments on an earlier version of the paper. The work was performed in the framework of the ‘Drought studies of the Netherlands’ in cooperation with the Institute for Inland Water Management and Waste Water Treatment (RIZA), Lelystad.

Chapter 5

Drought in the Netherlands - Regional frequency analysis versus time series simulation

Jules J. Beersma and T. Adri Buishand, 2006

Submitted to *Journal of Hydrology*

Abstract

The distribution of the annual maximum precipitation deficit is studied for six districts within the Netherlands. Gumbel probability plots of this precipitation deficit show a common extraordinary curvature in the upper tail. A regional frequency analysis yields a regional growth curve that can be approximated by a spline consisting of two linear segments on the standard Gumbel scale and a smooth transition between them. Alternatively, the application of a time series model based on nearest-neighbour resampling is explored. To reproduce the curvature of the precipitation deficit distributions it is necessary to include a 4-month memory term in the resampling model. This memory term leads, however, to a considerable increase of the standard error of large quantile estimates.

Much attention is given to the use of the bootstrap and the jackknife to determine the standard errors of quantile estimates based on nearest-neighbour resampling. A simulation experiment with a first-order autoregressive time series model shows that these standard errors can be biased, in particular for the bootstrap. The relative standard errors of quantile estimates are large in

the area of large curvature of the Gumbel probability plots. This holds both for nearest-neighbour resampling and regional frequency analysis. When the two methods are used for extrapolation, nearest-neighbour resampling clearly outperforms the regional frequency analysis. The latter then shows a strong increase in the relative standard error of quantile estimates with increasing return period due to the large uncertainty of the parameters in the spline model.

Using nearest-neighbour resampling and the bootstrap, confidence intervals are constructed for the return periods of the largest observed precipitation deficit for each of the six districts. Although these confidence intervals are quite wide, they are on average a factor of two narrower than the interval expected from the size of the sample only.

5.1 Introduction

The probability of drought events is a regularly recurring topic in drought studies. Quite often there is particular interest in the frequency of occurrence of the most extreme historic events. Since such events are by definition rare it is not so easy to obtain accurate estimates of their frequency of occurrence. Furthermore, one should keep in mind that the severity of a drought, and thus its associated probability, also depends on the sector that suffers from the drought. In addition, drought is usually not only controlled by lack of precipitation but also by evaporation.

In the Netherlands a frequently used measure of the severity of drought of a certain year is the maximum cumulative difference between potential evaporation and precipitation in the summer half of that year. This measure of drought is strongly related to moisture deficits for the vegetation during the growing season. In the previous chapter (Beersma and Buishand, 2004) it was used as a measure of drought for the country as a whole while in this chapter it is used to investigate and quantify regional differences within the Netherlands. In Section 4.3 it was already noted that a Gumbel probability plot of the annual maximum country-average precipitation deficit contains an extraordinary curvature at large deficits. For return periods beyond 20 years, the distribution is much heavier tailed than the Gumbel distribution (which would represent a straight line) while for very long return periods the tail is becoming thinner again (see Figure 5.1). To guide the eye Figure 5.1 also presents the distribution obtained with the time series simulation model that is introduced in Section 5.4.

To study the regional differences in the Netherlands, the country is divided into six geographic districts. It turns out that the extraordinary curvature

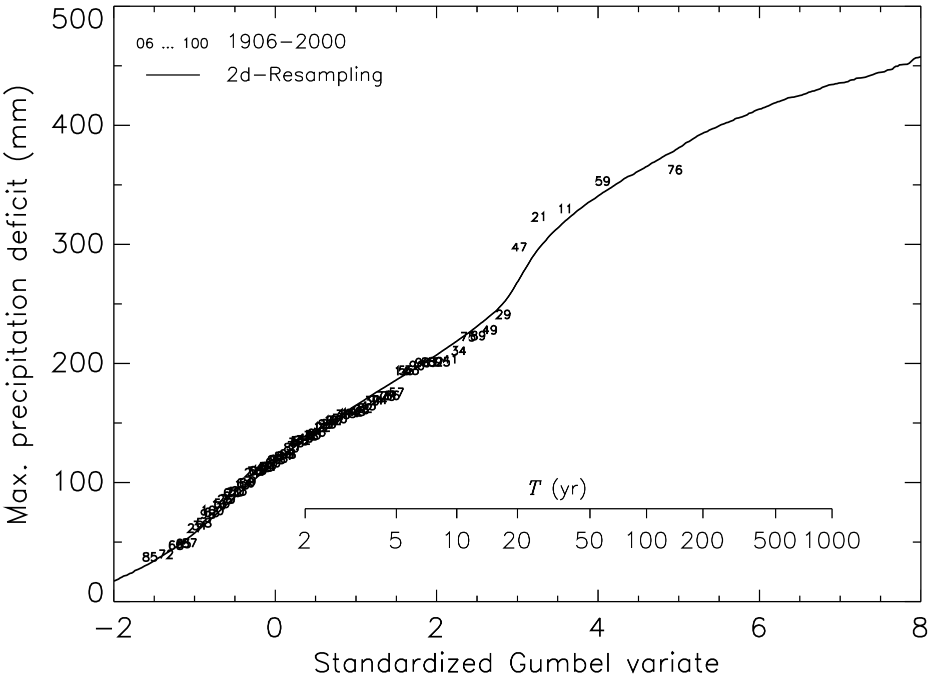

Figure 5.1. Gumbel probability plot of the annual maximum country-average precipitation deficit together with a plot of the annual maxima from a 100 000-year sequence generated with the time series model in Section 5.4.1. The numbers represent the year minus 1900; T refers to the return period.

in the upper tail of the distribution is a common characteristic for all six districts (Section 5.2). This curvature poses the question how to estimate large quantiles of the six distributions reliably and efficiently. One approach is a regional frequency analysis (e.g., Hosking and Wallis, 1997). For the precipitation deficits in the Netherlands this approach means that the six probability distributions have a common shape which implies that the most uncertain parameters of the six distributions can be taken equal for the six districts. To derive these common parameters the records of the six districts are combined. The resulting parameter estimates have a smaller uncertainty compared to those from a single record, which leads to more accurate estimates of extreme quantiles and exceedance frequencies. Time series simulation is another approach to obtain more accurate estimates. In Beersma et al. (2004) resampling of the original precipitation deficits was considered. A reduction of the variance of quantile and frequency estimates is then expected from the more efficient use of the available data (Buishand, 2006).

In this chapter a comparison is made between time series simulation and a regional frequency analysis. There is a focus on the accuracy of quantile estimates given the limited sample size of the historic data. In Section 5.2 the geographic districts are defined and the historic data that were used to calculate the precipitation deficits for these districts are described. The regional frequency analysis is performed in Section 5.3 and time series simulation based on nearest-neighbour resampling is presented in Section 5.4. Section 5.5 compares the accuracy of quantile estimates obtained with both methods for different quantiles, and concludes with a discussion and a summary of the results.

5.2 Historic data and choice of districts

The precipitation deficit in any period is the difference between precipitation and potential evaporation in that period. Potential evaporation which is routinely calculated for short grass is also known as the grass reference evaporation. Around early April the daily average potential evaporation becomes larger than the daily average precipitation in the Netherlands. The deficit is therefore accumulated from April 1 onward. After 30 September the average cumulative precipitation deficit tends to decrease because global radiation and thus potential evaporation are reduced. The annual maximum precipitation deficit is the largest precipitation deficit that occurs during the summer half-year (1 April − 30 September). This period largely coincides with the growing season in the Netherlands. The vegetation will generally not grow optimally during periods with a positive cumulative deficit. If there is no positive cumulative deficit, precipitation surpluses will lead to runoff. Without retention measures, these surpluses can not compensate for future positive deficits. The cumulative precipitation deficit is therefore reset to zero when it becomes negative.

Both for precipitation and evaporation daily values were available for the 95-year period 1906–2000. The grass reference evaporation was derived from temperature and sunshine duration at station De Bilt using the Makkink formula (e.g., de Bruin and Stricker, 2000). The global radiation in that formula was estimated from an empirical relation between global radiation and sunshine duration due to Frantzen and Raaff (1982). In Chapter 4 (Beersma and Buishand, 2004) it was shown that this reference evaporation compares very well with the original Makkink evaporation which is available from 1958 onwards only.

Daily precipitation for the 1906–2000 period was available for 18 stations spread over the country. Thirteen of these stations were used to calculate

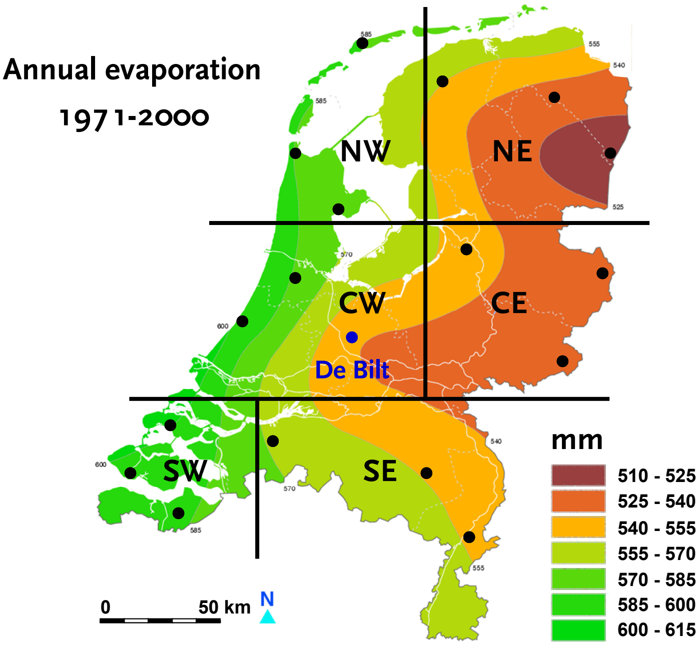

Figure 5.2. Map of annual mean evaporation together with the position of the 18 precipitation stations and the six geographic districts used in this chapter.

the daily country-average precipitation for the Netherlands in Section 4.2 and the country-average precipitation deficit in Figure 5.1. In this chapter, the Netherlands is divided into six geographic districts; North West (NW), North East (NE), Central West (CW), Central East (CE), South West (SW) and finally South East (SE). For each district, time series of daily average precipitation were obtained by averaging the precipitation amounts at three stations in that district. Time series of daily reference evaporation for each district were derived by adjusting the reference evaporation of station De Bilt using the spatial differences in annual mean evaporation for the 1971–2000 normal period. A map of the mean evaporation in the Netherlands including the position of the 18 precipitation stations and the six districts is presented in Figure 5.2. The evaporation adjustment factors that were used are: NE and CE: −1.5%; SE: +3%; CW: +5%; NW: +6% and SW: +9%.

Analogous to Figure 5.1 Gumbel plots of the annual maximum precipitation deficit for all six districts are presented in Figure 5.3. This figure clearly shows the common curvature in the tail of the distributions. Physically one can understand this behaviour. It is reasonable to assume that at a certain level of the precipitation deficit a positive feedback develops in which, as a

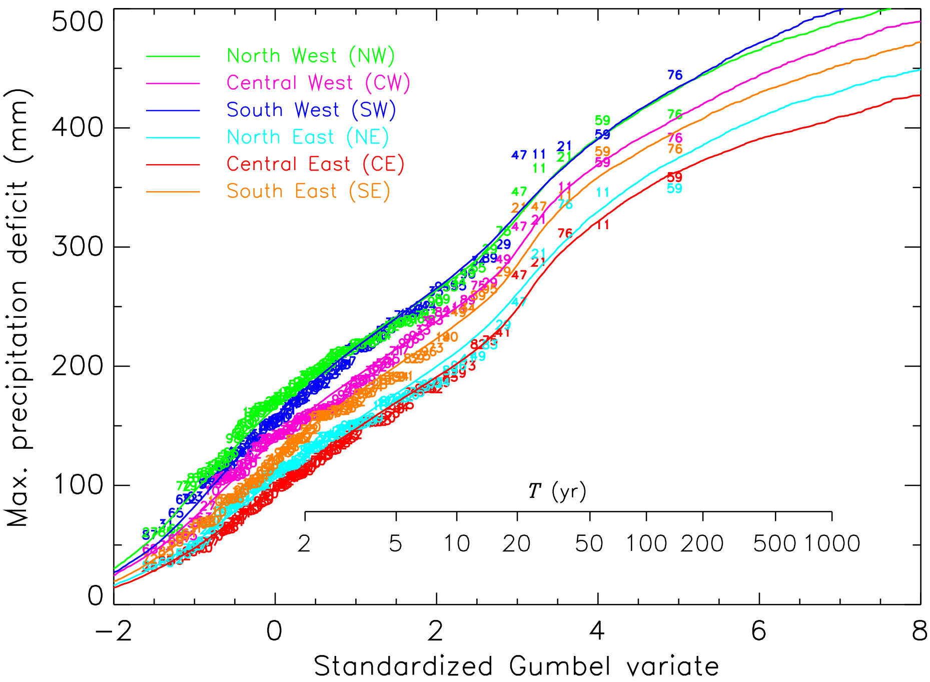

Figure 5.3. Gumbel probability plots of the annual maximum precipitation deficit for the six districts together with plots of the annual maxima from a 100 000-year sequence generated with the 2d-resampling model in Section 5.4.1. The coloured numbers represent the year minus 1900 for each district.

result of drying of the soil, cloudiness and the occurrence of precipitation are reduced, and temperature and global radiation (and thus the potential evaporation) are increasing. Both the reduction of precipitation and the increase of the potential evaporation enhance the precipitation deficit. It is also clear that such a feedback cannot go on indefinitely since it is bounded by zero precipitation and maximum potential evaporation (the latter of which is mainly bounded by global radiation). To determine the individual contributions of precipitation and potential evaporation to the assumed feedback, the annual maximum country-average precipitation deficits were also calculated for the hypothetical case where the reference evaporation is fixed according to its 1906–2000 climatology as well as the hypothetical case where the country-average precipitation is fixed according to its 1906–2000 climatology. In the first case the drought related reference evaporation feedback is eliminated but the precipitation feedback is still present while in the second case the precip-

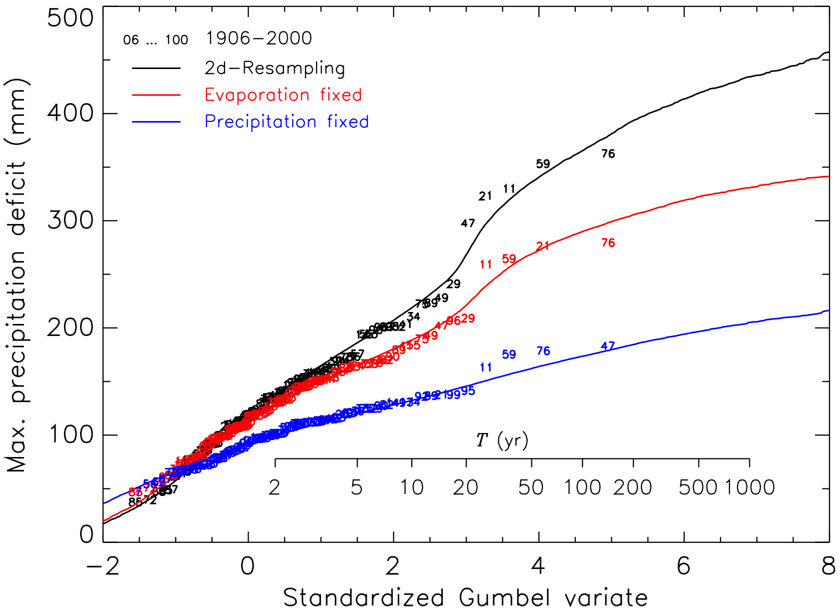

Figure 5.4. As Figure 5.1, together with the hypothetical case with fixed climatological reference evaporation (red) and the hypothetical case with fixed climatological precipitation (blue). See text for details.

itation feedback is eliminated but the evaporation feedback is still present. Figure 5.4 presents the Gumbel plots for these two hypothetical cases. These plots demonstrate that the annual maximum deficit is mainly controlled by precipitation variability. The four largest values in the case with fixed evaporation (red curve) show that it is most likely that a drought related precipitation feedback is responsible for the curvature of the Gumbel probability plot of the annual maximum precipitation deficit. The case with fixed precipitation (blue curve) shows that the effect of a potential evaporation feedback, if present at all, is much less pronounced than that of the precipitation feedback.

Besides similarities in the shape of the probability distribution of the annual maximum precipitation deficit for the six districts there are also some clear climatological differences between the coastal (western) and the inland (eastern) districts. Evaporation is on average larger in the western districts compared to the eastern districts (see Figure 5.2) while precipitation in the summer half-year is on average somewhat larger in the eastern districts compared to the western districts (not shown). As a result, the annual maximum precipitation deficit is on average larger in the western part of the country than in the eastern part. Table 5.1 summarizes the average annual maximum

Table 5.1. Sample mean, standard deviation, coefficient of variation (CV) and skewness of the annual maximum precipitation deficits in each of the six districts. Mean and standard deviation are given in mm, CV and skewness are dimensionless.

District	Mean	Std. deviation	CV	Skewness
North West (NW)	194.1	67.2	0.346	0.829
Central West (CW)	165.5	67.9	0.410	0.919
South West (SW)	186.9	73.2	0.392	1.044
North East (NE)	133.6	61.6	0.461	1.244
Central East (CE)	124.2	63.4	0.511	1.254
South East (SE)	151.7	70.4	0.464	1.134
Average over districts	159.3	67.3	0.431	1.071
Std. dev. between districts	28.1[a]	4.3[a]	0.059[a]	0.173

[a] Significant at the 5% level (see Appendix F)

precipitation deficit and a few other relevant statistics for each of the six districts. The largest differences, in terms of these statistics, are those between districts NW and CE. The NW district has the largest average precipitation deficit and the smallest coefficient of variation (CV, i.e., the standard deviation divided by the mean) and skewness, while district CE has the smallest precipitation deficit and the largest CV and skewness. It should be noted, however, that for a sample of size 95, the sample skewness is a biased and very variable statistic. Wallis et al. (1974) indicate that there is a negative bias of about 10%. The average skewness in Table 5.1 equals 1.07. After bias correction, this average is quite close to the theoretical value of 1.14 for the Gumbel distribution. Simulation shows that the differences between the skewness estimates for the six districts are not significant at the 5% level (see Appendix F). The differences between the means, standard deviations and CVs are however, significant at the 5% level.

Finally, note that prior to the analyses all daily data were converted into decades of days. Decades of days were obtained by dividing each calendar month into three decades; the first two decades in a month always represent 10 days and the third decade represents the remaining days. Each year thus contains 36 decades of days. The main reason for using decades of days instead of the daily data is that it saves a factor of ten on the computer running time of the resampling procedure (in Section 5.4.1) without loss of performance of the simulated precipitation deficits. The effect of the conversion into decades of days on the annual maximum precipitation deficit is negligible.

