CHALMERS | UNIVERSITY OF GOTHENBURG

PREPRINT 2008:28

Modelling Precipitation in Sweden Using Multiple Step Markov Chains and a Composite Model

JAN LENNARTSSON
ANASTASSIA BAXEVANI
DELIANG CHEN

Department of Mathematical Sciences
Division of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Göteborg Sweden 2008

Preprint 2008:28

Modelling Precipitation in Sweden Using Multiple Step Markov Chains and a Composite Model

Jan Lennartsson, Anastassia Baxevani,
Deliang Chen

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg, Sweden
Göteborg, September 2008

Preprint 2008:28
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2008

Modelling Precipitation in Sweden Using Multiple Step Markov Chains and a Composite Model

Jan Lennartsson[1], Anastassia Baxevani[1*†], Deliang Chen[2]

[1] Department of Mathematical Sciences, Chalmers University of Technology, University of Gothenburg, Gothenburg, Sweden

[2] Department of Earth Sciences, University of Gothenburg, Gothenburg, Sweden

Abstract

In this paper, we propose a new method for modelling precipitation in Sweden. We consider a chain dependent stochastic model that consists of a component that models the probability of occurrence of precipitation at a weather station and a component that models the amount of precipitation at the station when precipitation does occur. For the first component, we show that for most of the weather stations in Sweden a Markov chain of an order higher than one is required. For the second component, which is a Gaussian process with transformed marginals, we use a composite of the empirical distribution of the amount of precipitation below a given threshold and the generalized Pareto distribution for the excesses in the amount of precipitation above the given threshold. The derived models are then used to compute different weather indices. The distribution of the modelled indices and the empirical ones show good agreement, which supports the choice of the model.

KEY WORDS: High order Markov chain, generalized Pareto distribution, copula, precipitation process, Sweden

*Corresponding author

†Research supported partially by the Gothenburg Stochastic Center and the Swedish foundation for Strategic Research through Gothenburg Mathematical Modelling Center.

1 Introduction

Realistic sequences of meteorological variables such as precipitation are key inputs in many hydrologic, ecologic and agricultural models. Simulation models are needed to model stochastic behavior of climate system when historical records are of insufficient duration or inadequate spatial and /or temporal coverage. In these cases synthetic sequences may be used to fill in gaps in the historical record, to extend the historical record, or to generate realizations of weather that are stochastically similar to the historical record. A weather generator is a stochastic numerical model that generates daily weather series with the same statistical properties as the observed ones, see Liao *et al.* (2004).

In developing the weather generator, the stochastic structure of the series is described by a statistical model. Then, the parameters of the model are estimated using the observed series. This allows us to generate arbitrarily long series with stochastic structure similar to the real data series.

Parameter estimation of stochastic precipitation models has been a topic of intense research the last 20 years. The estimation procedures are intrinsically linked to the nature of the precipitation model itself and the timescale used to represent the process. There are models which describe the precipitation process in continuous time and models describing the probabilistic characteristics of precipitation accumulated on a given time period, say daily or monthly totals. Different reviews of the available models have been presented: see for example Woolshiser (1992), Cox and Isham (1988) and Smith and Robinson (1997).

Continuous time models for a single site with parameters related to the underlying physical precipitation process are particularly important for the analysis of data at short timescales, e.g. hourly. Some of these models are described in Rodríguez-Iturbe *et al.* (1987, 1988) and Waymire and Gupta (1981).

When only accumulated precipitation amounts for a particular time period (daily) are recorded then empirical statistical models, based on stochastic models that are calibrated from actual data are appealing. Empirical statistical models for generating daily precipitation data at a given site can be classified into four different types, chain dependent or two-part models, transition probability matrix models, resampling models and ARMA time series models, see Srikanthan and McMahon (2001) for a complete review of the

different models.

A generalization of the precipitation models for a single site is the spatial extension of these models for multiple sites, to try to incorporate the intersite dependence but preserving the marginal properties at each site. A more ambitious task is the modelling of precipitation continuously in time and space and original work on these type of models based on point process theory was presented by LeCam (1961) and further developed by Waymire *et al.* (1984) and Cox and Isham (1994). Mellor (1996) has developed the modified turning bands model which reproduces some of the physical features of precipitation fields in space as rainbands, cluster potential regions of rain cells.

In this study we concentrate on the chain-dependent model for the daily precipitation in Sweden which consists of two steps, first a model for the sequence of wet/dry days and second, a model for the amount of precipitation for the wet days. For the first, we use high-order Markov chains and for the second we introduce a composite model that incorporates the empirical distribution and the generalized Pareto distribution.

2 Data

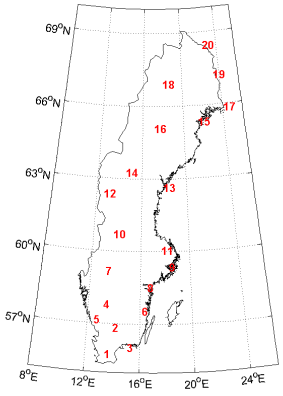

FIG.1: Location of the stations.

Precipitation data from 20 stations in Sweden have been used in the studies presented in this paper. The locations are shown in Fig. 1 and the names of the stations are given in Table 1. The data consist of accumulated daily precipitation collected during 44 years starting on the 1^{st} of January 1961 and ending the 31^{st} of December 2004 and are provided by the Swedish Meteorological and Hydrological Institute (SMHI). The number of missing

observations in all stations is generally low ($< 5\%$). The time plots of the annual number of wet days (above the threshold 0.1 mm) at the 20 stations are presented in Fig. 2.

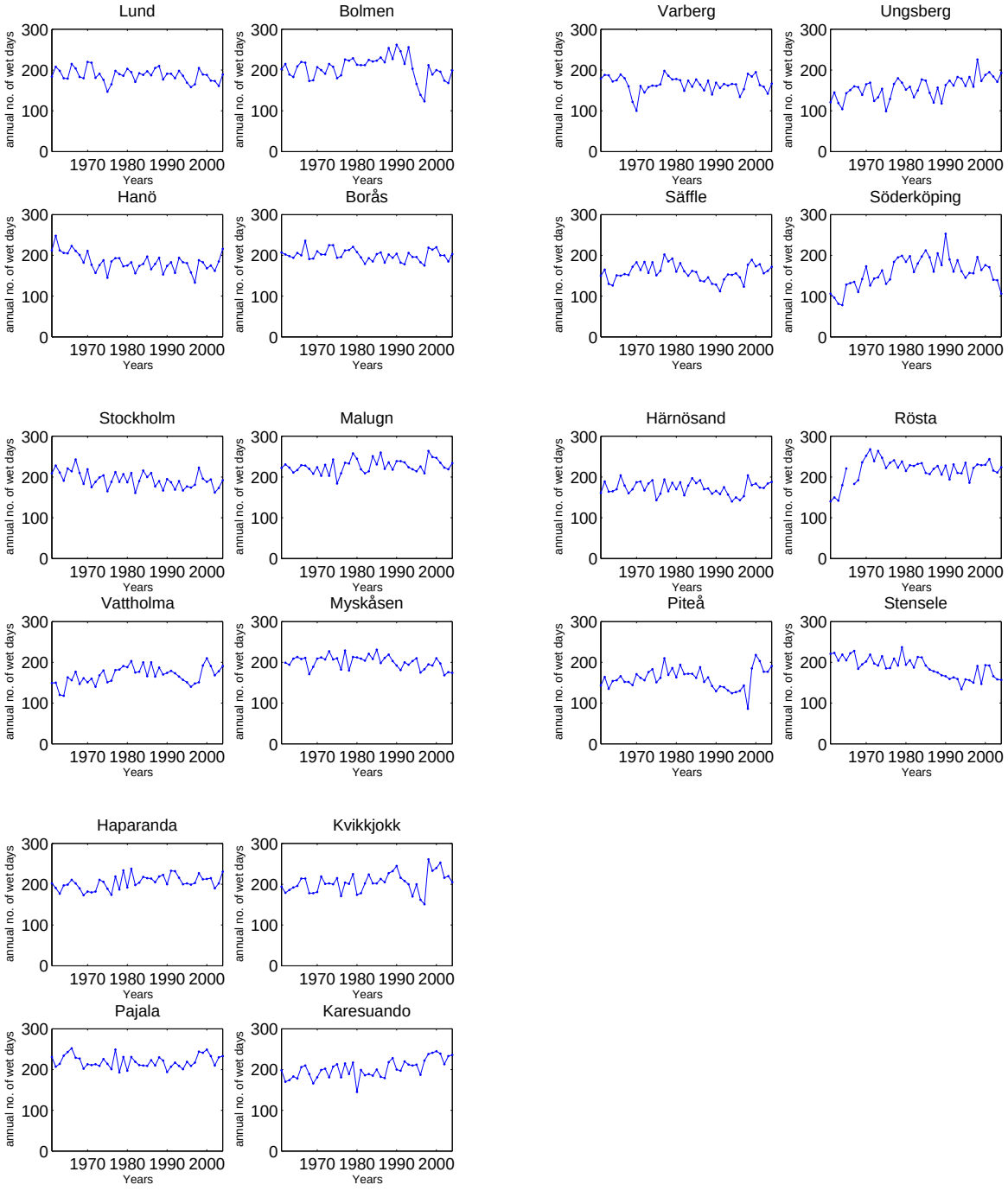

FIG.2: Time plot of annual number of wet days.

Time plots of annual number of wet days showed that the precipitation regime in some stations (namely, Söderköpping, Rösta and Stensele) contains possible trends. The

results presented in the next sections refer to the whole period of data from all stations, but attention should be paid when we refer to the above mentioned stations. In Fig. 3, time plots of the annual amount of precipitation of the wet days are presented. The total amounts of precipitation seem to be stationary over the different years.

3 Model

To model precipitation in Sweden, we have decided to use a chain dependent model. The first part of the model can be dealt with using Markov chains. Gabriel and Newman (1962) used a first-order stationary Markov chain. The models have since been extended to allow for non-stationarity, both by fitting separate chains to different periods of the year and by fitting continuous curves to the transition probabilities, see Stern and Coe (1984) and references within. The order of Markov chain required has also been discussed extensively, for example Chin (1977) and references therein, with the obvious conclusion that different sites require different orders. Still, the first order Markov chains are a popular choice since they have been shown to perform well for a wide range of different climates, see for example Bruhn *et al.* (1980), Lana and Burgueno (1998) and Castellvi and Stockle (2001). The main deficiency associated with the use of first order models is that long dry spells are not well reproduced, see Racsko *et al.* (1991), Guttorp (1995).

To model the amount of precipitation that has occurred during a wet day, different models have been proposed in the literature all of which assume that the daily amounts of precipitation are independent and identically distributed. Stidd (1973) and Hutchinson (1995) have proposed a truncated normal model for the amount of precipitation with a time dependent parameter, while the Gamma and Weibull distributions have been selected by Geng *et al.* (1986) as well as Selker and Heith (1990), because of their site-specific shape.