5.3 Regional frequency analysis

5.3.1 Basic extreme-value distributions

First, it was investigated if a common and simple probability distribution applies to all six districts. Although the curvature in the Gumbel probability plots (see Figure 5.3) suggests that the Gumbel distribution may not be appropriate, its skewness is close to the average bias-corrected skewness of the six districts. The Gumbel distribution was therefore fitted for all six districts together with the 2-parameter lognormal and Generalized Extreme Value (GEV) distributions, where the GEV distribution is defined by:

$$F(x) = \Pr(X \leq x) = \exp\left\{ - \left[1 - \frac{k(x-\mu)}{\sigma} \right]^{1/k} \right\} , \qquad (5.1)$$

with μ, σ and k respectively its location, scale and shape parameter. For $k = 0$ the GEV distribution reduces to the Gumbel distribution:

$$F(x) = \Pr(X \leq x) = \exp\left[-e^{-(x-\mu)/\sigma} \right] . \qquad (5.2)$$

The GEV distribution has a heavier upper tail than the Gumbel distribution if $k < 0$. For $k > 0$ it has a relatively light upper tail with an upper bound. The lognormal distribution assumes that the logarithm of the data are normally distributed. The GEV distribution was fitted using probability-weighted moments (Hosking et al., 1985) while the Gumbel and lognormal distributions were fitted by maximum likelihood (ML),

As in Chapter 4 (Beersma and Buishand, 2004) the fitted distributions were subjected to the Anderson-Darling (A-D) goodness-of-fit test (e.g., Kotz and Nadarajah, 2000). For the Gumbel and lognormal distributions the percentage points of the A-D statistic in Stephens (1986a) were used and for the GEV distribution those in Ahmad et al. (1988). The A-D test is known to be sensitive to deviations in the tails of the distribution. Despite this sensitivity the A-D test gave favourable results for the Gumbel distribution compared to the lognormal distribution. The lognormal distribution resulted in significant values (at the 5% level) for the districts NW, NE and CW while for the Gumbel distribution this was only the case for district NW. The fitted lognormal distributions have a heavier upper tail than the Gumbel distribution. This

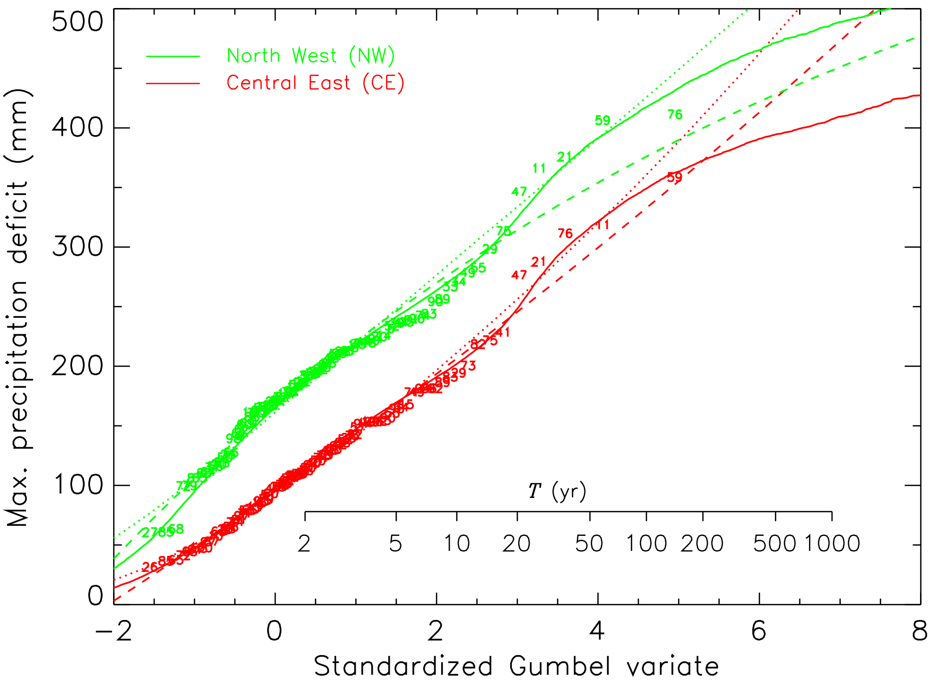

Figure 5.5. Gumbel probability plots of the annual maximum precipitation deficit for two districts together with plots of the annual maxima from a 100 000-year sequence generated with the 2d-resampling model in Section 5.4.1 (solid), the fitted lognormal distributions (dotted) and the fitted GEV distributions (dashed). The coloured numbers represent the year minus 1900.

is also expressed by their relatively large skewness which ranges between 1.17 and 1.81 compared to 1.14 for the Gumbel distribution. Apart from the NW district, the GEV distributions are close to the Gumbel distribution, with the shape parameter ranging between −0.035 and 0.032. For the NW district, the GEV distribution has the largest shape parameter ($k = 0.105$) and there is a considerable reduction in the value of the A-D statistic. This statistic remains, however, significant at the 5% level. Figure 5.5 shows the lognormal and GEV fits for the two districts that differ most, NW and CE. Note, in particular, that due to their different shapes, the two fitted GEV distributions intersect at a return period of about 500 years.

In summary, none of the three distributions provides an adequate fit to the observed distribution for all six districts. For the NW district the GEV distribution tends to underestimate the quantiles in the upper tail of the distribution (Figure 5.5), whereas the GEV fit is almost indistinguishable from the Gumbel fit for the other five districts.

5.3.2 Regional growth curve

A key assumption in regional frequency analysis is often that the distributions of the annual maxima become identical after some standardization. Here it is assumed that the distribution of the standardized annual maximum precipitation deficit $X^* = (X_i - \mu_i)/\sigma_i$, with X_i the annual maximum precipitation deficit for district i $(i = 1, \dots, 6)$, is the same for all districts. Here μ_i is a location parameter and σ_i is a scale parameter. A Gumbel plot of the standardized maxima is usually denoted as the regional growth curve, a dimensionless quantile function common to every site. The regional growth curve may form the basis of a regional frequency analysis.

For each district standardized values were derived by replacing μ_i and σ_i by the ML estimates of the location and scale parameter of the Gumbel distribution, and these were ranked in increasing order as $x_1^* \leq x_2^* \leq \ldots \leq x_{95}^*$. The regional growth curve in Figure 5.6 is obtained by averaging the x_j^* over the six districts for each j $(j = 1, \dots, 95)$. As expected from Figure 5.3, the regional growth curve deviates from the straight line for the Gumbel distribution due to its heavy upper tail.

Two Component Extreme Value (TCEV) distributions have been used in the literature to describe data with a heavy upper tail. In its most simple form the TCEV distribution function consists of the product of two Gumbel distribution functions; a basic component which covers most of the data and an outlier component which is more heavily tailed than the basic component (e.g., Rossi et al., 1984; Fiorentino et al., 1987). Given the curvature of our regional growth curve, the GEV distribution with a positive shape parameter k seems a more appropriate choice for the outlier component.

However, attempts to fit the TCEV distribution with this outlier component were not successful. A crucial point is that a rather large value of k $(k > 1)$ is needed to catch the amount of curvature of the growth curve. In this situation, ML estimation meets a serious problem (Coles, 2001) due to a singularity in the likelihood equation.

As an alternative, the regional growth curve can be approximated by linear segments on the standard Gumbel scale (Reed et al., 1999). In our case one linear segment was used for return periods up to about 15 years (which is standard Gumbel due to the standardization) and a second one was used for return periods of about 30 years onward. For return periods between 15 and 30 years a non-linear relationship is needed to obtain a smooth transition between the two linear segments. This leads to a cubic regression spline with

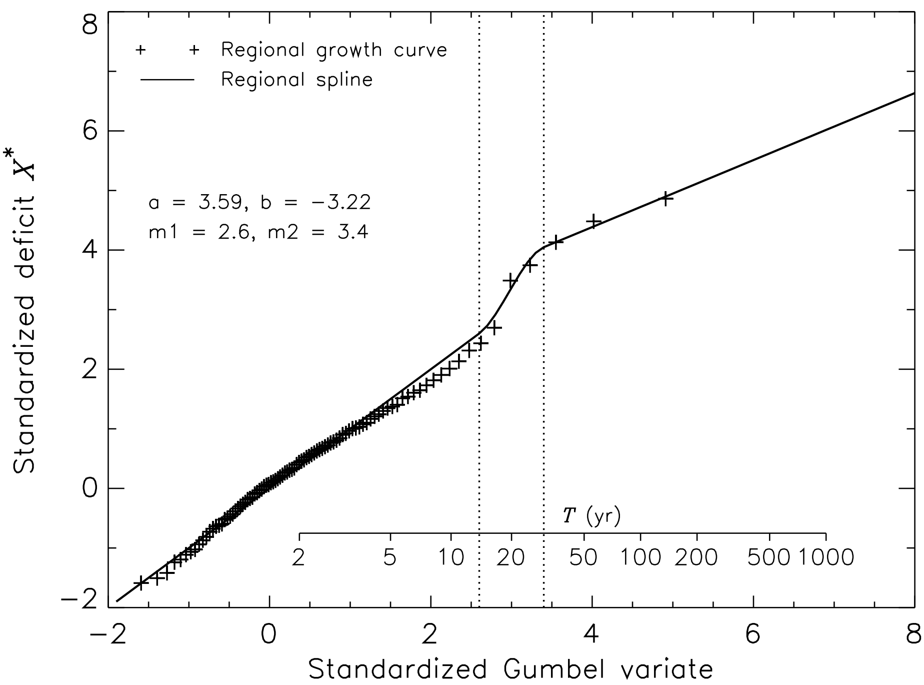

Figure 5.6. Gumbel probability plot of the standardized annual maximum precipitation deficit together with the fitted regional 2-parameter spline. The plusses represent the average standardized deficits of the six districts, and the dotted lines the position of the two knots of the spline model.

two knots, m_1 and m_2. In formula the spline model is given by:

$$X^* = \begin{cases} Y\,, & Y \leq m_1 \\ Y + a(Y - m_1)^2 + b(Y - m_1)^3\,, & m_1 < Y \leq m_2 \\ Y + a(m_2 - m_1)(2Y - m_1 - m_2) & \\ \quad + b(m_2 - m_1)^2(3Y - m_1 - 2m_2)\,, & Y > m_2 \end{cases} \tag{5.3}$$

where Y is the standard Gumbel variable, i.e.,

$$\Pr(Y \leq y) = \exp\left[-e^{-y}\right]\,. \tag{5.4}$$

Due to the cubic part the growth curve and its first derivative are continuous everywhere. Given the knots m_1 and m_2, the two spline parameters a and b were estimated by minimizing the sum of the squared differences:

$$\mathrm{SS} = \sum_{i=1}^{6}\sum_{j=1}^{95}(x^*_{i,j} - x^*_{\mathrm{model},j})^2\,, \tag{5.5}$$

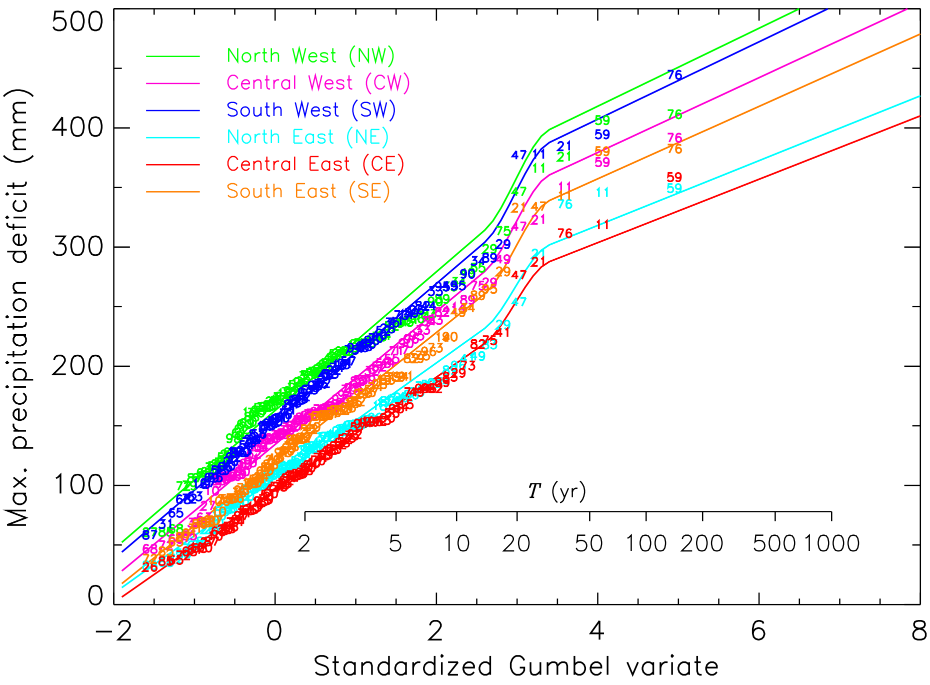

Figure 5.7. Gumbel probability plot of the annual maximum precipitation deficit for the six districts together with the corresponding rescaled fitted regional spline. The coloured numbers represent the year minus 1900 for each district.

with $x^*_{i,j}$ the jth ordered standardized annual maximum precipitation deficit for district i and $x^*_{\mathrm{model},j}$ the standardized value from equation (5.3) with $Y = -\ln[-\ln(p_j)]$ and p_j the plotting position, $p_j = (j - 0.3)/95.4$. The knots m_1 and m_2 were found by fitting the spline iteratively. Figure 5.6 shows the optimal spline, with $m_1 = 2.6$ and $m_2 = 3.4$, on the standard Gumbel scale (corresponding with return periods of 14.0 and 30.5 years, respectively). For each of the six districts, Figure 5.7 compares the Gumbel plot of the annual maximum precipitation deficits with the rescaled fitted regional spline. The latter fits generally well. Discrepancies are however found in some of the upper tails, in particular a tendency to underestimate large quantiles of the distribution for the CE district for which the skewness in Table 5.1 is rather large and a similar tendency to overestimate large quantiles for the NW district for which the skewness is rather small.

The improved correspondence in the upper tail between the regional spline and the data compared to the Gumbel distribution is, however, not sufficient to conclude that the spline model should be preferred to the Gumbel distribution. Compared to the Gumbel distribution the spline model has two additional parameters which unfortunately leads to additional uncertainty (i.e., larger standard errors, se). The spline model may therefore not necessarily be better in terms of root-mean-square error, rmse $= (\text{se}^2 + \text{bias}^2)^{1/2}$. To compare the uncertainty of the quantile estimates from the two distributions a simulation experiment was performed. In this experiment it was assumed that the fitted spline model represents the true distribution. The results from the Monte Carlo (MC) method were compared with those of the bootstrap and jackknife methods. The most essential details of the three methods are:

- MC simulation: 10 000 samples from the fitted spline model were generated. As in the original data, each MC sample consists of six 95-year sequences representing the annual maximum precipitation deficits of the six districts. The standardized Gumbel variables Y in the spline model (equation 5.3) were generated from the logistic multivariate Gumbel distribution (Stephenson, 2003). This distribution assumes that the precipitation deficits of the six districts are equicorrelated. The correlation coefficient of the multivariate Gumbel distribution was set equal to the average correlation coefficient of the annual maximum precipitation deficits (0.889). Both the Gumbel distribution and the spline model were fitted to the data in each MC sample. For the spline the knots were the same as in the fit to the observed data, i.e., $m_1 = 2.6$ and $m_2 = 3.4$ but the two spline parameters were estimated for each sample individually. For each MC sample the quantile estimates from the two fitted distributions were compared with the quantile estimates from the spline fitted to the original data.

- Bootstrap: 10 000 samples were drawn (with replacement) from the original data. Again, each bootstrap sample consists of six 95-year annual maximum precipitation deficits for the six districts. In order to preserve the correlation between the precipitation deficits of different districts, the same historical years are drawn for each district. The Gumbel distribution and the spline model were fitted to the data in each bootstrap sample in the same way as in the MC method. The comparison of the estimated quantiles was also identical to that in the MC method.

- Jackknife: The general idea behind the jackknife method is that a statistic of interest is recalculated repeatedly after omitting a part of the original data (e.g., Beersma and Buishand, 1999b; Chapter 2 of this thesis). In our case estimated quantiles from the Gumbel distribution and spline model were recalculated after omitting one year of historical data each time. So, for each fitted distribution we have in total 95 estimates, $\hat{\theta}_{-j}$, of the quantile of interest from (jackknife) samples in which the annual maximum precipitation deficit for one year is omitted. These 95 estimates are then used to obtain a jackknife estimate, $\hat{\mathrm{V}}_{\mathrm{jack}}^{1/2}$, of the standard error of the quantile estimate, where

$$\hat{\mathrm{V}}_{\mathrm{jack}} = \frac{J-1}{J} \sum_{j=1}^{J} \left(\hat{\theta}_{-j} - \hat{\theta}_{.} \right)^2 \tag{5.6}$$

with $\hat{\theta}_{.} = 1/J \sum_{j=1}^{J} \hat{\theta}_{-j}$ and $J = 95$. The jackknife further provides an alternative estimate of the quantile, $\hat{\theta}_{\mathrm{jack}} = J\hat{\theta} - (J-1)\hat{\theta}_{.}$ where $\hat{\theta}$ is the estimate from the complete sample.

Table 5.2 presents the relative bias, standard error and root-mean-square error of the 98% quantile (corresponding to a return period of 50 yr) for the fitted Gumbel distribution and the spline model based on MC and bootstrap simulations of 10 000 samples each as well as the jackknife method. For the MC simulation the same performance measures are also given for the second largest value of each simulated 95-year sequence as an empirical estimate of the 98% quantile. The table shows that despite the smaller bias for the regional

Table 5.2. Bias, standard error and root-mean-square error (rmse) of the 98% quantile of the annual maximum precipitation deficit for the fitted Gumbel distribution and the spline model based on MC, bootstrap and jackknife samples. The results are expressed as a percentage of the 98% quantile under the spline model and represent averages of the six districts.

		Monte Carlo (MC)	Bootstrap	Jackknife
Gumbel	bias	−6.1	−6.8	−6.3
	std. error	5.5	5.1	5.1
	rmse	8.2	8.5	8.1
Spline	bias	−0.6	−3.0	1.6
	std. error	8.5	8.4	9.3
	rmse	8.5	8.9	9.4
Empirical	bias	1.0	-	-
	std. error	8.3	-	-
	rmse	8.4	-	-

spline model the Gumbel distribution has slightly smaller root-mean-square errors. In the case of the regional spline model the root-mean-square errors are almost completely determined by the standard error while for the Gumbel distribution the standard error contributes less than half to the root-mean-square error. The bootstrap and jackknife standard errors agree quite well with those from the MC simulation. Only for the spline the jackknife overestimates the standard error by 10%. The MC simulation shows that the empirical estimate of the 98% quantile performs equally well as the fitted regional spline model.

Thus in terms of root-mean-square errors of estimates of the 98% quantile the regional spline gives comparable results as the Gumbel distribution and the empirical estimates. The root-mean-square errors for the regional spline may in fact be larger because the optimization of the locations of the knots m_1 and m_2 is not accounted for in the MC experiment and the bootstrap and jackknife methods. The next section deals with time series simulation and how quantile estimates based on time series simulation perform in this respect.

5.4 Time series simulation

5.4.1 Nearest-neighbour resampling

The previous section was restricted to the annual maxima of the precipitation deficit. In this section the time series of decade values of the precipitation deficit are considered. Synthetic sequences of this deficit were generated by nearest-neighbour resampling. The nearest-neighbour algorithm used is closely related to the one in Chapter 4 (Beersma and Buishand, 2004). In the nearest-neighbour method the precipitation deficit is resampled with replacement from the historical data. The simplest way to incorporate temporal correlation is to condition resampling on the latest simulated value. This is done by searching the historical precipitation deficits that are similar to that value. One of these nearest neighbours or analogs is then randomly selected and its historical successor is the next simulated value. Since only a single characteristic, i.e., the latest simulated precipitation deficit, is used to generate the next precipitation deficit, this type of resampling is referred to as 1d-resampling. To incorporate longer-term variability in the simulated time series, the search for nearest neighbours is not only based on the precipitation deficit simulated in the previous (time) step but also on the average of the precipitation deficits simulated during the preceding 4 months. The latter acts as a memory for the simulation as in, e.g., Harrold et al. (2003a), Beersma and Buishand (2004; see also Appendix D of this thesis), and Leander et al. (2005). This type of resampling, where basically two characteristics are used to find the nearest neighbours is therefore referred to as 2d-resampling.