In this study, we model the occurrence of wet/dry days using Markov chains of higher order and for the amount of precipitation we use a composite model, consisting of the empirical distribution function for values below a threshold and the distribution of excesses for values above the given threshold. Such a model is more flexible, describes better the tail of the distribution and additionally allows for dependence in the precipitation process.

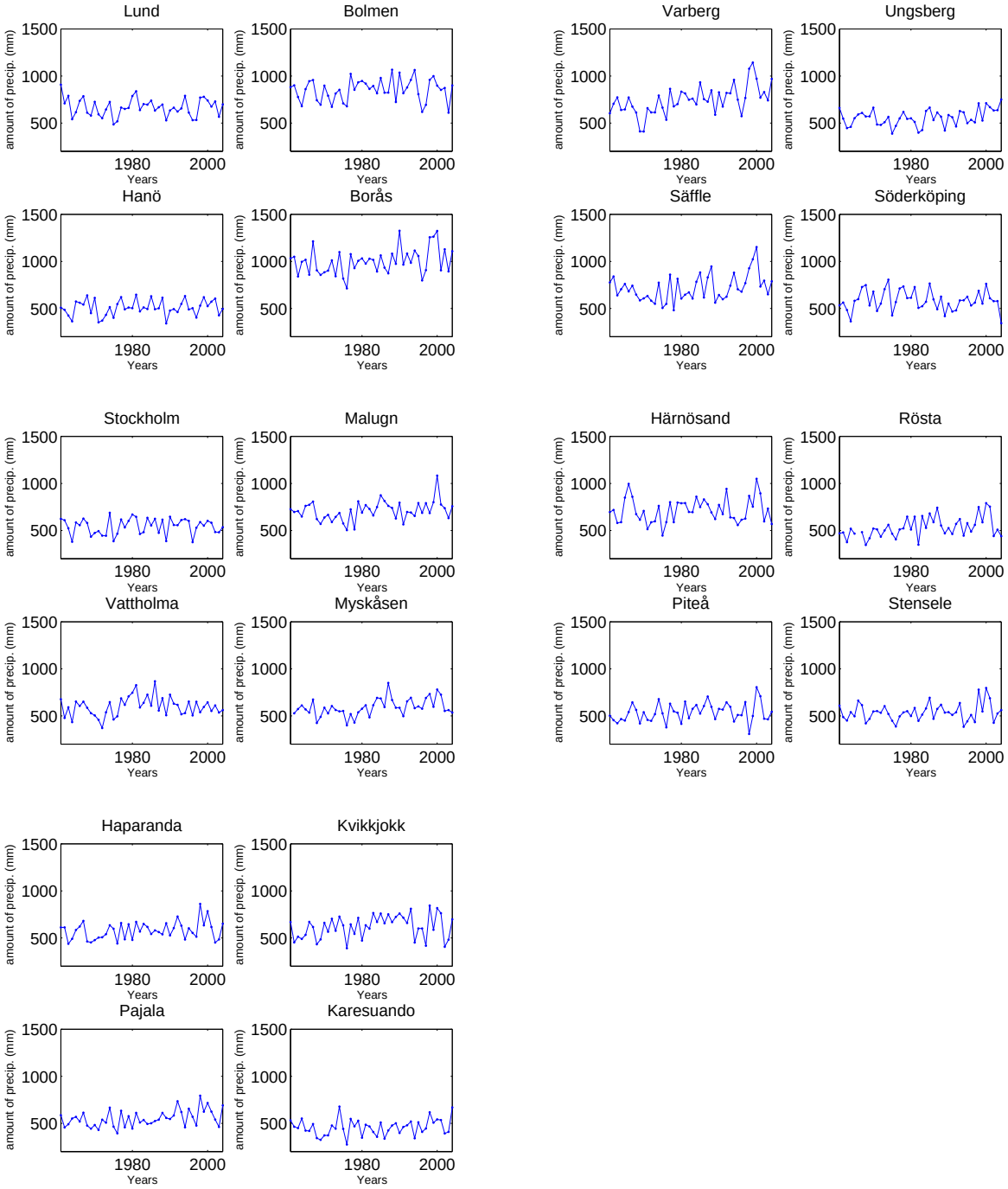

FIG.3: Time plot of annual amount of precipitation.

Let Z_t be the precipitation at a certain site at time t measured in days. Then, a chain-dependent model for the precipitation is given by,

$$Z_t = X_t W_t,$$

where X_t and W_t are stochastic processes such that X_t takes values in $\{0, 1\}$ and W_t

Number	Name
1	Lund
2	Bolmen
3	Hanö
4	Borås
5	Varberg
6	Ungsberg
7	Säffle
8	Söderköping
9	Stockholm
10	Malung
11	Vattholma
12	Myskelåsen
13	Härnösand
14	Rösta
15	Piteå
16	Stensele
17	Haparanda
18	Kvikkjokk
19	Pajala
20	Karesuando

Table 1: Names of weather stations.

takes values in $\mathbb{R}^+ \setminus 0$. The processes X_t and W_t will be referred to as the occurrence of precipitation and the amount of precipitation process, respectively.

The approach presented in this study provides a mechanism to make predictions of precipitation in time. This is particularly important for many applications in hydrology, ecology and agriculture. For example, at a monthly level, the amount of precipitation and the probability and length of a dry period are required quantities for many applications.

4 Models for the Occurrence of Precipitation

Let $\{X_t, t = t_1, \ldots, t_N\}$ denote the sequence of daily precipitation occurrence, i.e. $X_t = 1$, indicates a wet day and $X_t = 0$, a dry day. A wet day in the context of this study, occurs when at least $0.1mm$ of precipitation was recorded by the rain gauge. The level has been chosen above zero in order to avoid identifying dew and other noise as precipitation and to also avoid difficulties arising from the inconsistent recording of very small precipitation amounts. Moreover, daily precipitation amounts of less than $0.1mm$ can have relatively large observational errors, and including them would cause a significant change in the estimated transition probabilities of the occurrences. As a consequence this introduces additional errors into the fitted models. The model is fitted over different periods of the year, that is subsets of the N days of the year, that may be assumed stationary.

Before we continue any further we need to introduce some notation. Let $S = \{0, 1\}$ denote the state space of the k-Markov chain X_t. The elements of S are called letters and an ordering of letters $w \in S^l = S \times \cdots \times S$ is called a word of length l, while the words composed of the letters from position i to j in w for some $1 \leq i \leq j \leq l$, are denoted as $w_i^j = (w_i, w_{i+1}, ..., w_j)$. Finally, for $k \leq l$ let $\tau_k(w) = w_{l-k+1}^l$ denote the k-tail of the w word, i.e. $\tau_k(w)$ denotes the last k letters of w. If no confusion will arise when $k \leq j - i$, we also write $\tau_k(w^j)$ instead of $\tau_k(w_i^j)$.

It is assumed that the process X_t is a k-Markov chain: a model completely characterized by the transition probability

$$p_{w,j}(t) := P(X_t = j | \tau_k(X^{t-1}) = w), \quad j \in S, \quad t = t_1, \ldots, t_N,$$

where w is a word of length k and $X^{t-1} = \{\ldots, X_{t-2}, X_{t-1}\}$ is the whole process up to $t - 1$ so $\tau_k(X^{t-1})$ is the last k days up to and including X_{t-1}; that is, $\tau_k(X^{t-1}) = (X_{t-k}, \ldots, X_{t-1})$. Note that, for a 2−state Markov chain of any order $p_{w,1}(t) + p_{w,0}(t) = 1$. In the special case of time homogeneous Markov chain, $p_{w,j}(t) = p_{w,j}$, for $t = t_1, \ldots, t_N$, i.e. the transition probabilities are independent of time.

Let $n_{w,j}(t)$ denote the number of years during which day t is in state j and is preceded by the word w (i.e. $\tau_k(X^{t-1}) = w, w \in S^k$ and $X_t = j$). Then the probabilities $p_{w,j}(t)$ are estimated by the observed proportions

$$\hat{p}_{w,j}(t) = \frac{n_{w,j}(t)}{n_{w,+}(t)}, \quad w \in S^k, \quad j \in S, \quad t = t_1, \ldots, t_N,$$

where + indicates summation over the subscript. Note also that day 60 (February 29$^{\text{th}}$) has data only in leap years so day 59 precedes day 61 in non-leap years. Fig. 4 (left) shows the unconditional probability of precipitation, pooled over 5 days for clarity, plotted against t for the data from the station in Lund.

In the context of environmental processes, non-stationarity is often apparent, as in this case, because of seasonal effects or different patterns in different months. A usual practice is to specify different subsets of the year as seasons, which results to different models for each season, although the determination of an appropriate segregation into seasons is itself an issue.

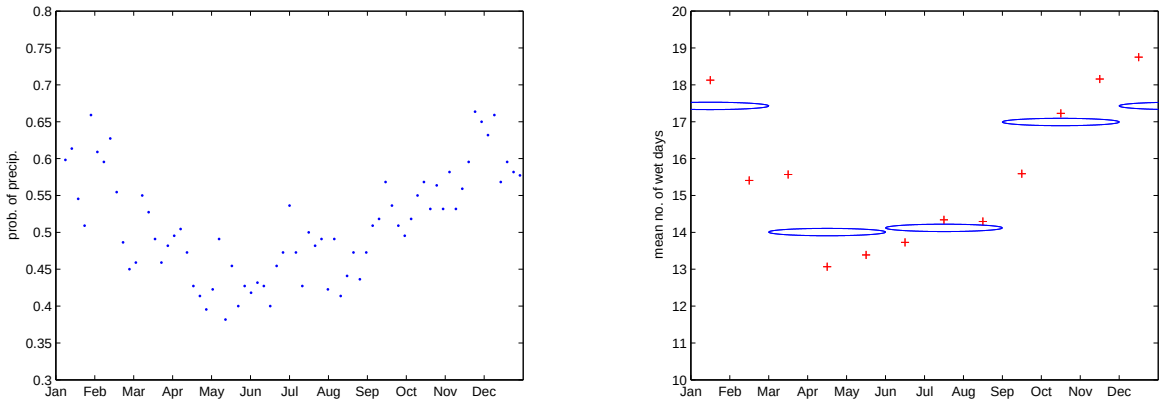

FIG.4: Lund, Sweden (data from 1961 to 2004). (Left): Observed $\hat{p}(t)$ pooled over 5 days. (Right): Mean number of wet days per month ("+"), and per season (solid lines).

4.1 Fitting Models to the Occurrence of Precipitation

There is an inter-annual variation in the annual number of wet days, as can be seen in Fig. 2. Moreover, there is also seasonal variation in the mean monthly number of wet days, see Fig. 4 (Right) for data from Lund, although this is not as prominent as in other regions of the world. It is possible that the optimum order of the chain describing the wet/dry sequence varies within the year and from one year to another. It is therefore important to properly identify the period of record that can be assumed as time homogeneous.

Moreover, the problem of finding an appropriate model for the occurrence of precipitation process, X_t, is equivalent to the problem of finding the order of a multiple step Markov chain. The Akaike Information Criterion (AIC), Bayesian Information Criterion

(BIC) and the Generalized Maximum Fluctuation Criterion (GMFC) order estimators, a short description of which can be found in the subsection 8.1, have been applied to the data for each of the stations. Various block lengths were considered for determining the order of the Markov chain, k, as suggested in Jimoh and Webster (1996).