The nearest neighbours in the resampling procedure are selected in terms of an Euclidean distance. For 2d-resampling the two contributions to the Euclidean distance have weights inversely proportional to the variance of the two characteristics. One of the $k = 5$ nearest neighbours is selected randomly using the decreasing kernel introduced by Lall and Sharma (1996) and applied by Buishand and Brandsma (2001) and Beersma and Buishand (2004). A 7-decade wide moving window, centered on the latest simulated decade, is used to restrict the search for nearest neighbours to the season of interest. Note that a resampling technique cannot produce smaller or larger decade values than those found in the historical record. However, for periods longer than a decade, such as the summer half-year, the precipitation deficit can be larger than the largest historical deficit because of rearranging extreme decade values from different parts of the historical record. Moreover, extreme annual precipitation deficits are in fact due to new combinations of decades with large or moderately large precipitation deficits rather than a single decade with an unprecedented precipitation deficit because of the boundedness of precipitation by zero and the light tail of the distribution of potential evaporation (which is bounded by global radiation). This makes nearest-neighbour resampling suitable for exploring the upper tail of the distribution of the annual maximum precipitation deficits. This use of nearest-neighbour resampling requires that the simulated series are much longer than the return periods of the quantiles of interest. To meet this requirement time series of the decade precipitation deficit of 100 000 years were generated with the resampling procedure.

5.4.2 Accuracy of quantile estimates

The annual maxima in the 100 000 year generated sequence of precipitation deficits were compared with the historical annual maxima. The annual maxima of the resampled sequences were obtained from the 18 consecutive decade precipitation deficits that constitute the summer-half year in the same way as for the observed data. Figure 5.8 presents Gumbel plots of the annual maxima from 2d- and 1d-resampling for the two districts that differ most. The Gumbel plots clearly demonstrate that the 4-month memory in the 2d-resampling algorithm is needed to reproduce the heavy upper tail of the annual maximum precipitation deficit distributions. Figure 5.3 (Section 5.2) presents for each of the six districts Gumbel plots of the annual maximum precipitation deficit from 100 000-year sequences generated with the 2d-resampling model.

Quantile estimates are derived 'empirically' from the ordered annual max-

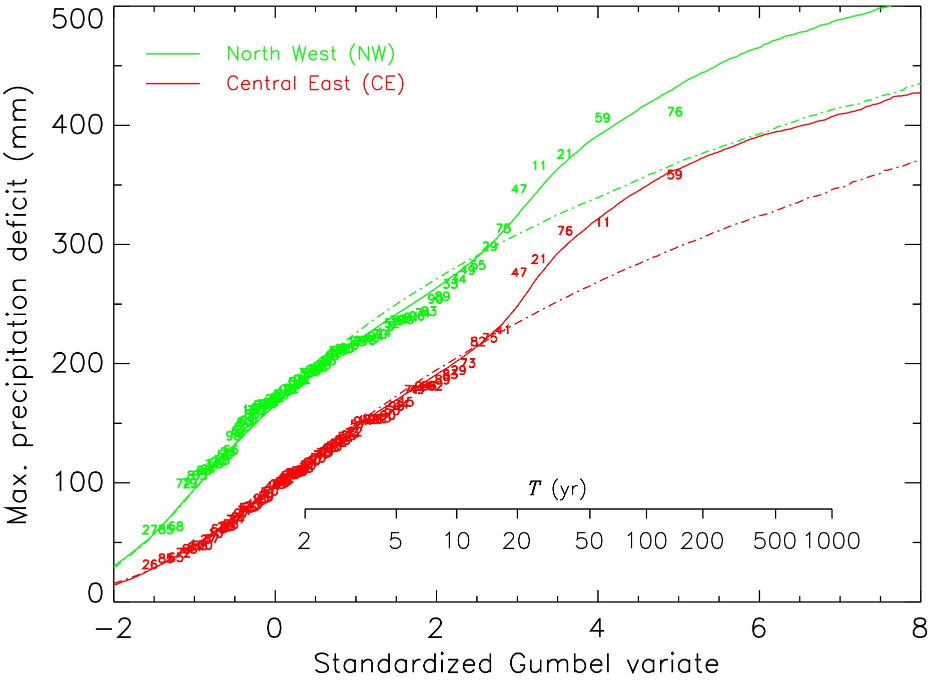

Figure 5.8. Gumbel probability plots of the annual maximum precipitation deficit for two districts together with plots of the annual maxima from 100 000-year sequences generated with the 2d-resampling (solid) and the 1d-resampling (dash-dot) models. The coloured numbers represent the year minus 1900.

imum precipitation deficits in the 100 000-year simulated series. In contrast to the regional frequency analysis of the previous section there is no underlying parametric model from which Monte-Carlo samples can be drawn to determine biases and standard errors. However, the bootstrap and jackknife can still be used to obtain the standard error of quantile estimates based on nearest-neighbour resampling. The combination of nearest-neighbour resampling with a bootstrap or jackknife procedure is closely related to the double-bootstrap and the jackknife-after-bootstrap methodologies (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). In the case of the bootstrap, N different 95-year (bootstrap) samples are generated by choosing years randomly with replacement from the historical record, and in the case of the jackknife 95 samples of 94 years are considered in which a single historical year is omitted. The resampling procedure then uses these (sub)samples of the historical data to generate long time series from which the quantiles are again empirically derived. For the jackknife thus 95 quantile estimates from the 95 jackknife samples are obtained while for the bootstrap N quantile estimates from the corresponding bootstrap samples are derived. Standard errors of quantile estimates, finally, are estimated in the same way as in Section 5.3.2. Note that these bootstrap and jackknife procedures are computer intensive, since instead of a single 100 000-year simulation with the nearest-neighbour resampling model, N (bootstrap) or 95 (jackknife) 100 000-year simulations are needed to determine the accuracy of the estimated quantiles. In the case of the bootstrap it is possible to reduce the number of simulations by requiring that all 95 historical years are equally represented. This variant of the bootstrap is known as the balanced bootstrap (Efron and Tibshirani, 1993). To obtain a comparable amount of simulated data for both procedures, $N = 499$ balanced bootstrap samples were resampled for 100 000 years while the 95 jackknife samples were resampled for 500 000 years. The bootstrap and jackknife standard errors of the 98% quantile obtained in this way are respectively 9.0 and 9.9%. These standard errors are slightly larger than those for the spline model in Table 5.2, and again the jackknife standard error is slightly larger than the bootstrap standard error. A statistical model that describes the distribution and dependence of the precipitation deficits is needed to determine how reliable these bootstrap and jackknife standard error estimates are. In the next subsection a systematic analysis of the combination of nearest-neighbour resampling with the bootstrap and jackknife procedures is given using synthetic data obtained with an autoregressive time series model.

5.4.3 Reliability of jackknife and bootstrap standard error estimates

In this section a first-order autoregressive, AR(1), process is introduced to investigate: (i) the accuracy of quantile estimates of the distribution of the annual maximum precipitation deficit based on resampling of different 95-year sequences of decade precipitation deficits, and (ii) the ability of the bootstrap and jackknife methods to estimate the accuracy of these quantile estimates from a single 95-year sequence.

The AR(1) model preserves the lag 1 autocorrelation coefficient ($r(1) = 0.128$) of the observed decade precipitation deficits during the summer half-year. In contrast to Sections 5.4.1 and 5.4.2, where precipitation deficits for different districts were considered, a single precipitation deficit is generated representing the country-average precipitation deficit. Details of the AR(1) model are given in Appendix G. The annual maxima of the simulated precipitation deficits were obtained from the 18 consecutive decade precipitation deficits that constitute the summer half-year in the same way as for the ob-

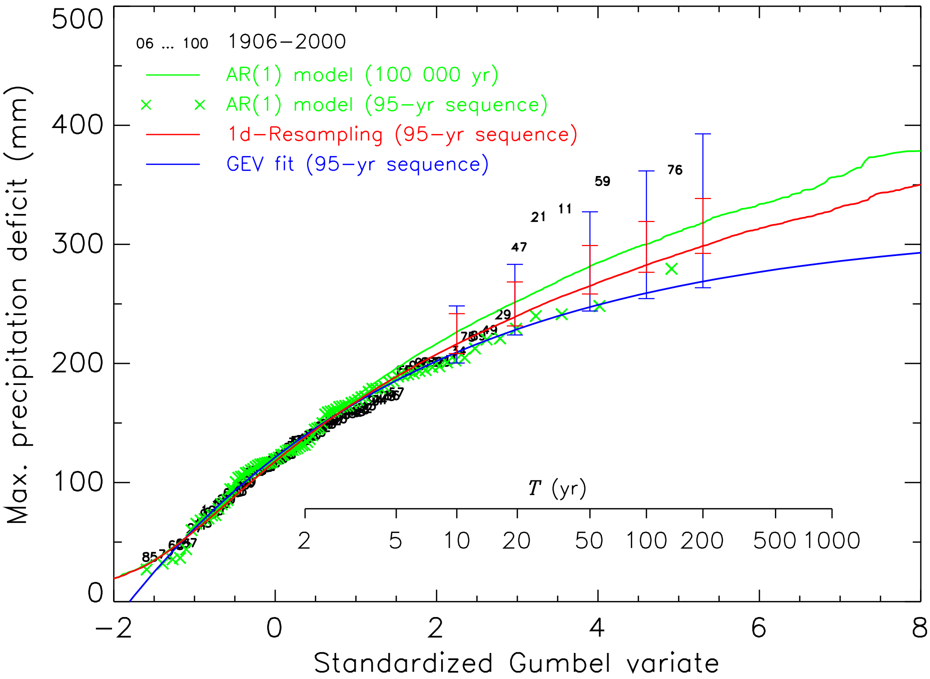

Figure 5.9. Gumbel probability plot of the annual maximum country-average precipitation deficit together with plots of the annual maxima of a 100 000-year sequence simulation with the AR(1) model fitted to the observed decade precipitation deficits (green), the annual maxima of a single 95-year sub-sequence of the 100 000-year AR(1) simulation (green crosses), the annual maxima of a 100 000-year sequence generated with the 1d-resampling model using the same 95-year sequence (red), and a GEV distribution fitted to this 95-year sequence (blue). Error bars indicate 98% confidence intervals based on the 999 different 95-year sequences of the 100 000-year AR(1) simulation (red for 1d-resampling and blue for the GEV distribution).

served and resampled data. Figure 5.9 presents a Gumbel probability plot of the annual maximum of the country-average precipitation deficit together with a plot of the annual maxima from a 100 000-year simulation with the AR(1) time series model (represented by the green curve). Up to return periods of about 20 years there is a very good correspondence between the AR(1) model and the observed precipitation deficits but the AR(1) model does not reproduce the heavy upper tail of the observations. Despite this difference the AR(1) model will be regarded to represent the true distribution of the annual maximum precipitation deficit for the remainder of this section. Figure 5.9 also shows Gumbel plots of the annual maxima of a single 95-year sub-sequence of the 100 000-year AR(1) simulation (green crosses), a GEV distribution fitted to this 95-year sequence (blue curve) and the annual maxima from a 100 000-year sequence generated with the 1d-resampling model using the same 95-year sequence (red curve). Note that the distribution obtained with the 1d-resampling model is much closer to the true distribution than the (fitted) GEV distribution. This is generally true as is demonstrated with the following experiment.

The first 94 905 years of the AR(1) simulation were split into 999 non-overlapping 95-year sequences. For each 95-year sequence, the 98% quantile of the annual maximum precipitation deficit was estimated by fitting a GEV distribution to the 95 annual maxima of the sequence and by generating a 100 000-year sequence of decade precipitation deficits with the 1d-resampling model. Fitting a GEV distribution results in a standard error of the 98% quantile of 6.1% while generation of a 100 000-year series with the 1d-resampling model results in a standard error of only 3.1%. For the 98% quantile, resampling is thus almost twice as accurate as fitting a GEV distribution; in other words resampling has a gain in accuracy of almost 50%. This gain depends on the quantile of interest because for the GEV distribution the relative standard error of the estimated quantile rapidly increases with increasing return period simply because of the growing influence of the uncertain shape parameter on the quantile. The gain of resampling varies from 20% for the 90% quantile to more than 60% for the 99.5% quantile. The 98% confidence intervals in Figure 5.9 also show these differences in accuracy of quantile estimates. The reduction of the variance of quantile estimates as a result of resampling is extensively described by Buishand (2006).

The 100 000-year AR(1) simulation was further used to test the reliability of the bootstrap and jackknife procedures that were employed in Section 5.4.2 to estimate the accuracy of quantile estimates derived from the resampled observations. For each of the above 999 different 95-year AR(1) sequences the bootstrap or jackknife procedure could be performed. However, to perform the full experiment with N bootstrap samples and 95 jackknife samples for each of the 999 different 95-year sequences is too expensive to run on most present day computers taking into account that for each bootstrap and jackknife sample series with a length of ∼100 000 years need to be generated with the resampling model. The bootstrap experiment (with $N = 199$ balanced samples and 100 000 years nearest-neighbour resampling) and the jackknife procedure (with 95 samples and 200 000 years nearest-neighbour resampling) were therefore only performed for the first 20 of the 999 different 95-year sequences. The average and the standard deviation of the bootstrap and jackknife estimates of

Table 5.3. Relative standard errors (%) of the 98% quantile of the annual maximum precipitation deficit in an AR(1) model for the country-average precipitation deficits. The quantile is estimated by nearest-neighbour resampling of 95-year AR(1) sequences, and its standard error is based on the sample standard deviation of the quantile estimate from the 999 different 95-year AR(1) sequences (MC), the bootstrap or the jackknife. The values for the bootstrap and the jackknife are the averages of 20 estimates from different 95-year AR(1) sequences; the values in parentheses give the standard deviations of these estimates.

	True (MC)	Bootstrap	Jackknife
1d-resampling	3.1	5.0 (1.5)	3.6 (0.6)
2d-resampling	5.1	5.3 (1.4)	6.0 (1.6)

the relative standard error of the 98% quantile from resampling these 20 different 95-year sequences are presented in Table 5.3 together with the true standard error obtained from resampling all 999 different 95-year sequences from the full AR(1) simulation. The table shows that in the case of 1d-resampling the bootstrap and the jackknife overestimate the standard error of the 98% quantile, in particular the bootstrap for which the overestimation is as large as 60%. Including a 4-month memory in the resampling model (2d-resampling), as was done for the precipitation deficit data of the six districts, leads to a 60% larger standard error. The difference in the standard error between the 1d- and 2d-resampling models depends, however, on the quantile of interest. It increases from 20% for the 90% quantile to 90% for the 99.5% quantile. In the case of 2d-resampling the bootstrap slightly outperforms the jackknife but more important, the standard errors in Table 5.3 are almost a factor of two smaller than the bootstrap and jackknife standard errors of respectively 9.0 and 9.9% found in Section 5.4.2 for nearest-neighbour resampling of the observed precipitation deficits. The standard deviations of 1.4 to 1.6% of the bootstrap and jackknife estimates in Table 5.3 indicate that this large difference is unlikely due to random variations of these estimates. Something that presumably contributes to this difference is that the distribution of the observed annual maximum precipitation deficits has a heavier upper tail than that simulated with the AR(1) model. Despite this discrepancy, the AR(1) simulation provides a good picture of the quality of the bootstrap and jackknife standard error estimates of estimated quantiles from nearest-neighbour resampling. In contrast with the results in Table 5.2, where the bootstrap and jackknife were used in combination with parametric models, the bootstrap and jackknife in combination with nearest-neighbour resampling give rather uncertain and occasionally somewhat biased results.

5.5 Discussion and conclusions

Regional frequency analysis and time series simulation by means of nearest-neighbour resampling were studied to estimate a large quantile of the distribution of the annual maximum precipitation deficit. A regression spline was used to describe the regional growth curve. Regional estimation of the common parameters of this spline only slightly reduces the uncertainty of large quantile estimates because of the relatively small number of districts (six) and the strong spatial correlation of the annual maximum precipitation deficits. To reproduce the heavy upper tail of the distribution of the annual maximum precipitation deficits with nearest-neighbour resampling, a 2d-resampling model was needed, i.e., a model with an additional 4-month memory to find the nearest neighbours. A disadvantage of this memory term is that the standard error of quantile estimates is considerably larger than for 1d-resampling. To determine the standard error of quantile estimates in the case of resampling, bootstrap or jackknife techniques are required. The combination of nearest-neighbour resampling with these techniques is computationally expensive and the results turn out to be rather uncertain and possibly somewhat biased.

Relative standard errors of the estimate of the 98% quantile were compared in Sections 5.3 and 5.4. For other quantiles the results can be quite different as is shown in Figure 5.10. The figure shows a strong increase of the relative standard errors for the spline model in the region of the curvature of the regional growth curve. After a local minimum near the 70-year event, the standard error increases steadily up to 17% for the 1000-year event. This increase is obviously related to the relatively large uncertainty of the parameters a and b in the spline model. Apart from a larger amplitude, the standard errors of quantile estimates obtained with 2d-resampling show a similar behaviour as those for the spline model up to the 100-year event, but for larger quantiles the relative standard error slowly decreases to approximately 6% for the 1000-year event, which is only slightly larger than the relative standard error for the 10-year event. For very large quantiles 2d-resampling thus clearly outperforms the spline model. In the region around the 30-year event, where the standard errors are large, the empirical sample quantile performs equally well as the spline model and 2d-resampling. For 2d-resampling, the bootstrap provides a considerably smaller standard error of the estimated quantiles in

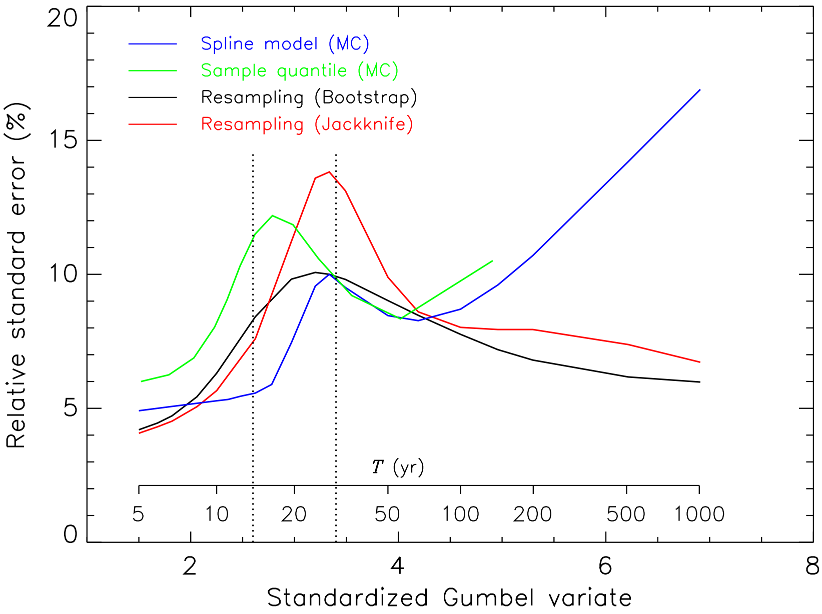

Figure 5.10. Relative standard errors of the estimated quantiles of the annual maximum precipitation deficit from a spline approximation to the regional growth curve, from the sample quantiles under the assumption that the regional spline represents the true distribution and from 2d-resampling. The relative standard errors are averages over the six districts. The dotted lines represent the position of the two knots of the spline model.

this region than the jackknife. Beyond this region the jackknife and bootstrap standard error estimates are of comparable size.

Thus, for return periods between 20 and 100 years, the standard errors of the estimated quantiles are relatively large, which seems to be related to the 'uncertain' curvature of the distribution of the annual maximum precipitation deficits. Up to the 97% quantile (which corresponds roughly with a return period of 30 years) the spline model turns out to be superior. But, for very large quantiles, where sample quantiles are unavailable and extrapolation is required, the uncertainty of quantile estimates obtained with 2d-resampling becomes much smaller than the uncertainty of quantile estimates from the spline model.