- 1 month blocks (i.e. January, February, ..., December),
- 2 month blocks (January - February, February - March, ..., December - January),
- 3 month blocks (January - March, February - April, ..., December - February).

The effect of block length on the order of the Markov chain can be seen in Figs. 5-7. We can notice that grouping the data in blocks of length more than one month, results in Markov chains of "smoother" order, in the sense that the order of the chain does not change so fast. It is also interesting to notice that while the order of the Markov chain for the stations 16-20, varies a lot according to the AIC and GMFC estimators it seems to be almost constant for the BIC order estimator. As it has been expected, the BIC order estimator underestimates the order k of the Markov chain relatively to both the AIC and GMFC order estimators for large k and moderate data sets, see Dalevi *et al.* (2006), while the values of the GMFC order estimator lie between the BIC and AIC ones. The results presented in Figs. 5-7, confirm that the model order is sensitive to the season (month) and the length of the season (number of months) considered, as well as the method used in identifying the optimum order. Possible dependence on the threshold used for identifying wet and dry days has not been studied here. For the rest of this study, we define as seasons the 3 month periods, December-February, March-May, June-August, September-November. As can be seen in Fig. 3 for the station in Lund, the rest of the stations provide with similar plots, the probability of precipitation is close to be constant during these periods, which makes the assumption of stationarity seem plaussible. The orders of the Markov chain for these periods can be found in Fig. 7. For the rest of this study the order k of the Markov chain is decided according to the GMFC order estimator.

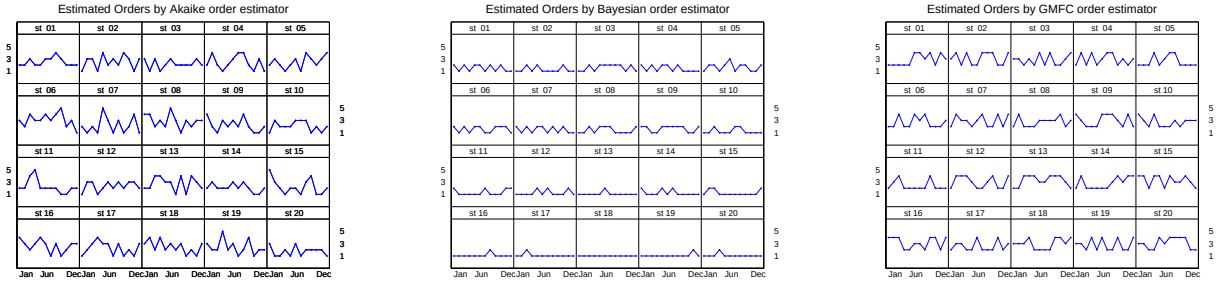

FIG.5: k-Markov chain orders for block lengths of one month, (Jan, Feb, ...).

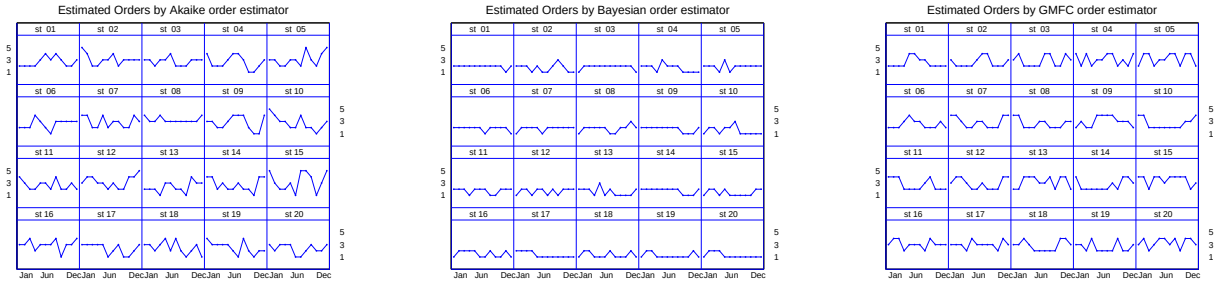

FIG.6 k-Markov chain orders for block lengths of two months, (Jan-Feb, Feb-Mar, ...).

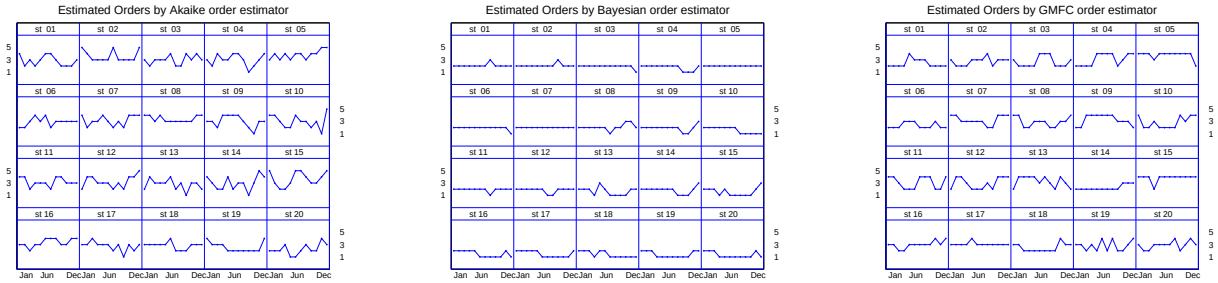

FIG.7 k-Markov chain orders for block lengths of three months, (Jan-Mar, Feb-Apr, ...).

4.2 Distribution of Dry Spell length

An interesting aspect of the wet/dry behavior, i.e. the process X_t, is the distribution of the dry (wet) spells, i.e., the number of consecutive dry (wet) days, which is an accessible property of multiple step Markov chains.

For a time homogeneous (stationary) k-Markov chain X_t, ($k \geq 2$), with state-space S let T be the first time the process X_t is such that $\tau_2(X^t) = (1,0)$, i.e.,

$$T = \inf\{t \geq 0 : \tau_2(Y^t) = (1,0)\}.$$

So T is the time of the start of the first dry period. Let also for the words $u, v \in S^k$

$$a_{u,v} = P(\tau_k(X^T) = v | \tau_k(X^0) = u)$$

denote the probability the process X_t has at time T a k-tail equal to v given that the k-tail at time 0 is equal to u. The probabilities $a_{u,v}$ are easily obtained for stationary processes, see Norris (1997). Note that at $t = 0$, there may be the start of a dry period, the start of a wet period, the continuation of a dry period or the continuation of a wet period. If $D(X_t)$ denotes the length of the first dry period that starts at time $t = 0$ for the k-Markov chain X_t, then assuming additionally that the process X_t is time homogeneous, the distribution of the first dry spell can be computed as

$$P(D(X_t) = m) = \sum_{\{u \in S^k\}} \pi_u \sum_{\{w \in S^k : \tau_2(w) = (1,0)\}} a_{u,w} P(\tau_m(X^{m-1}) = \mathbf{0}, X_m = 1 | \tau_k(X^0) = w), \tag{1}$$

where $\mathbf{0}$ is used to denote sequences of 0′s of appropriate length.

Now, if $v = w\mathbf{0}1$ is a word of length $m + k$ ($\mathbf{0}$ here is of order $m - 1$) and using the fact the process X_t is a k-Markov chain, Eq.1 can be rewritten as

$$P(D(X_t) = m) = \sum_{\{u \in S^k\}} \pi_u \sum_{\{w \in S^k : \tau_2(w) = (1,0)\}} a_{u,w} \prod_{i=1}^{m} P(X_i = v_{k+i} | \tau_k(X^{i-1}) = \tau_k(v^{k+i-1})). \tag{2}$$

Remark 1 *Here we should notice that the distribution of the first dry spell is different than the distribution of the subsequent dry spells for Markov chains of order greater than two. For one or two order Markov chains there is no need for this distinction. Moreover the equivalent of Eq. 1 for $k = 1$ is*

$$P(D(X_t) = m) = p_{0,0}^{m-1} p_{0,1}$$

while for $k = 2$, Eq. 2 simplifies to

$$P(D(X_t) = m) = \begin{cases} p_{10,1} & \text{for } m = 1 \\ p_{10,0} p_{00,0}^{m-2} p_{00,1} & \text{for } m \geq 2, \end{cases}$$

where $a_{u,v} = 1$ for all u,v in Eq. 1.

The distribution of the first dry spell can be also used for model selection or model validation purposes. For this, we use the Kolmogorov-Smirnov (KS) test, see Benjamin and Cornell (1970). The one sample KS test compares the empirical distribution function with the cumulative distribution function specified by the null hypothesis.

Assuming that $P_k(x)$ is the true distribution function (of a Markov chain of order k) the KS test is

$$D = \sup_{m \in \mathbb{N}^+} |P_k(D(X) \leq m) - F_{\text{emp}}(m)|,$$

where $F_{\text{emp}}(x)$ is the empirical cumulative distribution of the length of the first dry spell. If the data comes truly from a k order Markov chain and the transition probabilities are the correct ones, then by Glivenko-Cantelli theorem, the KS test converges to zero almost surely (a.s).

To apply the test, the transition probabilities have been estimated from the data using maximum likelihood for different values of the order k of the Markov chain. To obtain the empirical distribution of the length of the first dry spell, we have computed the length of the dry spells (sequence of zeros) following the first $(1, 0)$. (Here note that this is equivalent to computing the length of the first dry spell for Markov chains of order $k = 1$ or $k = 2$. In the case of $k = 3$, although the distribution of the first dry spell is not exactly the same as the distribution of any dry spell, we have still used all the dry spells available due to shortage of data.) The procedure has been applied separately to data from each station and season. If the first observations were zeros, they were ignored as the continuation of a dry spell. Also if a dry spell was not over by the end of the season then it was followed inside the next season.

To determine whether the theoretical model was correct or not, Monte Carlo simulations were performed. We have obtained the empirical distribution of the length of the first dry spell using 500 synthetic wet/dry records of 44 years of data (each station and season was treated separately), and the KS test was computed for each one of them, which resulted to the distribution of the KS statistic.

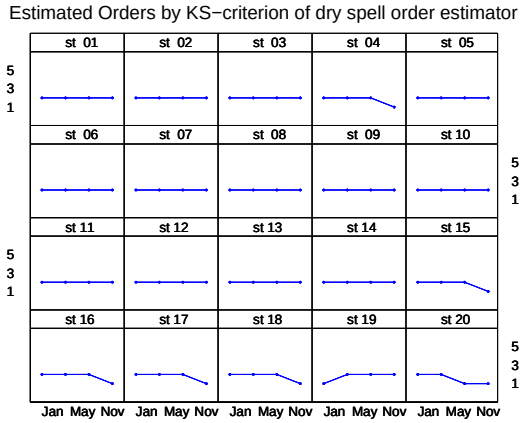

FIG.8: Order of Markov chain as suggested by the Kolmogorov-Smirnov statistic at 10% tail value for each station and season.