This discussion is concluded with a practical example of 2d-resampling in combination with the bootstrap. Bootstrap confidence intervals are given for the return period of the largest observed precipitation deficit for each of the six districts, i.e., the 1959 drought for the CE and NE districts and the 1976 drought for the other districts. The return periods of these events were

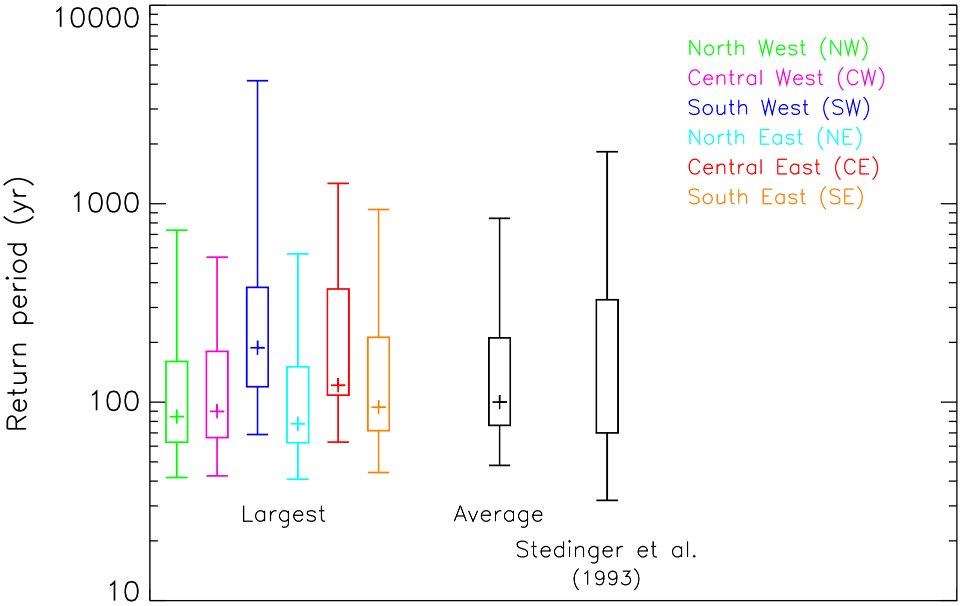

Figure 5.11. Confidence intervals for the return period of the largest precipitation deficit in each of the six districts obtained with the combination of 2d-resampling and the bootstrap procedure in Section 5.4.2. The boxes and the whiskers represent the 50% and 90% confidence interval respectively. The coloured plusses represent for each district the return period obtained from a 100 000-year sequence generated with the 2d-resampling model. For comparison the harmonic average of the six confidence intervals and the 50% and 90% confidence bounds of the return period associated with the largest value in a sample of size 95 (Stedinger et al., 1993) are given as well.

empirically derived from the 100 000-year sequences generated with the 2d-resampling model for each of the 499 bootstrap samples in Section 5.4.2. The bootstrap confidence bounds in Figure 5.11 are based on the percentiles of these bootstrap estimates. Besides the 50% and 90% confidence bounds for the largest precipitation deficit in each of the six districts and their harmonic averages, the figure also presents the corresponding bounds of the return period associated with the largest value in a sample of size 95 as given by Stedinger et al. (1993). The confidence intervals are quite wide, e.g., the 90% confidence interval for the 1976 drought in the NW district ranges between 42 and 735 years. The average 50% and 90% confidence intervals are narrower than the corresponding intervals for the return period of the largest value in a sample of size 95. Both intervals differ in particular regarding their upper limits. The width of the 2d-resampling bootstrap confidence interval is on average more than a factor of two smaller (on a linear scale) than that of the interval obtained without use of the information in the data.

Acknowledgments. We thank Prof. Bert Holtslag at Wageningen University for valuable comments on a draft of this paper, and for the suggestion to explore the nature of the drought related feedback in more detail. The work was performed in the framework of the ‘Drought studies of the Netherlands’ in cooperation with the Institute for Inland Water Management and Waste Water Treatment (RIZA), Lelystad under contracts RI-2726A and RI-4339.

Chapter 6

Summary and synthesis

6.1 Summary

Resampling techniques are used to determine statistical uncertainty and to simulate very long hydro-meteorological time series that contain unprecedented extreme events. An advantage of a frequency analysis of the extremes of such long time series is that the statistical uncertainty of the result (e.g., a T-year event) is generally smaller than for a frequency analysis of the observed extremes only. In addition, resampling techniques do not need various, more or less false, assumptions about the statistical properties of the data.

By combining time series of resampled hydro-meteorological data with suitable hydrological models, this time series resampling approach can be used for a wide range of studies regarding the effects of extreme hydro-meteorological events such as extremely high or low river discharges, extremely high or low ground-water levels or other extreme hydrological events. An additional advantage of this approach is that hydro-meteorological effects can be distinguished from human induced changes in the hydrological system such as canalization, urbanization and deforestation.

The first application of resampling (Chapter 2) involves the statistical uncertainty of variance estimates. The standard error (statistical uncertainty) of the variance is determined with the jackknife, and is used to develop a test for equality of variances of monthly values which can be useful for both model validation and climate change assessment. Since extreme events are related to the variance, it can even be used as a first indication in the detection of changes in extreme events. The test is of great practical use because it has a simple and straightforward multivariate extension, meaning that it can be applied to an area, consisting of several locations without complications. As an example, the multivariate test is applied to precipitation and temperature variances in the UKTR climate change simulation with the first Hadley Centre (UK) coupled ocean-atmosphere General Circulation Model in three regions: central North America, southern Europe and northern Europe. In particular, a significant increase (at the 5% level) of the variance of monthly precipitation over northern Europe is found in the climate change simulation for winter, summer and autumn. Apart from these increases, a significant decrease of the variance of the monthly near-surface temperature in spring is detected.

In Chapter 3 multi-site sequences of daily precipitation and temperature in the German part of the Rhine basin are resampled conditional on the large-scale atmospheric circulation. The atmospheric circulation then acts as a predictor for precipitation and temperature, i.e., the circulation determines largely whether a day will be wet or dry and whether a day will be relatively warm or cold. In the case of a systematic change in circulation, e.g., due to anthropogenic climate change, this will automatically lead to a change in (extreme) precipitation and temperature. Conditional nearest-neighbour resampling is therefore potentially useful for climate change applications. Different conditional nearest-neighbour resampling models are compared with each other and with the so-called analog method (Zorita et al., 1995; Zorita and von Storch, 1999). Despite the fact that the analog method is originally a deterministic method and conditional nearest-neighbour resampling a stochastic one the two methods are closely related. An important conclusion of the comparison of conditional nearest-neighbour resampling and a stochastic version of the analog method is that the simulation of precipitation and temperature for a new day should not only be conditioned on the circulation characteristics of that day (as happens by definition in the analog method) but also on the simulated precipitation and temperature for the previous day, in order to achieve the appropriate level of persistence and variability in the generated time series.

It should also be recognized that for climate change applications, which was a motivation for conditional nearest-neighbour resampling, there are a few limitations regarding its use. A serious one is that (future) changes in precipitation may not result from changes in the atmospheric circulation alone as is shown by several climate change simulations with general circulation models (van Ulden and van Oldenborgh, 2006; van Ulden et al., 2007), and by simulations of historical precipitation trends by means of conditional nearest-neighbour resampling (Beersma and Buishand, 1999a).

In 980-year multi-site simulations of daily precipitation and temperature, representative of the current climate, up to 35% larger 10-day area-average precipitation amounts are produced than observed in the historical 35-year reference period. In combination with rainfall-runoff models, such unprecedented precipitation extremes can be very useful for determining extremely rare river discharges and thus for planning and design of the hydrological infrastructure.

Chapter 4 deals with (extreme) drought. The joint probability of precipitation deficits in the Netherlands and discharge deficits of the river Rhine is the central theme in this chapter. These joint probabilities are estimated using nearest-neighbour resampling and compared with estimates obtained from fitting bivariate probability distributions. The asymptotic dependence structure between precipitation and discharge deficits was found to play a crucial role in estimating the joint exceedance probabilities. A clear advantage of nearest-neighbour resampling is that it does not need assumptions about the dependence structure since it is inherited from the data. In the framework of fitting bivariate probability distributions, satisfactory results could be achieved only by introducing a new bivariate distribution which is a mixture of a bivariate normal and a bivariate Gumbel distribution, i.e., a bivariate normal distribution with a (logistic) Gumbel dependence structure. This bivariate distribution gives a better estimate of the probability that the precipitation and discharge deficits are both extreme than nearest-neighbour resampling. This is due to the fact that the upper tails of both marginal (i.e., univariate) distributions are not properly reproduced by the nearest-neighbour resampling model.

Thus, nearest-neighbour resampling performs superior regarding the dependence structure, but overall the novel bivariate distribution turns out to be more suitable to estimate the joint exceedance probabilities of large precipitation and discharge deficits. Based on this bivariate distribution small probabilities are found (once every 200–300 years) for the joint exceedance of the precipitation deficit and the discharge deficit that occurred during the driest historical years (1921 and 1976). When a failure region based on the economic damage is used the probability of these droughts increases to about once every 100 years. Besides, the differences between the probabilities of drought obtained with the different methods are smaller with such a failure region than with the ordinary joint exceedances of the corresponding precipitation and discharge deficits.

Spatial variation in the probability distribution of precipitation deficits in the Netherlands is the subject of Chapter 5. This spatial variation is addressed by dividing the Netherlands into six districts. Apart from differences in the return levels, the probability distributions for all of the six districts have a common extraordinary curvature in the upper tail. These distributions resemble neither lognormal nor extreme value distributions. To reproduce both the differences in return levels between the districts and the extraordinary curvature in the tail two alternatives are considered: a regional frequency analysis and time series simulation by nearest-neighbour resampling. The regional frequency analysis yields a regional growth curve that can be approximated by a spline on the Gumbel scale. To reproduce the common curvature in the upper tail of the distribution the resampled data need additional long-term persistence. This is achieved by introducing a 4-month memory term in the resampling procedure. This memory term leads to an increase of the statistical uncertainty (standard error) of large return levels of the precipitation deficit. But when the two methods are used for extrapolation (larger return levels than observed), nearest-neighbour resampling still outperforms the regional frequency analysis. This is due to a strong increase in the relative standard error of return level estimates with increasing return period as a result of the large uncertainty in the parameters of the spline that approximates the regional growth curve. In terms of statistical uncertainty nearest-neighbour resampling thus outperforms a regional frequency analysis. And this is the type of results that demonstrates the added value of resampling.

Using nearest-neighbour resampling and the bootstrap procedure, confidence intervals are constructed for the return periods of the largest observed precipitation deficit for each of the six districts. Depending on the district, the largest precipitation deficit occurred in 1959 or 1976. Although these confidence intervals are quite wide, they are on average a factor of two narrower than the expected interval of the largest observation in a series of this length.

6.2 Synthesis

The length and the specific period of the data sample that is used for any statistical analysis has an influence on the accuracy of the end-result. This also holds for results obtained with resampling procedures. The statistical uncertainty generally decreases if the length of the reference record that is used for resampling increases. A related point is the representativity of the reference period used for resampling. Without trends, one can generally say that the longer the reference period, and the more natural variability may be captured, the more representative the data.

In practice, we generally have not much more than one hundred years of recorded instrumental hydro-meteorological data. The record length usually decreases to only 30–50 years when the data are needed with sufficient temporal (e.g., daily) resolution and sufficient spatial coverage. In Chapter 3 the resampling models make use of 35 years (1961–1995) of daily precipitation and temperature for a number of stations in the German part of the Rhine basin. In Chapters 4 and 5 daily precipitation and (potential) evaporation data for 95 years (1906–2000) are used which are supplemented in Chapter 4 with daily discharges of the river Rhine at gauging station Lobith for the same 95-year period.

The bootstrap and jackknife procedures that are presented and applied in Chapter 5 make it possible to systematically investigate the statistical uncertainty resulting from the reference period being used. In combination with nearest-neighbour resampling the bootstrap is used to construct confidence intervals for the return periods of the largest observed precipitation deficit in each district. In general, it should be a challenge to supply extreme quantiles, return levels or return periods from nearest-neighbour resampling with such bootstrap or jackknife uncertainty analyses. This can and should make users more aware of the (large) uncertainty involved in 'extrapolating' extreme events from relatively small historical samples. In future work, it would be interesting in this respect to perform similar uncertainty analyses based on the, relatively short, 35-year reference period that is used in Chapter 3. One should however be aware that such uncertainty analyses can become quite computer intensive. Moreover, the results in Chapter 5 show that the resulting uncertainty estimates can be biased.

In Chapter 4 daily precipitation data for the 95-year reference period from thirteen stations are used. For the analysis in Chapter 5 this set is augmented with five stations. Since no long (homogeneous) series of daily evaporation exist, such a long series is constructed from the Makkink formula for potential evaporation using daily global radiation *estimated* from the daily sunshine duration at station De Bilt. Sunshine duration and global radiation are closely related since the latter mainly depends on the presence of clouds and the solar elevation. This is confirmed by measurements of daily global radiation at De Bilt which are available from 1958 onwards. For the period 1958–2000 there is a good correspondence between the Makkink evaporation based on the measured global radiation and that based on the global radiation estimated from the sunshine duration at De Bilt (Beersma and Buishand, 2004; Chapter 4 of this thesis, page 60). For the regional analysis in Chapter 5 the single evaporation series is scaled based on the spatial variation in the climatological mean (1971–2000) of the Makkink evaporation. And although this scaling of the evaporation series can be justified, since the spatial variability of precipitation is much larger than that of evaporation, this situation is not desirable.

The lack of long (daily) evaporation series has a historical background. For several practical reasons KNMI decided in 1987 (CHO-TNO, 1988) to provide routinely daily (potential) evaporation based on the Makkink formula rather than the Penman open water evaporation which had been provided until then. For De Bilt the Makkink evaporation series was extended back until 1958, i.e., the year in which the global radiation measurements started. For extension further back in time (both useful for trend analysis and extreme value analysis) there seems no other option than to use estimated global radiation from sunshine duration as performed in Chapter 4 and used in Chapter 5.

Besides for De Bilt, daily sunshine duration is available from 1900 or slightly earlier for four other principal climatological stations in the Netherlands (Beek, De Kooy, Eelde and Vlissingen). Some corrections for the homogeneity of these series might be necessary. With a reasonable effort it seems possible to supplement the long series of estimated daily Makkink evaporation for De Bilt with four similar series for different locations in the Netherlands. Apart from this, there is at least one known measurement site in the Netherlands, i.e., station 'Duivendaal-Haarweg' from Wageningen University, for which a much longer daily global radiation series exists than that for De Bilt. For the growing season (April–September) this series already starts in 1928. From 1938 onwards the data are available throughout the year. Although, the design of the solarimeter instrument changed a few times, the site was relocated in 1980, and the series contains some periods with missing data (de Bruin et al., 1995), it would be interesting to check its quality and homogeneity. Due to its length this series is quite unique and could be very useful for the type of analyses described in Chapters 4 and 5.

Very long time series simulated with nearest-neighbour resampling are *stationary* in nature, unless resampling is conditioned on a non-stationary predictor. There is a good reason for the emphasis on 'stationary' in the previous sentence. Although stationary time series are simulated we know that climate is in fact not stationary, at least not in the long run. Stationarity largely depends on the time scale that one considers. For practical reasons climatologists often look at 30-year periods, for which they assume that the climate is stationary. But even though the climate might seem stationary in a 30-year period this is not necessarily so for a 100-year or longer period. And if the time scale is extended to that of the ice-ages climate is definitely not stationary. This brings us to the issue of climate variability versus climate change. Two terms both related to non-stationarity but with a very thin distinction in meaning. In the IPCC view climate change refers to a persistent and statistically significant change in the mean climate or its variability, while climate variability refers to variations (either natural of anthropogenic) in the mean state (see glossary in

IPCC, 2001). Other terminology refers to internal forcings, or fluctuations, of the climate system as climate variability and to external forcings, e.g., due to anthropogenic emission of greenhouse gases and aerosols, as climate change. In reality both type of forcings act simultaneously and one effect might mask the other which makes it difficult to distinguish, or even detect, them.

Climate is thus not stationary in the long run and, due to anthropogenic climate change, probably not even in the near future. A relevant example of the non-stationarity of climate in the recent history is the winter rainfall in the Netherlands and over large parts of the Rhine basin which shows significant increasing trends over the 20th century (Rapp and Schönwiese, 1995; Widmann and Schär, 1997; Schmidli et al., 2002; Klein Tank and Sluijter, 2003; Hundecha and Bárdossy, 2005). The causes of these trends are not fully understood yet but they are assumed to be related to both natural variability and global warming. Apart from the causes, these trends are often thought to be the result of changes in the large scale atmospheric circulation. By conditioning nearest-neighbour resampling of precipitation and temperature on the observed atmospheric circulation for the period 1891–1995, in a similar way as described in Chapter 3 for the sub-period 1961–1995, Beersma and Buishand (1999a) showed that the (change in) atmospheric circulation can explain on average only slightly more than 50% of the trends in mean winter rainfall over the German part of the Rhine basin.

Should such trends in the observations be corrected if these observations are used to generate long stationary time series by means of nearest-neighbour resampling? Detrending of the observed data is only useful if a trend is caused by a persistent change of climate. In that case the data should be detrended towards 'the time period in the observations for which the resampled data should be representative'. For the resampling conditional on the 35 years of historical circulation indices in Chapter 3, the trends in the winter precipitation over the German part of the Rhine basin were not corrected because part of these trends can be explained by long-term changes in these indices (see above) and because the cause for the remaining trend is unknown. In the historical time series of summer precipitation and evaporation in the Netherlands that were used for nearest-neighbour resampling in Chapters 4 and 5 no trends were detected. A trend in the observed data used for resampling that cannot be explained adds (some unknown) uncertainty to the end-result. The uncertainty obtained with jackknife and bootstrap procedures then underestimates the true uncertainty.

Despite the intrinsic non-stationarity of climate, the need for long stationary hydro-meteorological time series remains. Such long time series are certainly not meant as 'predictions' into the far future, rather they are representative of the ‘current’ or a particular historical state of the climate system. The great length of the time series is needed to represent extreme events with a very low probability of occurrence which in turn are needed for the design of protective measures. In a period of (anthropogenic) climate change the generated time series may be representative of relatively short periods only, say a few decades, depending on the speed of the climate change.

Apart from the application of the variance test (in Chapter 2) and recognition of the potential use of conditional nearest-neighbour resampling (Chapter 3) in the context of climate change, future climate change was not considered in this thesis. At the same time, the demand for information regarding the future changes in extreme (hydro-meteorological) events due to anthropogenic climate change is steadily increasing. And although resampling can definitely be of help here, it is considered a subject on its own. Work on this topic for the Meuse basin is in progress (Leander and Buishand, 2007).

Appendix A

Estimation of correlation between pseudovalues

The estimates of the correlation coefficients ρ in Tables 2.2 and 2.3 were obtained from the Monte Carlo experiment as follows. Let $\overline{\theta^*_{j,m}}$ be the average pseudovalue for year j in the mth simulation ($j = 1, \ldots, J$; $m = 1, \ldots, M$), taken over N independent sequences. A natural estimate of ρ is then:

$$\hat{\rho} = \frac{2 \sum_{m=1}^{M} \sum_{i=1}^{J} \sum_{j=1}^{i-1} \left(\overline{\theta^*_{i,m}} - \overline{\theta^*}\right)\left(\overline{\theta^*_{j,m}} - \overline{\theta^*}\right)}{MJ(J-1)\hat{v}}, \tag{A.1}$$

where

$$\overline{\theta^*} = \frac{1}{JM} \sum_{j=1}^{J} \sum_{m=1}^{M} \overline{\theta^*_{j,m}} \quad \text{and}$$

$$\hat{v} = \frac{1}{JM} \sum_{j=1}^{J} \sum_{m=1}^{M} \left(\overline{\theta^*_{j,m}} - \overline{\theta^*}\right)^2 .$$

For computational purposes it is more convienient to obtain $\hat{\rho}$ by (Koch, 1983):

$$\hat{\rho} = (J\hat{v}_{\mathrm{b}} - \hat{v})/[(J-1)\hat{v}], \tag{A.2}$$

where

$$\hat{v}_{\mathrm{b}} = \frac{1}{M} \sum_{m=1}^{M} \left(\overline{\theta^*_{\cdot m}} - \overline{\theta^*}\right)^2, \tag{A.3}$$

with

$$\overline{\theta^*_{\cdot m}} = \frac{1}{J} \sum_{j=1}^{J} \overline{\theta^*_{j,m}} \tag{A.4}$$

the mean of the pseudovalues in the mth simulation. Equation (A.2) can not be used to estimate ρ from a single record, because the result $\hat{\rho} = -1/(J-1)$ for $M = 1$ does not depend on the true value of ρ.