The suggested orders of the Markov chain using the Kolmogorov-Smirnov statistic at the 10% tail value are collected in Fig. 8. The resulting orders of the Markov chain appear to be close to those obtained by the BIC order estimator. In Table 2, we have collected information on how many data sets have passed the Kolmogorov-Smirnov test at the 10% tail value for the different seasons. Observe that the KS test suggests that the 1-Markov chain, although widely used, is an inadequate model for the majority of the stations in Sweden over the different seasons.

	Season			
Model	$S1$	$S2$	$S3$	$S4$
$k=1$	1	0	1	6
$k=2$	20	20	20	20
$k=3$	20	20	20	20

Table 2: Number of data sets that have passed the Kolmogorov-Smirnov test at the 10% tail value for different orders of the Markov chain. $S1$ stands for Dec.-Feb., $S2$ for Mar.-May, $S3$ for Jun.-Aug. and $S4$ for Sep.-Nov.

4.3 Distribution of Long Dry Spells

Let us now define as long dry spell, a dry spell with length longer or equal to the order k of the Markov chain. Then it is easy to show that the distribution of the long dry spell is

actually geometric. Indeed, let a long dry spell that starts at time i have length $m \geq k$ and let us also assume that we know that the length of the dry spell is at least l. Then, for $m \geq l \geq k$

$$P(D(X_t) = m | \tau_l(X^{i+l-1}) = \mathbf{0}) = p_{\mathbf{0},1} p_{\mathbf{0},0}^{m-l} = p_{\mathbf{0},1}(1 - p_{\mathbf{0},1})^{m-l},$$

where as before

$$p_{\mathbf{0},1} = P(X_{n+1} = 1 | \tau_k(X^n) = \mathbf{0}), \quad \forall n.$$

Therefore, the expected length of long dry spells is given by

$$E(D(X_t) | \tau_l(X^{i+l-1}) = \mathbf{0}) = l + \frac{1 - p_{\mathbf{0},1}}{p_{\mathbf{0},1}}. \tag{3}$$

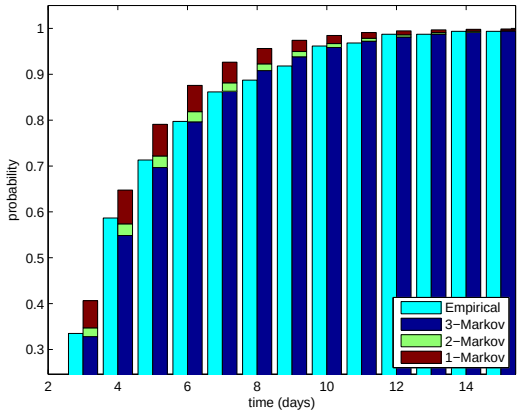

FIG. 9: Conditional distribution of Dry Spell given the Dry Spell is longer or equal to 3 days for k-Markov chain models of order $k = 1$, $k = 2$ and $k = 3$ and the data from Lund. Data are from the winter months December-February.

Fig. 9 shows the conditional distribution of dry spell given that it has lasted for more than two days for the first season and the data from Lund. The estimated order of the Markov chain for this data set is 2 using both the GMFC and the KS criterion. A first order Markov chain, the popular model of choice in this case would obviously underestimate the risk of a long dry spell. A two order Markov chain seems to be the best choice for this particular data set.

It is clear from Table 3, that underestimation of the order k of the Markov chain leads to underestimation of the expected length of the long dry spells, where again a dry spell is defined as long if it has length larger than or equal to the order of the Markov chain.

Model	$l=1$	$l=2$	$l=3$
$k=1$	2.49	3.49	4.49
$k=2$	-	3.91	4.91
$k=3$	-	-	5.11
Observed mean value	2.56	3.97	5.23

Table 3: Expected length of long dry spells for season Dec-Feb in Lund.

5 Modeling the Amount Precipitation Process

In this section we model the amounts of daily precipitation. This is done in two steps. Firstly we model the dependence structure of the amount precipitation process and secondly we estimate the marginal distribution.

One of the important features of any climatological data set, is that they exhibit dependence between nearby stations or successive days. In this work we are interested in the latter case and the dependence structure is modelled using two-dimensional Gaussian copula.

After the copula has been estimated, we remove the days with precipitation below the cut-off level of $0.1mm$. That is, we let Y_t be the thinning process resulting from the amount of precipitation process W_t when we consider only the wet days, i.e., $Y_t := W_t|X_t = 1$. Then, the marginal distribution of the amounts of daily precipitation is modelled following an approach that combines the fit of the distribution of excesses over a high threshold with the empirical distribution of the thinned data below the threshold.

5.1 Copula

Almost every climatological data set exhibit dependence between successive days. To model the temporal dependence structure of the data we use the two-dimensional Gaussian copula C given by

$$\begin{aligned} C(u,v;\rho) &= \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}} dxdy \qquad (4) \\ &= \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)), \end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution and Φ_ρ is the joint cumulative distribution function of two standard normal random variables with correlation coefficient ρ.

To estimate the copula, let

$$A = \{t : Y_t > 0 \text{ and } Y_{t+1} > 0\},$$

be the set of all days with non zero precipitation that were followed by days also with non zero precipitation (greater than $0.1mm$) and

$$\mathbf{u} = [Y_{a_1}, Y_{a_2}, \dots], \quad \mathbf{v} = [Y_{a_1+1}, Y_{a_2+1}, \dots], \quad a_1, a_2, \dots \in A$$

be the vectors consisting of the amounts of precipitation during the days indicated in the set A and the following days respectively, both with marginal distribution $F(x)$. Then, transforming the vectors $\mathbf{u}$ and $\mathbf{v}$ by taking the empirical cumulative distribution corrected by the factor $\frac{n}{n+1}$, (n is the number of days with positive precipitation in the data set) results to vectors $\mathbf{U}$ and $\mathbf{V}$ respectively that follow the discrete uniform distribution in $(0, 1)$. If the Gaussian copula in Eq. 4 describes correctly the dependence structure of the data, then

$$\left(\Phi^{-1}(\mathbf{U}), \Phi^{-1}(\mathbf{V})\right) \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2 \end{bmatrix}\right).$$

Finally the copula parameter ρ is estimated using Pearson's correlation coefficient. An analytic description of the method and its application can be found in Lennartsson and Shu, (2005). The dependence between successive days is demonstrated in Fig. 13 where the transformed data from Lund are plotted.