Appendix B

Properties of kurtosis estimates

In a Monte Carlo study, Pearson (1935) observed that kurtosis estimates can be heavily biased. For the normal distribution, it can be shown (Cramér, 1946) that $E(\hat{\gamma}_2) = -6/(J+1)$. Table B.1 compares the mean of $\hat{\gamma}_2$ with the true kurtosis γ_2 for sample sizes encountered in Chapter 2. For leptokurtic

Table B.1. Mean (first row) and standard deviation (second row) of kurtosis estimates for sequences of independent observations from various distributions (5000 simulations).

		$\hat{\gamma}_2$, Eq. (2.3)			$\hat{\gamma}_2$, Eq. (2.16)
Distribution	Kurtosis	$J = 5$	$J = 10$	$J = 30$	$J = 10, n_s = 3$
Uniform	−1.2	−1.11	−1.01	−1.11	−0.97
		0.53	0.54	0.26	0.33
Normal	0	−1.00	−0.54	−0.19	−0.21
		0.50	0.76	0.71	0.69
χ^2_{12}	1	−0.98	−0.42	0.26	0.17
		0.53	0.95	1.35	1.21
χ^2_4	3	−0.94	−0.16	1.12	0.89
		0.56	1.21	2.12	1.86
Laplace	3	−0.87	0.08	1.41	1.15
		0.54	1.13	1.86	1.67
Exponential	6	−0.88	0.19	2.27	1.88
		0.62	1.46	2.94	2.56

distributions ($\gamma_2 > 0$) the true kurtosis is seriously underestimated. The bias grows with increasing γ_2. For $J = 5$, $E(\hat{\gamma}_2) \approx -1$ for all distributions considered in Table B.1, no matter their true kurtosis. This bias is partly caused by the boundedness of $\hat{\gamma}_2$. Expressions for the bounds of standardized sample moments are given in Dalén (1987). The upper bound of $\hat{\gamma}_2$ is 0.25, 5.11 and 25.03 for $J = 5$, 10 and 30, respectively. The sample kurtosis of a sample of size 5 from a Laplace distribution is thus always smaller than the true kurtosis.

The last column in Table B.1 gives the mean of the pooled estimate $\hat{\gamma}_2$ in equation (2.16) for $n_s = 3$ samples of size 10 from the same distribution. The bias is roughly of the same order as that in a single sample of size 30. Note that for this sample size, $E(\hat{\gamma}_2) \approx 1$ for the two distributions with $\gamma_2 = 3$.

Besides the bias, the large variability of kurtosis estimates is a point of concern. For the leptokurtic distributions in Table B.1, the standard deviation of $\hat{\gamma}_2$ increases with increasing J as a result of the growth of its upper bound. It is only for larger samples than those in Table B.1 that $\text{var}(\hat{\gamma}_2)$ becomes proportional to $1/J$. Further, the standard deviation of the pooled estimate of 3×10 observations is somewhat smaller than that of a single sample of size 30. Spatial averaging over grid points will strongly reduce the standard deviation of $\hat{\gamma}_2$.

Appendix C

Simulation of daily circulation indices

C.1 Introduction

A typical feature of air pressure is that the day-to-day variability is relatively small during periods of high pressure and generally large during periods of low pressure. Such state-dependent behaviour cannot be reproduced by classical autoregressive (AR) processes. Al-Awadhi and Jollife (1998) therefore studied the use of threshold autoregressive (TAR) models to describe time series of surface pressure in the UK. Zwiers and von Storch (1990) applied this class of models to time series of the Southern Oscillation index. It seems reasonable to suspect that the statistical properties of the circulation indices (which are based on air pressure maps) are also state dependent. In contrast to the univariate applications of the TAR models mentioned above usually more than one index is needed to characterize the atmospheric circulation. Lall and Sharma (1996) showed that nearest-neighbour resampling is able to reproduce the nonlinear behaviour of a TAR model. Because the extension of nearest-neighbour resampling to the multivariate situation is straightforward, this method was used to simulate time series of the daily circulation indices Z, W and S.

C.2 Model construction

Three unconditional nearest-neighbour simulation models for generating Z, W and S were examined. The different feature vectors $\mathbf{D}_t$ are schematically

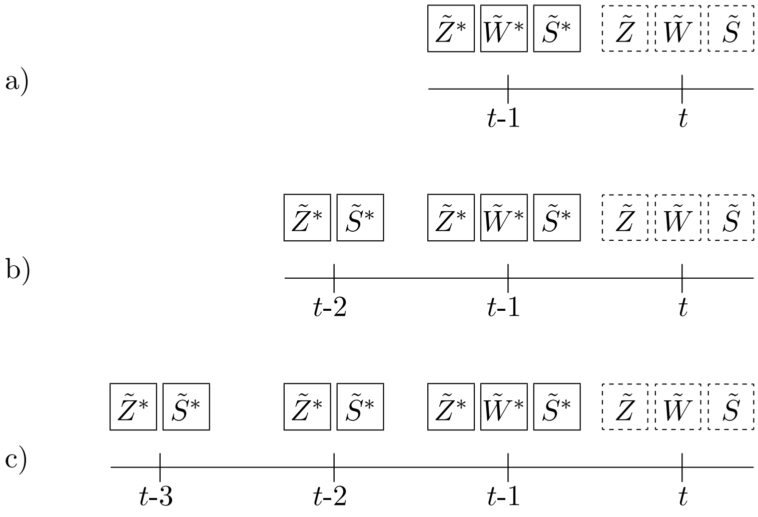
a)
$\tilde{Z}^*$ $\tilde{W}^*$ $\tilde{S}^*$ $\tilde{Z}$ $\tilde{W}$ $\tilde{S}$
t-1
t
b)
$\tilde{Z}^*$ $\tilde{S}^*$ $\tilde{Z}^*$ $\tilde{W}^*$ $\tilde{S}^*$ $\tilde{Z}$ $\tilde{W}$ $\tilde{S}$
t-2
t-1
t
c)
$\tilde{Z}^*$ $\tilde{S}^*$ $\tilde{Z}^*$ $\tilde{S}^*$ $\tilde{Z}^*$ $\tilde{W}^*$ $\tilde{S}^*$ $\tilde{Z}$ $\tilde{W}$ $\tilde{S}$
t-3
t-2
t-1
t

Figure C.1. Elements of the feature vector (solid boxes) for unconditional simulation of circulation indices Z, W, S (dashed boxes): (a) CIRC 1; (b) CIRC 2 and (c) CIRC 3. Asterisk indicates that the circulation index was simulated in a previous time step; tilde refers to a standardized value.

shown in Figure C.1. In the first-order model (CIRC 1) the feature vector contains the three standardized circulation indices on day $t-1$ with equal weights w_i. The feature vector of the second-order model (CIRC 2) contains standardized circulation indices for day $t-1$ and for day $t-2$. The weights for $\tilde{Z}_{t-1}$, $\tilde{W}_{t-1}$, $\tilde{S}_{t-1}$, $\tilde{Z}_{t-2}$, $\tilde{W}_{t-2}$ and $\tilde{S}_{t-2}$ are 1, 2, 1, 1, 0 and 1 respectively. The weight for $\tilde{W}_{t-2}$ was set to zero because the autocorrelation structure of the W index closely resembles that of a first-order AR process. Finally, there is a third-order model (CIRC 3) for which $\tilde{Z}_{t-3}$ and $\tilde{S}_{t-3}$ are included in the feature vector both with unit weight (and the weight for $\tilde{W}_{t-1}$ was set to 3). All models made again use of the decreasing kernel in equation (3.2) with $k = 5$, and the search for nearest neighbours was restricted to days within a 91-day moving window, centered on the calendar day of interest.

C.3 Model results

C.3.1 Autocorrelation of circulation indices

With the three models 980-year simulations were performed by resampling from the historical circulation indices of the 35-year period 1961–1995. Ta-

Table C.1. Differences between the lag 1 and lag 3 autocorrelation coefficients of daily circulation indices in 980-year simulations and the historical data. Bottom lines: historical (1961–1995) estimates of $r(1)$ and $r(3)$ with their standard errors, *se*. Z, W and S denote the relative vorticity, the west component and the south component of the flow respectively. Values in **bold** refer to statistically significant differences.

	$r(1)$			$r(3)$		
Model	Z	W	S	Z	W	S
CIRC 1	−0.008	−0.008	−0.008	**−0.091**	0.008	**−0.086**
CIRC 2	−0.018	−0.008	−0.021	−0.014	0.017	−0.013
CIRC 3	**−0.033**	−0.010	**−0.036**	−0.015	0.011	−0.022
Historical	0.498	0.755	0.521	0.182	0.393	0.190
se	0.012	0.006	0.011	0.010	0.012	0.011

ble C.1 presents the differences between the lag 1 and lag 3 autocorrelation coefficients of the simulated circulation indices and those of the historical record. The historical estimates and their jackknife standard errors (Buishand and Beersma, 1993) are also given.

In the first-order model, the lag 3 autocorrelation coefficients of the Z and S indices are significantly underestimated. The other autocorrelation coefficients differ only slightly from the historical ones. For the second-order model there are no significant differences. The third-order model is no improvement compared to the second-order model because the lag 1 autocorrelation coefficients of Z and S are significantly underestimated.

C.3.2 Run lengths of circulation types

Additionally, the reproduction of the average run length of six typical circulation types was examined. Days were classified as cyclonic, strong westerly and southerly if the standardized values of respectively Z, W and S were larger than 1.0, and as anticyclonic, easterly and northerly if $\tilde{Z}$, $\tilde{W}$ and $\tilde{S}$ were smaller than −1.0. Table C.2 presents the percentage differences between the average run lengths of the simulated indices and those of the historical indices.

The first-order model somewhat underestimates the average run lengths. The underestimation of the run lengths in the second-order model is worse but significant only for northerly flows. The third-order model significantly

Table C.2. Relative differences (%) between average run lengths of six typical circulation types in the 980-year simulations and the historical data. Bottom lines: historical (1961–1995) estimates of the average run lengths (days) with their relative standard errors, *se* (%). Values in **bold** refer to statistically significant differences.

Model	Cyclonic	Strong westerly	Southerly	Anti-cyclonic	Easterly	Northerly
CIRC 1	−1.20	−1.00	−0.65	−0.39	−2.17	−1.35
CIRC 2	−2.77	−2.17	−2.62	−2.47	−2.00	**−4.53**
CIRC 3	**−5.26**	−4.25	**−4.00**	**−4.71**	−3.50	**−6.87**
Historical	1.74	2.14	1.69	1.62	2.70	1.81
se (%)	1.97	2.32	1.83	1.74	2.86	2.10

underestimates the average run lengths for the cyclonic and anticyclonic flows as well as for the southerly and northerly flows. In the historical record the average run length for easterly flows is about 25% larger than for strong westerly flows. All models are able to reproduce this asymmetry between the average duration of strong westerly and easterly flows.

C.4 Conclusions

In conclusion, the third-order model (CIRC 3) performs worse both in terms of lag 1 autocorrelation coefficients and run lengths of circulation types. The second-order model performs better with respect to reproduction of the autocorrelation coefficients and the first-order model produces the run length statistics somewhat better. For the simulations conditional on simulated circulation indices in Sections 3.3.2 and 3.3.3, the second-order model (CIRC 2) was used to generate 980 years of daily circulation indices.

Appendix D

Nearest-neighbour resampling

D.1 Introduction

In the nearest-neighbour method the variables of interest are sampled simultaneously with replacement from the historical data. To incorporate temporal correlation, resampling is restricted to the historical values that have similar characteristics as those of the last simulated decade. One of these nearest neighbours or analogs is randomly selected and its historical successor is the next simulated decade.

A feature vector (or state vector) $\mathbf{D}_t$ is used to find the nearest neighbours in the historical data. $\mathbf{D}_t$ is formed from standardized (weather) variables generated for decade $t-1$ and earlier decades. The latter is necessary to reproduce longer-term variability (e.g., Harrold et al., 2003a,b). The nearest neighbours of $\mathbf{D}_t$ are selected in terms of a weighted Euclidean distance. For two q-dimensional vectors $\mathbf{D}_t$ and $\mathbf{D}_u$ this distance is defined as:

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

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. To obtain an equal contribution of all feature vector elements to the Euclidean distance, the weights w_j are inversely proportional to the variance of those elements. The weights are calculated separately for each of the 36 calendar decades to account for the seasonal variation in the variance. A decreasing kernel (Lall and Sharma, 1996) is used to select one of the k nearest neighbours:

$$p_j = \frac{1/j}{\sum\limits_{i=1}^{k} 1/i}, \quad j = 1, ..., k \tag{D.2}$$

with p_j the probability that the jth closest neighbour is resampled, and $k = 5$ (Buishand and Brandsma, 2001). To impose a realistic seasonal cycle upon the simulated data the search for nearest neighbours was restricted to a 7 decade wide 'moving window', centered on the calendar decade to be simulated. This window prevents that 'summer decades' are simulated during winter and 'winter decades' during summer (Buishand and Brandsma, 2001). For the historical record of 95 years, the nearest neighbours are thus selected from $7 \times 95 = 665$ decades.

A resampling technique cannot produce smaller or larger decade values than those found in the historical record. However, for periods longer than a decade, the precipitation or discharge deficit can be larger than the largest historical deficit because of rearranging extreme decade values from different parts of the historical record. In fact, this is the property that can make resampling methods useful. In a number of simulation studies of daily precipitation the generated extreme multi-day precipitation amounts were well beyond the extreme historical amounts and followed a Gumbel distribution, even outside the range of the historical data (Brandsma and Buishand, 1998; Wójcik and Buishand, 2003).

D.2 Model identification

Since the objective of the resampling model is to simulate values of precipitation P, evaporation E and discharge Q simultaneously, certain characteristics of these variables should be included in the feature vector. Several simulations were performed with different feature vectors. Best results, regarding the upper tails of the distributions of both the precipitation deficit and the discharge deficit (Figures 4.3 and 4.4) are obtained when the feature vector contains the following three elements: (i) the standardized discharge (Q) of the latest simulated decade, (ii) the average standardized discharge during the previous 18 decades, and (iii) the average of the standardized difference between precipitation and evaporation ($E - P$) in the 13 decades prior to the decade of interest. The discharge was standardized by dividing by the mean discharge for the calendar decade of interest. The variable $E - P$ was standardized by subtracting the mean and dividing by the sample standard deviation for the calendar decade of interest.

Besides this model, which is used in Chapter 4, also models with Q, E and P as individual feature vector elements, and models with different memory lengths (ranging between 2 and 12 months) were investigated but all of them gave poorer results.

D.3 Model results

To give an impression of the model performance, Table D.1 compares the average values and the standard deviations of evaporation minus precipitation ($E - P$) and of the Rhine discharge (Q) for the summer half-year in the simulated series with those in the historical series.

The averages and the standard deviations at various time scales in the simulated data are generally within one standard error from those in the historical data, pointing to a good correspondence between the simulated and historical data.

Figure D.1 presents, also for the summer half-year, the autocorrelation functions of $E - P$ and Q for the historical and simulated data. As expected, the autocorrelation is much larger for the Rhine discharge (Q) than for $E - P$. For all lags the autocorrelation of Q is very well reproduced by the resampling model. For $E - P$ the lag 1 autocorrelation is somewhat underestimated while the lag 2 autocorrelation is slightly overestimated. Overall, the autocorrelation functions are well reproduced by the resampling model.

Table D.1. Averages and standard deviations of evaporation minus precipitation ($E - P$) and of the Rhine discharge (Q) in the historical records and the simulated series for the summer half-year (April–September). $E - P$ is in mm decade^{-1}, and Q is in m^3 s^{-1}. For the historical data the standard errors are given between parentheses. The standard errors of the standard deviations were calculated following Buishand and Beersma (1996) and Beersma and Buishand (1999b; Chapter 2 of this thesis).

	Historical		Simulated	
	$E - P$	Q	$E - P$	Q
Average	4.1 (0.5)	2114 (50)	3.7	2117
σ_{decade}	18.7 (0.4)	731 (33)	18.6	720
σ_{month}	11.7 (0.4)	657 (31)	11.2	652
σ_{summer}	5.2 (0.5)	488 (33)	5.2	472

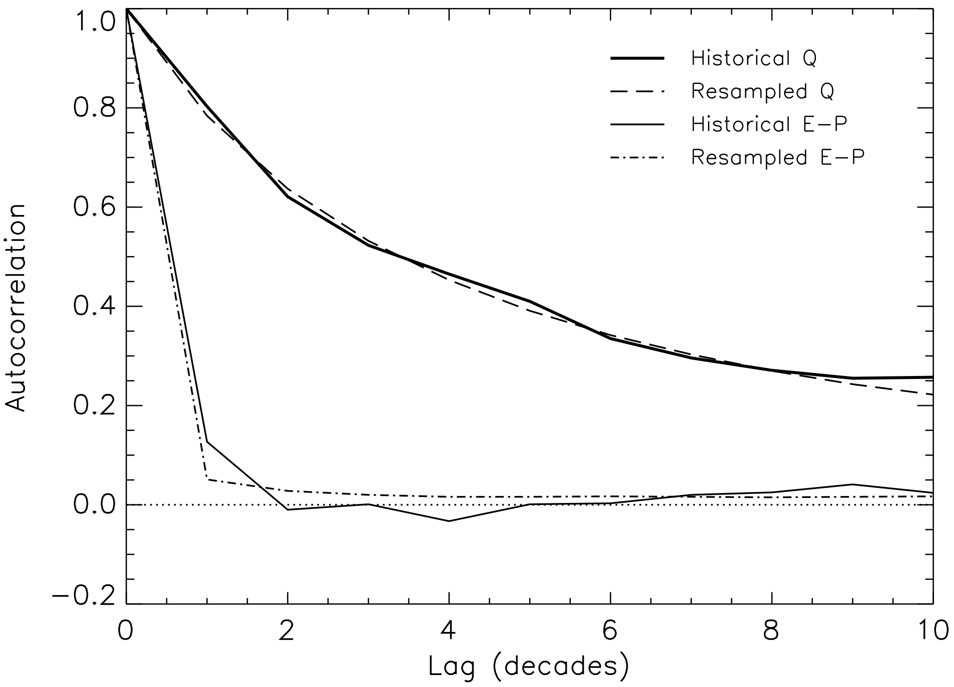

Figure D.1. Autocorrelation functions of $E - P$ and Q for the historical and simulated data.

Appendix E

Maximum likelihood estimation of the dependence parameters ρ and α

The estimation of the dependence parameters for the bivariate distributions, in Chapter 4, was carried out after the estimation of the marginal distributions. The likelihood is then based on the joint distribution of the standardized data, i.e., equation (4.2) for the bivariate normal distribution and equation (4.5) for the bivariate logistic Gumbel distribution. The standardized data are denoted as $(x_1,\ y_1),\ (x_2,\ y_2), \ldots,\ (x_N,\ y_N)$.

Censoring was necessary to estimate the parameters of the marginal distributions for the discharge deficit (Section 4.4). Let $y_{n+1} < t,\ y_{n+2} < t, \ldots,\ y_N < t$ correspond to the censored data. The likelihood for the parameter α in the logistic dependence model is then given by (Smith, 1994; Ledford and Tawn, 1996; Coles, 2001):

$$L(\alpha) = \prod_{i=1}^{n} f(x_i, y_i) \prod_{i=n+1}^{N} \left.\frac{\partial F}{\partial x}\right|_{(x_i,t)}, \tag{E.1}$$

where $f(x,y) = \frac{\partial^2 F}{\partial x \partial y}$ is the joint density:

$$\begin{aligned} f(x,y) \quad = \quad & e^{-(x+y)/\alpha} \left(e^{-x/\alpha} + e^{-y/\alpha}\right)^{\alpha-2} \left[\left(e^{-x/\alpha} + e^{-y/\alpha}\right)^{\alpha} + 1/\alpha - 1\right] \\ & \times \ \exp\left[-\left(e^{-x/\alpha} + e^{-y/\alpha}\right)^{\alpha}\right] . \end{aligned} \tag{E.2}$$

For the bivariate normal distribution the likelihood (E.1) can be written:

$$L(\rho) = \prod_{i=1}^{n} \phi_2(x_i, y_i) \prod_{i=n+1}^{N} \Pr(Y \leq t \mid X = x_i)\, \phi(x_i)\,, \tag{E.3}$$

where $\phi(x)$ is the standard normal density. Because the conditional distribution of Y in equation (E.3) is normal with mean ρx_i and variance $1 - \rho^2$, the likelihood becomes:

$$L(\rho) = \prod_{i=1}^{n} \phi_2(x_i, y_i) \prod_{i=n+1}^{N} \Phi\left(\frac{t - \rho x_i}{\sqrt{1 - \rho^2}}\right) \phi(x_i)\,, \tag{E.4}$$

with $\Phi(x) = \int_{-\infty}^{x} \phi(x)\mathrm{d}x$ the distribution function of a standard normal variable.

Appendix F

Significance of differences between districts

The differences in mean, standard deviation, CV and skewness between the six districts were tested using the standard deviations in the bottom row of Table 5.1. The MC experiment in Section 5.3.2 was used to determine the statistical significance of these values. In this experiment 10 000 samples from the fitted spline model were generated. Each MC sample consists of six 95-year sequences representing the standardized precipitation deficits of the six districts. These standardized 95-year sequences were for each district rescaled with their estimates of the location and scale parameter of the Gumbel distribution, respectively μ_i and σ_i, $i = 1, \ldots, 6$ (see also the second paragraph of Section 5.3.2). Since under the spline model all six districts have the same skewness, the null hypothesis for a test for differences in skewness is fulfilled automatically. A test for differences in the mean, standard deviation or CV requires equal means, standard deviations or CVs under the null hypothesis, which was achieved by adjusting the μ_i and σ_i to have a common (i.e. the district-average) mean, standard deviation or CV. For each of the generated 10 000 MC samples and each of the statistics in Table 5.1 the standard deviation between districts was calculated. The significance level was then approximated as the proportion of MC samples in which this statistic exceeded the corresponding value in the bottom row of Table 5.1.

For the skewness the standard deviation between districts in Table 5.1 is 0.173. This value is exceeded in 12% of the 10 000 samples and is thus not significant at the 5% level. For the mean, standard deviation and CV the standard deviations between districts in Table 5.1 are respectively 28.1, 4.3 and 0.059. These values correspond respectively with significance levels of < 0.1, 1.4 and < 0.1%.

Appendix G

AR(1) simulation model

Prior to fitting a first-order autoregressive, AR(1), time series model to the decade of days time series of the country-average precipitation deficits, these deficits were standardized by subtracting the long-term decade-average and subsequently dividing by the decade standard deviation. The AR(1) model for the standardized values is given by:

$$y_t = r(1)y_{t-1} + a_t\,, \tag{G.1}$$

with y_t the standardized precipitation deficit for the t-th decade of days during the summer half-year, $r(1)$ the lag 1 autocorrelation coefficient of the standardized precipitation deficits and a_t uncorrelated random noise with $\mathrm{E}(a_t) = 0$. The a_t's are known as random shocks or innovations. Since the y_t's are standardized, $\mathrm{E}(y_t) = 0$ and $\mathrm{var}(y_t) = 1$.

Estimates of the innovations a_t for the summer half-years of the period 1906–2000 were obtained by substituting the observed lag 1 autocorrelation coefficient of 0.128 in equation (G.1). Figure G.1 presents a normal probability plot of these estimates. The figure shows that the distribution of the innovations from the observed data deviates from the normal distribution. With a simple transformation of the normal distribution a much better correspondence with the distribution of the observed innovations is achieved:

$$a_t = \begin{cases} 0.095 + e_t - 0.08e_t^2 & \text{if } e_t \leq 0 \\ 0.095 + e_t - 0.11e_t^2 & \text{if } e_t > 0 \end{cases} \tag{G.2}$$

with e_t a standard normal variable. The transformed normal distribution based on equation (G.2) is presented in Figure G.1 as well. The AR(1) model for simulating time series of precipitation deficits used in Section 5.4.3 is obtained by combining equations (G.1) and (G.2). The standardized precipitation deficits simulated with this model are finally rescaled to their original

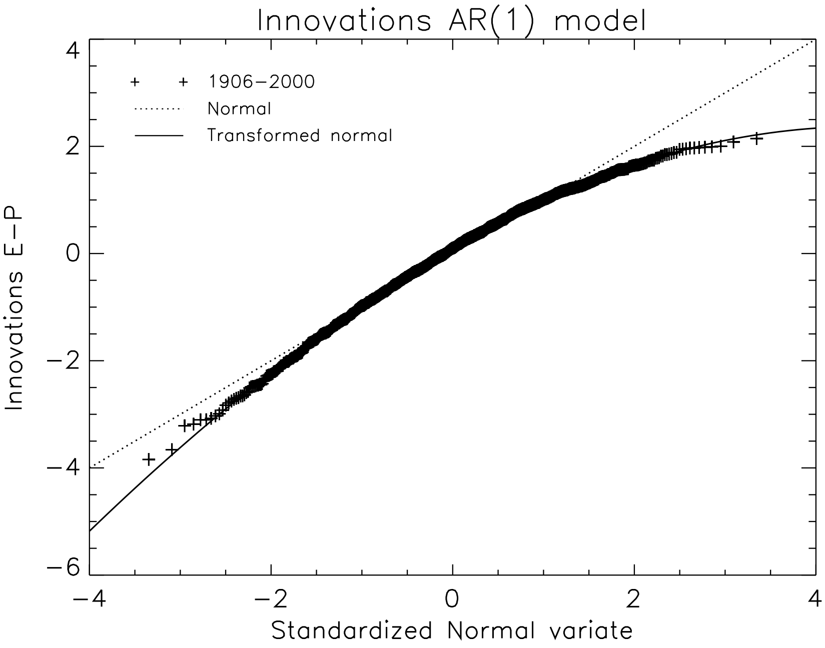

Figure G.1. Normal probability plot of the AR(1) innovations of the standardized country-average precipitation deficit during the summer half-year together with the standard normal distribution (dotted curve) and the transformed normal distribution defined by equation (G.2) (solid curve).

level by inverting the standardization procedure (using the same decade averages and decade standard deviations that were used in the standardization procedure).

Bibliography

Ahmad, M.I., C.D. Sinclair, and B.D. Spurr, 1988. Assessment of flood frequency models using empirical distribution function statistics. *Water Resour. Res.*, **24,** 1323–1328.

Al-Awadhi, S. and I. Jollife, 1998. Time series modelling of surface pressure data. *Int. J. Climatol.*, **18,** 443–455.

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

Beersma, J.J. and T.A. Buishand, 1999a. 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, 34 pp.

Beersma, J.J. and T.A. Buishand, 1999b. A simple test for equality of monthly variances in climate time series. *J. Climate*, **12,** 1770–1779.

Beersma, J.J. and T.A. Buishand, 2003. Multi-site simulation of daily precipitation and temperature conditional on the atmospheric criculation. *Clim. Res.*, **25,** 121–133.

Beersma, J.J. and T.A. Buishand, 2004. Joint probability of precipitation and discharge deficits in the Netherlands. *Water Resour. Res.*, **40,** W12508, doi:10.1029/2004WR003265.

Beersma, J.J. and T.A. Buishand, 2006. Drought in the Netherlands - Regional frequency analysis versus time series simulation. Submitted to *J. Hydrol.*

Beersma, J.J., T.A. Buishand, and H. Buiteveld, 2004. Droog, droger, droogst; KNMI/RIZA bijdrage aan de tweede fase van de Droogtestudie Nederland. KNMI-publicatie 199-II, KNMI, De Bilt, 52 pp.

Bellone, E., J.P. Hughes, and P. Guttorp, 2000. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate Res.*, **15,** 1–12.

Benestad, R.E., 2002. Empirically downscaled temperature scenarios for northern Europe based on a multi-model ensemble. *Climate Res.*, **21,** 105–125.

Bhaskaran, B., J.F.B. Mitchell, J.R. Lavery, and M. Lal, 1995. Climatic response of the Indian subcontinent to doubled CO_2 concentrations. *Int. J. Climatol.*, **15,** 873–892.

Boos, D.D. and C. Brownie, 1989. Bootstrap methods for testing homogeneity of variances. *Technometrics*, **31,** 69–82.

Boos, D.D., P. Janssen, and N. Veraverbeke, 1989. Resampling from centered data in the two sample problem. *J. Statist. Planning and Inference*, **21,** 327–345.

Bortot, P., S.G. Coles, and J.A. Tawn, 2000. The multivariate Gaussian tail model: an application to oceanographic data. *Appl. Statist.*, **49,** 31–49.

Brandsma, T. and T.A. Buishand, 1998. Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling. *Hydrol. Earth Syst. Sci.*, **2,** 195–209 (Corrigendum, *Hydrol. Earth Syst. Sci.*, **3**, 319, 1999).

Brillinger, D.R., 1964. The asymptotic behaviour of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. *Rev. Inst. Int. Statist.*, **32,** 202–206.

Brown, M.B. and A.B. Forsythe, 1974. Robust tests for the equality of variances. *J. Amer. Statist. Assoc.*, **69,** 364–367.

Buishand, T.A., 1978. The binary DARMA(1,1) process as a model for wet-dry sequences. Technical Note 78-01, Agricultural University Wageningen, The Netherlands, 49 pp.

Buishand, T.A., 1984. Bivariate extreme-value data and the station-year method. *J. Hydrol.*, **69,** 77–95.

Buishand, T.A., 2006. Estimation of a large quantile of the distribution of multi-day seasonal maximum rainfall: Can stochastic simulation be of use? Submitted to *Climate Res.*

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

Buishand, T.A. and J.J. Beersma, 1996. Statistical tests for comparison of daily variability in observed and simulated climates. *J. Climate*, **9,** 2538–2550. (Corrigendum, *J. Climate*, **10**, 818, 1997).

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

Cao, H.X., J.F.B. Mitchell, and J.R. Lavery, 1992. Simulated diurnal range and variability of surface temperature in a global climate model for present and doubled CO_2 climates. *J. Climate*, **5,** 920–943.

Charles, S.P., B.C. Bates, and J.P. Hughes, 1999. A spatiotemporal model for downscaling precipitation occurence and amounts. *J. Geophys. Res.*, **104,** 31 567–31 669.

CHO-TNO, 1988. Van Penman naar Makkink; een nieuwe berekeningswijze voor de klimatologische verdampingsgetallen. Rapporten en nota's No. 19, Commissie voor Hydrologisch Onderzoek TNO, 's-Gravenhage, The Netherlands, 67 pp.

Coles, S.G., 2001. *An introduction to statistical modeling of extreme values.* Springer, London, 208 pp.

Coles, S.G., J.E. Heffernan, and J.A. Tawn, 1999. Dependence measures for multivariate extremes. *Extremes*, **2,** 339–365.

Coles, S.G. and J.A. Tawn, 1994. Statistical methods for multivariate extremes: an application to structural design (with discussion). *Appl. Statist.*, **43,** 1–48.

Conway, D., R.L. Wilby, and P.D. Jones, 1996. Precipitation and air flow indices over the British Isles. *Climate Res.*, **7,** 169–183.

Corte-Real, J., H. Xu, and B. Qian, 1999. A weather generator for obtaining daily precipitation scenarios based on circulation patterns. *Climate Res.*, **13,** 61–75.

Cramér, H., 1946. *Mathematical Methods of Statistics.* Princeton University Press, 575 pp.

Dalén, J., 1987. Algebraic bounds on standardized sample moments. *Statist. Prob. Lett.*, **5,** 329–331.

Davison, A.C. and D.V. Hinkley, 1997. *Bootstrap methods and their applications.* Cambridge University Press, 582 pp.

de Bruin, H.A.R. and J.N.M. Stricker, 2000. Evaporation of grass under non-restricted soil moisture conditions. *Hydrol. Sci. J.*, **45,** 391–406.

de Bruin, H.A.R., B.J.J.M. van den Hurk, and D. Welgraven, 1995. A series of global radiation at Wageningen for 1928-1992. *Int. J. Climatol.*, **15,** 1253–1272.

Diaconis, P. and B. Efron, 1983. Computer-intensive methods in statistics. *Scientific American*, **248,** 96–108.

Downton, M.W. and R.W. Katz, 1993. A test for inhomogeneous variance in time-averaged temperature data. *J. Climate*, **6,** 2448–2464.

Efron, B. and R.J. Tibshirani, 1993. *An introduction to the Bootstrap.* Chapman & Hall, New York, 436 pp.

Favre, A.-C., S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée, 2004. Multivariate hydrological frequency analysis using copulas. *Water Resour. Res.*, **40,** W01101, doi:10.1029/2003WR002456.

Feller, W., 1968. *An introduction to probability theory and its applications.* John Wiley & Sons, New York, 509 pp.

Fink, A., U. Ulbrich, and H. Engel, 1996. Aspects of the January 1995 flood in Germany. *Weather*, **51,** 34–39.

Fiorentino, M., S. Gabriele, F. Rossi, and P. Versace, 1987. Regional flood frequency analysis using the two-component extreme value distribution. a key reference abstract. *Excerpta*, **2,** 39–50.

Fowler, H.J., C.G. Kilsby, and P.E. O'Connell, 2000. A stochastic rainfall model for the assessment of regional water resource systems under changed climatic conditions. *Hydrol. Earth Sys. Sci.*, **4,** 263–282.

Frantzen, A.J. and W.R. Raaff, 1982. De relatie tussen de globale straling en de relatieve zonneschijnduur in Nederland. Wetenschappelijk rapport W.R. 82-5, KNMI, De Bilt, 45 pp.

Giorgi, F., B. Hewitson, J. Christensen, M. Hulme, H. von Storch, P. Whetton, R. Jones, L. Mearns, and C. Fu, 2001. Regional Climate Information - Evaluation and Projections. In: *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change,* Houghton, J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell, and C.A. Johnson (eds.). Cambridge University Press, 881 pp.

Goodess, C.M. and J.P. Palutikof, 1998. Development of daily rainfall scenarios for southeast Spain using a circulation-type approach to downscaling. *Int. J. Climatol.*, **18,** 1051–1083.

Gordon, H.B. and B.G. Hunt, 1994. Climate variability within an equilibrium greenhouse simulation. *Climate Dyn.*, **9,** 195–212.

Gregory, J.M. and J.F.B. Mitchell, 1995. Simulation of daily variability of surface temperature and precipitation over Europe in the current $2\times CO_2$ climates using the UKMO climate model. *Q.J.R. Meteorol. Soc.*, **121,** 1451–1476.

Hall, P. and N. Tajvidi, 2000. Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli*, **6,** 835–844.

Harrold, T.I., A. Sharma, and S.J. Sheather, 2003a. A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resour. Res.*, **39,** 1343, doi:10.1029/2003WR002570.

Harrold, T.I., A. Sharma, and S.J. Sheather, 2003b. A nonparametric model for stochastic generation of daily rainfall occurence. *Water Resour. Res.*, **39,** 1300, doi:10.1029/2003WR002182.

Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, **3,** 1163–1174.

Hosking, J.R.M and J.R. Wallis, 1988. The effect of intersite dependence on regional flood frequency analysis. *Water Resour. Res.*, **24,** 588–600.

Hosking, J.R.M and J.R. Wallis, 1997. *Regional frequency analysis; An approach based on L-moments.* Cambridge University Press, 224 pp.

Hughes, J.P., D.P. Lettenmaier, and P. Guttorp, 1993. A stochastic approach for assessing the effect of changes in synoptic circulation patterns on gauge precipitation. *Water Resour. Res.*, **29,** 3303–3315.

Hundecha, Y. and A. Bárdossy, 2005. Trends in daily precipitation and temperature extremes across western Germany in the second half of the 20th century. *Int. J. Climatol.*, **25,** 1189–1202.

Huth, R., J. Kyselý, and M. Dubrovský, 2001. Time structure of observed, GCM-simulated, downscaled, and stochastically generated daily temperature series. *J. Climate*, **14,** 4047–4061.

IPCC, 1990. *Climate Change: The IPCC Scientific Assessment,* J.T. Houghton, G.J. Jenkins and J.J. Ephraums (eds.). Cambridge University Press, 365 pp.

IPCC, 1996. *Climate Change 1995: The Science of Climate Change,* J.T. Houghton, L.G. Meira Filho, B.A. Callander, N. Harris, A. Kattenberg and K. Maskell (eds.). Cambridge University Press, 572 pp.

IPCC, 2001. *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change,* Houghton, J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell, and C.A. Johnson (eds.). Cambridge University Press, 881 pp.

Jenkinson, A.F. and F.P. Collinson, 1977. An initial climatology of gales over the North Sea. Synoptic Climatology Branch Memorandum 62, Meteorological Office, Bracknell, 18 pp.

Johnson, R., 2001. An introduction to the bootstrap. *Teaching Statistics*, **23,** 49–54.

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

Kelly, K.S. and R. Krzysztofowicz, 1997. A bivariate meta-Gaussian density for use in hydrology. *Stoch. Hydrol. Hydraul.*, **11,** 17–31.

Kendall, M. and A. Stuart, 1973. *The Advanced Theory of Statistics*, volume 2: *Inference and Relationship* (3rd ed.). Charles Griffin, London, 723 pp.

Keselman, H.J., P.A. Games, and J.J. Clinch, 1979. Tests for homogeneity of variance. *Commun. Stat.-Simul. Comput.*, **B8,** 113–129.

Klein Tank, A.M.G. and R. Sluijter, 2003. *Nederland is verder opgewarmd.* In: *De toestand van het klimaat in Nederland 2003*, K. Verbeek (ed.). pp. 6–14, KNMI, De Bilt.

Koch, G.G., 1983. *Intraclass correlation coefficient.* In: *Encyclopedia of Statistical Sciences*, vol. 4, S. Kotz, N.L. Johnson, and C.B. Read (eds.). pp. 212–217, John Wiley & Sons, New York.

Kotz, S. and S. Nadarajah, 2000. *Extreme value distributions: Theory and applications.* Imperial College Press, London, 185 pp.

Kroll, C.N. and J.R. Stedinger, 1998. Regional hydrologic analysis: Ordinary and generalized least squares revisted. *Water Resour. Res.*, **34,** 121–128.

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

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

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

Ledford, A. and J.A. Tawn, 1996. Statistics for near independence in multivariate extreme values. *Biometrika*, **83,** 169–187.

Leese, M.N., 1973. Use of censored data in the estimation of Gumbel distribution parameters for annual maximum flood series. *Water Resour. Res.*, **9,** 1534–1542.

Leytham, K.M., 1987. A joint rank test for assessing multivariate normality in hydrologic data. *Water Resour. Res.*, **23,** 2311–2317.

Liang, X.-Z., W.-C. Wang, and M.P. Dudeck, 1995. Interannual variability of regional climate and its change due to the greenhouse effect. *Global and Planetary Change*, **10,** 217–238.

Linderson, M.-L., 2001. Objective classification of atmospheric circulation over southern Scandinavia. *Int. J. Climatol.*, **21,** 155–169.

Mearns, L.O., S.H. Schneider, S.L. Thompson, and L.R. McDaniel, 1990. Analysis of climate variability in general circulation models: Comparison with observations and changes in variability in 2×CO_2 experiments. *J. Geophys. Res.*, **95,** 20 469–20 490.

Meehl, G.A. and W.M. Washington, 1993. South Asian summer monsoon variability in a model with doubled atmospheric carbon dioxide concentration. *Science*, **260,** 1101–1104.

Meehl, G.A., M. Wheeler, and W.M. Washington, 1994. Low-frequency variability and CO_2 transient climate change. Part 3. Intermonthly and interannual variability. *Climate Dyn.*, **10,** 277–303.

Middelkoop, H. and C.O.G. van Haselen, 1999. Twice a river. Rhine and Meuse in the Netherlands. RIZA report No. 99.003, RIZA, Arnhem, The Netherlands, 127 pp.

Miller, R.G., 1968. Jackknifing variances. *Ann. Math. Statist.*, **39,** 567–582.

Murphy, J.M., 1995. Transient response of the Hadley Centre Coupled Ocean-Atmosphere Model to increasing carbon dioxide. Part I: Control climate and flux adjustment. *J. Climate*, **8,** 36–56.

Murphy, J.M. and J.F.B. Mitchell, 1995. Transient response of the Hadley Centre Coupled Ocean-Atmosphere Model to increasing carbon dioxide. Part II: Spatial and temporal structure of response. *J. Climate*, **8,** 57–80.

O'Brien, R.G., 1978. Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika*, **43,** 327–342.

Palutikof, J.P., C.M. Goodess, S.J. Watkins, and T. Holt, 2002. Generating rainfall and temperature scenarios at multiple sites: examples from the Mediterranean. *J. Climate*, **15,** 3529–3548.

Pearson, E.S., 1935. A comparison of β_2 and Mr. Geary's W_n criteria. *Biometrika*, **27,** 333–352.

Pickands, J., 1981. Multivariate extreme value distributions. *Bull. Int. Statist. Inst.*, **49,** 859–878.

Preisendorfer, R.W. and T.P. Barnett, 1983. Numerical model-reality intercomparison tests using small-sample statistics. *J. Atmos. Sci.*, **40,** 1884–1896.

Qian, B., J. Corte-Real, and H. Xu, 2002. Multisite stochastic weather models for impact studies. *Int. J. Climatol.*, **22,** 1377–1397.

Quenouille, M.B., 1956. Notes on bias in estimation. *Biometrika*, **43,** 353–360.

Raisanen, J., 1995. A comparison of the results of seven GCM experiments in northern Europe. *Geophysica*, **30,** 3–30.

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

Rapp, J. and C.-D. Schönwiese, 1995. *Atlas der Niederslags- und Temperaturtrends in Deutschland 1891-1990 und 1961-1990.* Frankfurther Geowissenschaftliche Arbeiten, Serie B, Band 5. J.W. Goethe Universität, Frankfurt am Main, Deutschland.

Reed, D.N., 1986. Simulation of time series of temperature and precipitation over eastern England by an atmospheric general circulation model. *J. Climatol.*, **6,** 233–253.

Reed, D.W., D.S. Faulkner, and E.J. Stewart, 1999. The FORGEX method of rainfall growth estimation II: Description. *Hydrology and Earth System Sciences*, **3,** 197–203.

Rind, D., R. Goldberg, and R. Ruedy, 1989. Change in climate variability in the 21st century. *Climate Change*, **14,** 5–37.

Rossi, F., M. Fiorentino, and P. Versace, 1984. Two-component extreme value distribution for flood frequency analysis. *Water Resour. Res.*, **20,** 847–856.

Santer, B.D. and T.M.L. Wigley, 1990. Regional validation of means, variances, and spatial patterns in general circulation model control runs. *J. Geophys. Res.*, **95,** 829–850.

Schmidli, J., C. Schmutz, C. Frei, H. Wanner, and C. Schär, 2002. Mesoscale precipitation variability in the region of the European Alps during the 20th century. *Int. J. Climatol.*, **22,** 1049–1074.

Schubert, S., 1994. A weather generator based on the European 'Grosswetterlagen'. *Climate Res.*, **4,** 191–202.

Shiau, J.T., 2003. Return period of bivariate distributed extreme hydrological events. *Stochast. Envir. Res. and Risk Assess.*, **17,** 42–57.

Shumway, R.H., A.S. Azari, and P. Johnson, 1989. Estimating mean concentrations under transformation for environmental data with detection limits. *Technometrics*, **31,** 347–356.

Sibuya, M., 1960. Bivariate extreme statistics. *Ann. Inst. Statist. Math.*, **11,** 195–210.

Smith, R.L., 1994. *Multivariate Threshold Methods.* In: *Extreme Value Theory & Application*, J. Galambos, J. Lechner, and E. Simiu (eds.). pp. 225–248, Kluwer, Dordrecht.

Stedinger, J.R., R.M. Vogel, and E. Foufoula-Georgiou, 1993. *Frequency analysis of extreme events.* In: *Handbook of Hydrology*, D. Maidment (ed.). pp. 18.1–18.66, McGraw-Hill, New York.

Stehlík, J. and A. Bárdossy, 2002. Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation. *J. Hydrol.*, **256,** 120–141.

Stephens, M.A., 1986a. *Tests based on EDF statistics.* In: *Goodness-of-fit Techniques*, R.B. D'Agostino and M.A. Stephens (eds.). pp. 97–193, Marcel Dekker, New York.

Stephens, M.A., 1986b. *Tests based on regression and correlation.* In: *Goodness-of-fit Techniques*, R.B. D'Agostino and M.A. Stephens (eds.). pp. 195–233, Marcel Dekker, New York.

Stephenson, A., 2003. Simulating multivariate extreme value distributions of logistic type. *Extremes*, **6,** 49–59.

Tawn, J.A., 1988. Bivariate extreme value theory: Models and estimation. *Biometrika*, **75,** 397–415.

Tiago de Oliveira, J., 1980. *Bivariate extremes: Foundations and statistics.* In: *Multivariate Analysis*, vol. 5, P.R. Krishnaiah (ed.). pp. 349–366, North-Holland Publishing, Amsterdam.

van Ulden, A.P., G. Lenderink, B.J.J.M. van den Hurk, and E. van Meijgaard, 2007. Circulation statistics and climate change in Central Europe: PRUDENCE simulations and observations. *Climatic Change.* (in press).

van Ulden, A.P. and G.J. van Oldenborgh, 2006. Large-scale atmospheric circulation biases and changes in global climate model simulations and their importance for climate change in Central Europe. *Atmos. Chem. Phys.*, **6,** 863–881.

Vogel, R.M., 1986. The probability plot correlation test for the normal, lognormal, and Gumbel distributional hypotheses. *Water Resour. Res.*, **22,** 587–590.

von Storch, H., 1982. A remark on Chervin-Schneider's algorithm to test significance of climate experiments with GCM's. *J. Atmos. Sci.*, **39,** 187–189.

Wallis, J.R., N.C. Matalas, and J.R. Slack, 1974. Just a moment! *Water Resour. Res.*, **10,** 211–219.

Walsh, J.E., 1947. Concerning the effect of intraclass correlation on certain significance tests. *Ann. Math. Statist.*, **18,** 88–96.

Welch, B.L., 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika*, **29,** 350–361.

Widmann, M. and C. Schär, 1997. A principal component and long-trend analysis of daily precipitation in Switzerland. *Int. J. Climatol.*, **17,** 1333–1356.

Wigley, T.M.L. and B.D. Santer, 1990. Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *J. Geophys. Res.*, **95,** 851–865.

Wilby, R.L., H. Hassan, and K. Hanaki, 1998. Statistical downscaling of hydrometeorological variables using general circulation model output. *J. Hydrol.*, **205,** 1–19.

Wilson, L.L., D.P. Lettenmaier, and E. Skyllingstad, 1992. A hierarchical stochastic model of large-scale atmospheric circulation patterns and multiple station daily precipitation. *J. Geophys. Res.*, **97,** 2791–2809.

Wójcik, R. and T.A. Buishand, 2003. Simulation of 6-hourly rainfall and temperature by two resampling schemes. *J. Hydrol.*, **273,** 69–80.

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

Yue, S., 2001. The Gumbel logistic model for representing a multivariate storm event. *Adv. in Water Resour.*, **24,** 179–185.

Yue, S., T.B.M.J. Ouarda, B. Bobée, P. Legendre, and P. Bruneau, 1999. The Gumbel mixed model for flood frequency analysis. *J. Hydrol.*, **226,** 88–100.

Zorita, E., J.P. Hughes, D.P. Lettenmaier, and H. von Storch, 1995. Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *J. Climate*, **8,** 1023–1042.

Zorita, E. and H. von Storch, 1999. The analog method as a simple statistical downscaling technique: Comparison with more complicated models. *J. Climate*, **12,** 2474–2489.

Zwiers, F.W., 1987. Statistical considerations for climate experiments. Part II: Multivariate tests. *J. Climate Appl. Meteor.*, **26,** 477–487.

Zwiers, F.W. and H. von Storch, 1990. Regime-dependent autoregressive time series modeling of the Southern Oscillation. *J. Climate*, **3,** 1347–1363.

Zwiers, F.W. and H.J. Thiébaux, 1987. Statistical considerations for climate experiments. Part I: Scalar tests. *J. Climate Appl. Meteor.*, **26,** 464–476.

Publications by the author related to this thesis

In this chapter an overview is given of publications by the author that are related to or relevant for this thesis. The references are followed by a short (Dutch) summary or the original abstract and are presented in chronological order.

Jules J. Beersma, 1992. GCM control run of UK Meterological Office compared with the real climate in the NW European winter. KNMI publication: WR-92-02, KNMI, De Bilt, 32 pp.
(Related to Chapter 2, Beersma and Buishand, 1999b).

Abstract A method is presented to compare the statistical properties of surface air temperature, sea surface temperature, precipitation, global radiation, 500 mbar height, sea level pressure, and wind from a GCM with those of observations. As an illustration the control run of an 11-layer GCM from the UK Met Office [Hadley Centre] for five successive winters (DJF) is compared with observations for NW Europe. In this comparison large differences are found in (monthly) mean values, standard deviations and autocorrelation coefficients for various elements. In general, this GCM winter run creates too low temperatures over land, too high temperatures over sea, large temperature variability but small pressure variability and too small autocorrelation coefficients for time-lags of more than three days. Too many wet days are created with too little precipitation. The 500 mbar circulation is veered with respect to reality and shows a peculiar through over the North Sea. The wind at 100 m is backed and the geostrophic wind is somewhat underestimated. No realistic transition of the surface temperature from land to sea is observed. These results suggest that care has to be taken in interpreting direct model output on a regional scale such as NW Europe. Given the limitations of this study, the relatively short GCM run and difficulties related to comparing GCM grid points with station observations, further work along these lines is desirable, preferably on the basis of the output of more recent GCM versions.

T. Adri Buishand and Jules J. Beersma, 1993. Jackknife tests for differences in autocorrelation between climate time series. J. Climate, 6, 2490–2495.
(Related to Chapter 2, Beersma and Buishand, 1999b).

Abstract Two tests for differences in the lag 1 autocorrelation coefficient based on jackknife estimates are proposed. These tests are developed for the pooled sample of all daily values in a certain calendar month (e.g., all January data). Jackknife estimates of the autocorrelation coefficients and their standard errors from such a sample are obtained by omitting each year once and recomputing the autocorrelation estimates. Monte Carlo results for several distributions show that the critical values of the test statistics can be based on the Student's t-distribution. Regional analogs of these test statistics are derived from the jackknife estimate of the mean lag 1 autocorrelation coefficient for the sites of interest. In a similar way one can get a single test statistic for a season or the whole year. As an illustration it is shown that the lag 1 autocorrelation coefficients of the simulated daily temperatures of the Canadian Climate Centre second-generation general circulation model are significantly below those of the observed temperatures at De Bilt for most seasons. Over western Europe there is no statistical evidence of differences in autocorrelation between the $1\times CO_2$ and $2\times CO_2$ runs of this model.

T. Adri Buishand and Jules J. Beersma, 1996. Statistical tests for comparison of daily variability in observed and simulated climates. J. Climate, 9, 2538–2550. (Corrigendum; J. Climate, 10, 818, 1997).
(Related to Chapter 2, Beersma and Buishand, 1999b).

Abstract Tests for differences in daily variability based on the jackknife are presented. These tests properly account for the effect of autocorrelation in the data and are reasonably robust against departures from normality. Three measures for the daily variability are considered; process, within-month and innovation variance. The jackknife statistic compares the logarithm of these measures. The standard errors of this logarithm are obtained by recomputing the variance estimates for all subsamples wherein one month is omitted from the complete sample. A simple extension of the jackknife procedure is given to obtain a powerful multivariate test in situations that the differences in variance

have the same sign across the region considered or over the year.

As an illustration the tests are applied to near-surface temperatures over Europe simulated by the coupled ECHAM/LSG model [from the Max-Planck-institute for Meteorology in Hamburg, Germany]. It is shown that the control run of the model significantly overestimates the process variance in winter and spring, and the within-month variance in all seasons. Significant differences are also found for the innovation variances of the daily temperatures, but the sign of the differences varies over the year. In a perturbed run with enhanced atmospheric greenhouse gas concentrations the daily temperature variability over Europe significantly decreases in winter and spring compared with the control run.

Jules J. Beersma and T. Adri Buishand, 1999. 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, 34 pp.
(Related to Chapter 3, Beersma and Buishand, 2003).

Summary This study deals with multi-site conditional simulation of precipitation and temperature for 25 locations in the German part of the Rhine basin. Nearest-neighbour resampling is also used to generate synthetic sequences of daily circulation indices that are needed for long-duration conditional simulations. Conditional simulations are performed to reconstruct precipitation statistics for the period 1891–1995. These simulations explain on average slightly more than 50% of the trends in mean winter precipitation at five stations for which monthly data during this century were available. The sensitivity of simulated precipitation to changes in circulation indices is studied by performing three simulations conditional on the 1961–1995 circulation indices, in which in each simulation only one of the three circulation indices is systematically changed. These simulations show that the simulated precipitation is most sensitive to changes in the westerly flow index W, followed by changes in the vorticity index Z. The mean precipitation is typically much more sensitive to systematic changes in W and Z than the precipitation extremes. This is because a large part in the change in the mean precipitation is due to a change in the number of wet days, which has less influence on the extremes.

Jules J. Beersma, T. Adri Buishand and Rafał Wójcik, 2001. Rainfall generator for the Rhine basin; multi-site simulation of daily weather variables by

nearest-neighbour resampling. In: Generation of hydrometeorological reference conditions for the assessment of flood hazard in large river basins, P. Krahe and D. Herpertz (Eds.). CHR-Report no. I-20, International Commission for the Hydrology of the Rhine basin (CHR), Lelystad, The Netherlands, pp. 69–77.
(Related to Chapter 3, Beersma and Buishand, 2003).

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).

Jules J. Beersma en T. Adri Buishand, 2002. Droog, droger, droogst - Bijdrage van het KNMI aan de eerste fase van de Droogtestudie Nederland. KNMI publicatie 199-I, ISBN 90-369-2221-6, KNMI, De Bilt, 42 pp.
(Related to Chapter 4, Beersma and Buishand, 2004).

Samenvatting In de zomer van 2000 heeft de commissie 'Waterbeheer 21^{e} eeuw' (WB21) advies uitgebracht over de organisatie en inrichting van het waterbeheer in de 21^{e} eeuw. De commissie gaf o.a. aan dat behalve te veel water ook te weinig water een bedreiging vormt voor de (toekomstige) waterhuishouding van Nederland. In de 'startovereenkomst Waterbeleid 21^{e} eeuw' wordt ten behoeve van de laagwaterproblematiek (extreme droge situaties) een gezamenlijke verkenning van Rijk, provincies en waterschappen naar droogte aangekondigd die later is omgedoopt tot de 'Droogtestudie Nederland'.

De tot dan toe gebruikte schattingen van de kans op extreme droogte stammen uit de PAWN studie van 1985. Dit rapport voorziet in de behoefte van de Droogtestudie om de kans op (extreme) droogte opnieuw te schatten mede op basis van de meest recente historische gegevens. Naast droogte in Nederland wordt in deze analyse ook naar droogte in termen van lage afvoeren van de

Rijn gekeken in verband met het belang van de Rijnafvoer bij het voorzien in de Nederlands waterbehoefte. Verder wordt voor het eerst ook uitvoerig gekeken naar de kans op het simultaan optreden van extreme droogte en extreme lage afvoeren van de Rijn, een combinatie die de grootste economische schade tot gevolg heeft.

De herhalingstijden van historische 'schadejaren' zijn bepaald op basis van frequentie analyses van zowel historisch waargenomen extremen als van extremen uit lange gesimuleerde reeksen op basis van tijdreeks resampling. De gesimuleerde reeksen blijken minder geschikt om de kans op simultane gebeurtenissen te schatten. Nader onderzoek zal moeten uitwijzen of de simulatie van met name het neerslagtekort door middel van tijdreeks resampling verbeterd kan worden. Het gebruik van een getransformeerde twee-dimensionale normale kansverdeling heeft als bezwaar dat de afhankelijkheid tussen extreme neerslag- en afvoertekorten wordt onderschat, waardoor ook de kans op simultane extreme gebeurtenissen wordt onderschat. Op basis van een van de (geschatte) economische schade afgeleid faalgebied en verschillende twee-dimensionale normale verdelingen zijn de herhalingstijden voor de simultane neerslag- en afvoertekorten van de schadejaren bepaald. Historisch gezien zijn 1976 en 1921 de droogste jaren gevolgd door 1959 en 1947. Voor 1976 is een herhalingstijd van bijna 200 jaar gevonden.

Indien de aanvoer van de Rijn geen rol van betekenis speelt zijn de herhalingstijden gebaseerd op een getransformeerde normale verdeling van het neerslagtekort die op het oog de 'beste' fit voor de grootste neerslagtekorten geeft. Het droogste jaar is dan wederom 1976 (met een herhalingstijd van ongeveer 75 jaar) gevolgd door 1959, 1911 en 1921.

Jules J. Beersma, T. Adri Buishand en Hendrik Buiteveld, 2004. Droog, droger, droogst - KNMI/RIZA bijdrage aan de tweede fase van de Droogtestudie Nederland. KNMI publicatie 199-II, ISBN 90-369-2260-7, KNMI, De Bilt, 52 pp.
(Related to Chapters 4 and 5, Beersma and Buishand, 2004, 2006).

Samenvatting Dit rapport vormt het vervolg op het gelijknamige rapport uit 2002 en beschrijft de KNMI bijdrage aan de tweede fase van de Droogtestudie Nederland, inclusief een bijdrage van RIZA die nauw bij het thema van deze rapportage aansluit.

De statistiek voor het bepalen van de kans op het simultaan optreden van extreme droogte in Nederland en extreem lage afvoeren van de Rijn is verder uitgewerkt met als resultaat verbeterde schattingen van simultane overschrijdingskansen. Voor de droogte van 1976 leidt dit tot een herhalingstijd van ongeveer 110 jaar (190 jaar in het vorige rapport).

Daarnaast is een verbeterd resampling model ontwikkeld dat goed bruikbaar is voor het schatten van de overschrijdingskansen van het (maximale) neerslagtekort in Nederland. Voor de droogte van 1976 in termen van alleen het neerslagtekort in Nederland wordt nu een herhalingstijd gevonden van ongeveer 90 jaar (was 75 jaar). Datzelfde model is ook gebruikt voor het schatten van de overschrijdingskansen van het neerslagtekort in een zestal regio's binnen Nederland. De regionale differentiatie laat zien dat er binnen Nederland naast systematische verschillen ook verschillen van jaar tot jaar zijn. Zo is in het westen van het land het maximale neerslagtekort meestal groter dan in het oosten van het land. 1976 was het droogst in Zeeland en het minst droog in oost Nederland. In de regio's noordoost Nederland en oost Nederland was 1959 het droogste jaar, voor de overige regio's was dat 1976.

De droogte van 2003 kreeg veel aandacht (in de media) vanwege een aantal (bijna) problemen. Toch was de zomer van 2003 niet extreem droog. Kijken we alleen naar het (maximale) neerslagtekort van 2003 dan vinden we gemiddeld over Nederland een herhalingstijd van ongeveer 10 jaar. In noordwest Nederland was met een herhalingstijd van 13 jaar de droogte relatief het grootst en in het Maasgebied met een herhalingstijd van 4 jaar relatief het kleinst. Houden we ook rekening met de (lage) afvoer van de Rijn dan neemt gemiddeld over Nederland de herhalingstijd iets toe; van ongeveer 10 tot ongeveer 12 jaar.

In dit rapport worden ook de klimaatscenario's en toekomstscenario's gepresenteerd waarmee in de tweede fase van de Droogtestudie is gewerkt. Voor Nederland worden de standaard KNMI klimaatscenario's toegepast terwijl voor de stroomgebieden van de Rijn en de Maas klimaatscenario's op basis van de UKHI/IS92a projecties worden gebruikt. Voor het Controlist klimaatscenario (1°C temperatuurtoename in 2050) is er voldoende overeenstemming tussen beide typen klimaatscenario's. Onder dat scenario neemt de gemiddelde Rijnafvoer aan het einde van de zomer en in het begin van het najaar met ruim 5% af, terwijl de gemiddelde Maasafvoer dan nauwelijks verandert. Daarnaast neemt onder het Controlist klimaatscenario het maximale neerslagtekort voor alle schadejaren toe. Voor de meest extreme jaren (1959 en 1976) is de toename iets kleiner ($\sim$5%) dan de gemiddelde toename van 6.2%. Ondanks deze systematische toename van het maximale neerslagtekort zijn de herhalingstijden voor de getransformeerde schadejaren vrijwel onveranderd. Dit komt doordat de kansverdeling van de maximale neerslagtekorten opschuift en wel zodanig dat de rangorde van de schadejaren niet verandert.

Samenvatting

In dit proefschrift zijn resampling technieken gebruikt voor het bepalen van de statistische onzekerheid van kenmerken van extreme waarden, en om zeer lange hydro-meteorologische tijdreeksen te simuleren met extreme waarden die nog niet eerder zijn waargenomen. Een frequentieanalyse van de extremen in zulke lange tijdreeksen heeft als voordeel dat de statistische onzekerheid van het resultaat in het algemeen kleiner is dan bij een frequentieanalyse van uitsluitend de waargenomen extremen. Een resampling techniek heeft daarnaast als voordeel dat het niet nodig is om allerlei, min of meer onjuiste, aannames te maken over de statistische eigenschappen van de data.

Door tijdreeksen van geresamplede hydro-meteorologische data te combineren met geschikte hydrologische modellen, kan de tijdreeks resampling aanpak toegepast worden op een breed scala aan studies met betrekking tot de gevolgen van extreme hydro-meteorologische gebeurtenissen, zoals extreem hoge of lage rivierafvoeren, extreem hoge of lage grondwaterstanden of andere extreme hydrologische gebeurtenissen. Een bijkomend voordeel van deze aanpak is dat hydro-meteorologische effecten kunnen worden gescheiden van door de mens veroorzaakte veranderingen in het hydrologische systeem, zoals kanalisering, verstedelijking en ontbossing.

Het eerste hoofdstuk vormt de inleiding waarin de bredere context van het proefschrift geschetst wordt. Tevens wordt in dit hoofdstuk een overzicht gegeven van de historische ontwikkelingen van de resampling technieken waarvan gebruik wordt gemaakt. Met een simpel voorbeeld op basis van het gooien met een dobbelsteen wordt de kracht van resampling technieken geïllustreerd. Hoofdstuk 1 sluit af met een overzicht van de onderwerpen, en de daarbij gebruikte resampling technieken, die in de volgende hoofdstukken aan bod komen.

De eerste toepassing van resampling (Hoofdstuk 2) betreft de statistische onzekerheid van variantie schattingen. De ‘standard error’ (een maat voor de statistische onzekerheid) van de variantie wordt bepaald met behulp van de 'jackknife' en wordt gebruikt voor het opzetten van een toets die de gelijkheid van varianties van maandwaarden kan vaststellen. Zo'n toets kan nuttig zijn bij het valideren van modellen of bij het detecteren van klimaatverandering. En omdat het veel of weinig optreden van extreme gebeurtenissen afhangt van de variantie kan zo'n toets zelfs als een eerste indicatie gebruikt worden bij de detectie van verschillen of veranderingen in de frequentie van extreme gebeurtenissen. Deze toets heeft een brede praktische toepasbaarheid mede omdat hij een simpele en degelijke multivariate uitbreiding heeft. Hiermee wordt bedoeld dat de toets zonder problemen kan worden toegepast op een gebied waarvoor op verschillende locaties varianties bepaald zijn. Door de varianties op de verschillende locaties in één (multivariate) toets te combineren neemt het onderscheidend vermogen (zeg maar, de gevoeligheid) van de toets in het algemeen sterk toe.

Ter illustratie is de multivariate toets toegepast op neerslag en temperatuur varianties in een simulatie van klimaatverandering met het Engelse Hadley Centre klimaatmodel. Voor verschillende gebieden en seizoenen zijn significante verschillen (op het 5% niveau) tussen de varianties in twee verschillende 10-jaar periodes geconstateerd. Meest opvallend daarbij is de significante toename van de variantie van maandwaarden van de neerslag over Noord-Europa in drie van de vier seizoenen: zomer, herfst en winter.

In Hoofdstuk 3 worden, voorwaardelijk op de grootschalige atmosferische circulatie, reeksen van dagwaarden van de neerslag en de temperatuur simultaan gegenereerd voor verschillende locaties in het Duitse deel van het stroomgebied van de Rijn. In dit type voorwaardelijke tijdreeks resampling treedt de atmosferische circulatie op als een predictor voor neerslag en temperatuur. Met andere woorden, de atmosferische circulatie bepaalt in belangrijke mate of een dag nat of droog is en of een dag relatief warm of koud is voor de tijd van het jaar. In het geval er een systematische verandering optreedt in de atmosferische circulatie, bijvoorbeeld als gevolg van antropogene klimaatverandering, zal dit automatisch tot een verandering in (extreme) neerslag en temperatuur leiden.

Voorwaardelijke tijdreeks resampling is daarom mogelijk nuttig in toepassingen waar veranderingen in (extreme) neerslag en/of temperatuur van belang zijn. Voorwaardelijke tijdreeks resampling is ook vergeleken met de nauw verwante en al langer in gebruik zijnde analoge methode die tevens populair is bij het maken van weersverwachtingen. Een belangrijke conclusie van die vergelijking is dat de simulatie van neerslag en temperatuur voor een nieuwe dag niet alleen gebaseerd moet worden op de circulatie karakteristieken van die dag maar ook op de gesimuleerde neerslag en temperatuur van de vorige dag. Hiermee wordt de persistentie en variabiliteit in de gesimuleerde tijdreeksen aanzienlijk verbeterd.

In gesimuleerde lange (980-jarige) tijdreeksen, van dagwaarden van de neerslag en de temperatuur voor verschillende locaties in het Duitse deel van het stroomgebied van de Rijn, die representatief zijn voor het huidige klimaat, komen tot 35% grotere 10-daagse gebiedsgemiddelde neerslagsommen voor dan waargenomen in de 35-jarige historische referentie periode. In combinatie met neerslag-afvoer modellen zijn zulke niet eerder waargenomen neerslagextremen zeer nuttig bij het bepalen van extreem zeldzame rivierafvoeren en bij het plannen en ontwerpen van de hydrologische infrastructuur.

Hoofdstuk 4 handelt over (extreme) droogte in Nederland. Omdat grote delen van Nederland tijdens droge periodes voorzien kunnen worden van water uit de Rijn, dient bij de beoordeling van droogte rekening te worden gehouden met het simultaan optreden van neerslagtekorten in Nederland en afvoertekorten van de Rijn. De kans op het simultaan optreden van (grote) neerslag- en afvoertekorten is bepaald op basis van tijdreeks resampling en is vergeleken met schattingen verkregen uit het fitten van tweedimensionale (bivariate) kansverdelingen.

De asymptotische afhankelijkheidsstructuur tussen neerslag- en afvoertekorten blijkt een cruciale rol te spelen bij het bepalen van simultane overschrijdingskansen. Een evident voordeel van tijdreeks resampling is dat geen aannames over de afhankelijkheidsstructuur gemaakt hoeven te worden omdat deze afhankelijkheid automatisch van de historische data wordt overgenomen. Bij het fitten van bivariate kansverdelingen konden alleen bevredigende resultaten worden verkregen door een nieuwe bivariate kansverdeling te introduceren die bestaat uit een mengsel van een bivariate normale en een bivariate Gumbel verdeling. Meer precies, een bivariate normale verdeling met de afhankelijkheidsstructuur van een bivariate Gumbel verdeling. Deze nieuwe tweedimensionale kansverdeling blijkt uiteindelijk het meest geschikt om de simultane overschrijdingskansen van grote neerslag- en afvoertekorten te schatten.

Op basis van deze tweedimensionale kansverdeling worden kleine kansen gevonden (eens in de 200 tot 300 jaar) voor een simultane overschrijding van het neerslagtekort en het afvoertekort zoals voorgekomen in de meest droge historische jaren (1921 en 1976). Bij een faalgebied op basis van de economische schade nemen de kansen op een dergelijke droogte toe tot ongeveer eens in de 100 jaar. Daarnaast zijn de verschillen in de kans op droogte op basis van de verschillende methodes bij zo'n faalgebied kleiner dan bij (standaard) simultane overschrijdingen van het bijbehorende neerslag- en afvoertekort.

Ruimtelijke variatie in de kansverdeling van het neerslagtekort in Nederland is het onderwerp van Hoofdstuk 5. De ruimtelijke variatie wordt in kaart gebracht door Nederland op te delen in zes regio's. Naast verschillen in de grootte van het neerslagtekort, hebben de kansverdelingen voor deze zes regio's een opmerkelijke maar gemeenschappelijke kromming in de rechter staart. De kansverdelingen lijken daardoor noch op een lognormale noch op een extreme waarden verdeling. Om zowel de verschillen in terugkeerniveau[1] van de regio's als de opmerkelijke kromming in de staart van de verdelingen te reproduceren zijn twee alternatieve methodes beschouwd: een regionale frequentieanalyse en tijdreeks resampling.

De regionale frequentieanalyse levert een voor Nederland representatieve 'growth curve' op die benaderd kan worden met een differentieerbare (spline) functie op de schaal van de standaard Gumbel variabele. De tijdreeks resampling methode heeft een extra geheugenterm van vier maanden nodig om de gemeenschappelijke kromming in de staart van de neerslagtekortverdelingen te kunnen reproduceren. Door deze geheugenterm neemt de statistische onzekerheid (standard error) in schattingen van het terugkeerniveau ongewild toe. Echter, wanneer de beide methodes worden gebruikt voor extrapolatie, d.w.z. wanneer het terugkeerniveau groter is dan historisch waargenomen, dan presteert tijdreeks resampling beter dan de regionale frequentieanalyse. Dit komt door de sterke toename van de relatieve standard error van het terugkeerniveau met toenemende herhalingstijd als gevolg van de grote onzekerheid in de parameters van de functie die de 'growth curve' benadert. Dus in termen van statistische onzekerheid is tijdreeks resampling te verkiezen boven de regionale frequentieanalyse. Dit is het type resultaten waaruit de meerwaarde van tijdreeks resampling blijkt.

Door tijdreeks resampling te combineren met de bootstrap procedure is, voor elk van de zes regio's een betrouwbaarheidsinterval afgeleid voor de herhalingstijd behorende bij het grootst waargenomen neerslagtekort. Afhankelijk van de regio is dit het neerslagtekort van 1959 of 1976. Hoewel deze betrouwbaarheidsintervallen tamelijk breed zijn, zijn ze gemiddeld een factor twee smaller dan verwacht mag worden indien alleen rekening wordt gehouden met de lengte van de waargenomen reeks.

In de synthese in Hoofdstuk 6 passeert een aantal onderwerpen de revue die mede van invloed zijn op de onzekerheid van statistische analyses in het algemeen en van statistische analyses op basis van tijdreeks resampling in het

---

[1] Het niveau van het neerslagtekort dat hoort bij een bepaalde herhalingstijd of terugkeertijd.

bijzonder. Zo wordt uitvoerig stilgestaan bij de lengte en de periode van de waarnemingsreeksen die worden gebruikt, en bij de gevolgen van de mogelijke aanwezigheid van (historische) trends in die reeksen.

In Hoofdstuk 6 wordt ook aandacht geschonken aan het gebrek aan met name lange waarnemingsreeksen van de potentiële (Makkink) verdamping in Nederland. De oorzaak hiervoor is dat de benodigde globale straling voor de bepaling van de Makkink verdamping pas vanaf 1958 structureel door het KNMI wordt gemeten. In Hoofdstuk 4 is voor station De Bilt een lange reeks (1906–2000) van de Makkink verdamping geconstrueerd door van (goede) schattingen van de globale straling op basis van de zonneschijnduur uit te gaan. Voor vier andere klimatologische hoofdstations in Nederland zijn ook lange zonneschijnduur reeksen beschikbaar en daarnaast bestaat er nog een door Wageningen Universiteit gemeten reeks van de globale straling die al in 1928 begint. Voor beide geldt echter dat, voor de hier beoogde toepassing, de kwaliteit en de homogeniteit van deze reeksen onderzocht zou moeten worden.

Nu duidelijk is gemaakt dat tijdreeks resampling zeer nuttig kan zijn bij het modelleren en analyseren van (nog niet waargenomen) extreme hydrometeorologische gebeurtenissen, zou bij toepassing van deze methode ook structureel een onzekerheidsanalyse op basis van jackknife of bootstrap procedures, zoals beschreven in Hoofdstuk 5, moeten plaatsvinden. Dit kan, en moet, gebruikers bewuster maken van de (grote) onzekerheid die ‘extrapolatie’ van extreme gebeurtenissen uit relatief korte historische reeksen met zich meebrengt.

Curriculum Vitae

Al vanaf mijn geboorte op 2 mei 1964 word ik, Julius Johan Beersma, Jules genoemd. In mijn geboortestad Zwolle doorliep ik achtereenvolgens de kleuter, lagere en middelbare school. Op de laatste, de Thomas a Kempis Scholengemeenschap, heb ik in 1982 een Atheneum-B diploma behaald. Een studie natuurkunde — destijds mijn favoriete vak — aan de Rijksuniversiteit Groningen, waar zowel in technische, experimentele als theoretische richting afgestudeerd kon worden, volgde. Mijn keus viel uiteindelijk op theoretische natuurkunde met een maximum aan informatica bijvakken. Na een korte omzwerving in de wetenschapsfilosofie, ben ik in 1989 afgestudeerd op een theoretisch onderzoek bij de experimentele vakgroep Vaste Stof Fysica, met als titel ‘Colloid growth theory for irradiated NaCl’. In zekere zin ben ik die mengeling van theorie en praktijk tot op heden trouw gebleven.

In 1990 kon ik in het kader van een ‘tewerkstelling’ als erkend gewetensbezwaarde aan de slag bij het KNMI in De Bilt. (Eigenlijk had ik gehoopt dat van uitstel — van militaire dienst — afstel zou komen maar zover is het, achteraf gezien, gelukkig niet gekomen.) In die tijd kwam het onderzoek naar antropogene klimaatverandering sterk in de belangstelling. In 1992, kort voor het einde van de tewerkstelling, wist ik een toegevoegde vaste aanstelling in het kader van de versterking van het KNMI klimaatprogramma te verwerven. Sindsdien ben ik onafgebroken als onderzoeker in dienst van het KNMI geweest. Eerst bij de hoofdafdeling Wetenschappelijk Onderzoek en op dit moment bij de daaruit ontstane sector Klimaat en Seismologie. Echter vanaf het begin in dezelfde ‘onderzoeksgroep’ die in de loop der tijd zeker zeven verschillende namen heeft gehad en sinds de reorganisatie van het afgelopen jaar afdeling Klimaatdata en -advies heet. Daarnaast ben ik, vanaf het einde van de jaren negentig, nog een deel van mijn tijd bij de afdeling Voorspelbaarheidsonderzoek (VO) gedetacheerd geweest. Daar heb ik een simpel diagnostisch wolkenschema en een nieuwe parameterisatie voor kortgolvige straling voor het ‘intermediate complexity’ klimaatmodel ECBilt ontwikkeld.

In mijn eerste KNMI jaren heb ik vooral aan de ontwikkeling en verbetering van toetsen ten behoeve van het detecteren van klimaatveranderingen gewerkt en statistische analyses van de uitvoer van state-of-the-art klimaatmodellen gedaan die van belang waren voor het construeren van klimaatscenario's. Die analyses bestonden zowel uit het valideren van de uitvoer van de controle runs van klimaatmodellen als het kwantificeren van veranderingen in het (model)klimaat ten gevolge van de toename van de concentraties van broeikasgassen in de atmosfeer (inclusief het vaststellen van de statistische significantie van zulke veranderingen). Vrij uniek in die tijd was dat we niet alleen naar veranderingen in de gemiddelden keken maar ook naar veranderingen in variabiliteit en naar veranderingen in de statistische eigenschappen van dagwaarden van klimaatvariabelen, hetgeen voor het construeren van klimaatscenario's van wezenlijk belang is. Mede door de ervaring met resampling technieken die ik in die periode heb opgedaan ben ik later betrokken geraakt bij het zogenaamde 'Neerslaggenerator voor de Rijn' project (1996–heden) waarin het KNMI nauw met RIZA samenwerkt en waarin nearest-neighbour resampling van meteorologische tijdreeksen de basis vormt. Hieruit zijn weer andere projecten en opdrachten ontstaan waaraan ik heb meegewerkt zoals die voor de kans op droogte in Nederland ten behoeve van de door RIZA gecoördineerde 'Droogtestudie Nederland' (2001–2005).

Acknowledgments. A large part of this thesis is directly or indirectly the product of the formal collaboration[†] between the sister-institutes RIZA[‡] and KNMI, which are both part of the Dutch Ministry of Transport and Public Works, where RIZA asks the questions and KNMI tries to provide scientifically sound answers. The topic of Chapter 3 fitted in the framework of the "Neerslaggenerator voor de Rijn" (Precipitation generator for the river Rhine) project, a joint RIZA-KNMI effort. The work in Chapters 4 and 5 is the result of the KNMI contribution to the "Droogtestudie Nederland" (Drought studies of the Netherlands) and the "Regionale droogtestudie Midden-west Nederland" (Regional drought study for the Mid-west of the Netherlands) which were both co-ordinated by RIZA.

The author is very grateful to Kees Floor for making the cover photo available and to Marna Slappendel for the artistic photo editing.

[†] Under contracts: RI-2041, RI-2726(A), RI-3414A, RI-4129 and RI-4339.

[‡] Institute for Inland Water Management and Waste Water Treatment, Lelystad.