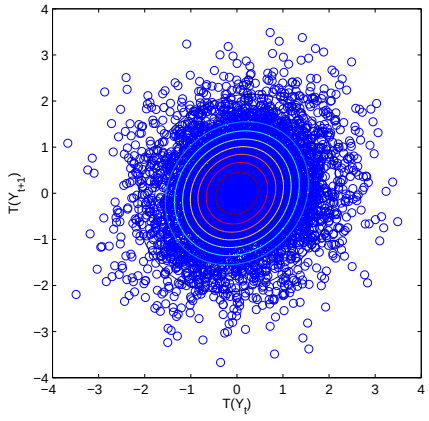

FIG.13: Plot of the dependence structure with the marginal distributions transformed to standard normal.

For a thorough coverage of bivariate copulas and their properties see Hutchinson and Lai (1990), Joe (1997), Nelsen (2006), and Trivedi and Zimmer (2005) who provide with a copula tutorial for practitioners. The values of the correlation coefficient ρ, estimated for each station are collected in Table 4. Notice that all the estimates of the correlation coefficient ρ are statistically significant, which makes the assumption of independence between the data points to seem unreasonable.

5.2 Marginal Distribution

Finally, to model the amount precipitation process we propose an approach that combines the fit of the distribution of excesses over a high threshold with the empirical distribution of the original data below the threshold. We commence our analysis by introducing some notation followed by some introductory remarks. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables having marginal distribution $F(x)$. Let us also denote by

$$F_u(x) = P(X \leq x | X > u),$$

for $x > u$, the conditional distribution of X given that it exceeds level u and assume that $F_u(x)$ can be modelled by means of a generalized Pareto distribution, that is

$$F_u(x) = 1 - \left(1 + \xi \left(\frac{x-u}{\sigma}\right)\right)^{-\frac{1}{\xi}}, \tag{5}$$

for some $\mu, \sigma > 0$ and ξ over the set $\{x : x > u \text{ and } 1 + \xi\frac{x-u}{\sigma} > 0\}$, and zero otherwise. Let also, $F_{\text{emp}}(x)$ denote the empirical distribution i.e.,

$$F_{\text{emp}}(x) = \frac{1}{n}\sum_{i=1}^{n}\{X_i \leq x\},$$

where $\{\cdot\}$ denotes the indicator function of an event, i.e. the $0-1$ random variable which takes value 1 if the condition between brackets is satisfied and 0 otherwise.

Finally, define the function

$$F_{\text{C}}(x; u) = F_{\text{emp}}(x \wedge u) + (1 - F_{\text{emp}}(u))F_u(x),$$

which, as can be easily checked, is a probability distribution function that will be used to model the amount precipitation process. Thus what needs to be addressed is the choice of the level u above which the excesses can be accurately modelled using a generalized Pareto distribution as well as methods for the estimation of the distribution parameters.

5.2.1 Choice of Threshold Level

Selection of a threshold level u, above which the generalized Pareto distribution assumption is appropriate is a difficult task in practice see for example, McNeil (1996), Davison and Smith (1990) and Rootzen and Tajvidi (1997). Frigessi *et al.* (2002), suggest a dynamic mixture model for the estimation of the tail distribution without having to specify a threshold in advance. Once the threshold u is fixed, the model parameters ξ and σ are estimated using maximum likelihood, although there exists a number of other alternative methods, see for instance Resnick (1997) and Crovella and Taqqu (1999) and references therein.

5.2.2 Extreme Value Analysis for Dependent Sequences

The generalized Pareto distribution is asymptotically a good model for the marginal distribution of high excesses of independent and identically distributed random variables, see Coles (2001), Leadbetter *et al.* (1983). Unfortunately, this is a property that is almost unreasonable for most of the climatological data sets since dependence in successive days is to be expected. A way of dealing with the dependence between the excesses is either to choose the level u high enough so that enough time has past between successive excesses to make them independent or to use declustering, which is probably the most widely adopted method for dealing with dependent exceedances; it corresponds to filtering the dependent observations to obtain a set of threshold excesses that are approximately independent, see Coles (2001). A simple way of determining m-clusters of extremes, after specifying a threshold u, is to define consecutive excesses of u to belong to the same m-cluster as long as they are separated by less than $m+1$ time days. It should be noted that the separation of extreme events into clusters is likely to be sensitive to the choice of u, although we do not study this effect in this work. The effect of declustering to the generalized Pareto distribution in Eq. 5 is the replacement of the parameters σ and ξ by $\sigma\theta^{-1}$ and ξ, where

θ is the so-called extremal index and is loosely defined as

$$\theta = (\text{limiting mean cluster size})^{-1}.$$

5.3 Method Application

In this subsection we apply the method described in subsection 5.2 to model the thinning of the amount of precipitation process, i.e. Y_t. To demonstrate the method we use data from the station in Lund. The rest of the stations give similar results.

As we have already seen, the data exhibit temporal dependence. The correlation coefficient ρ, using the Gaussian copula for the data from Lund was estimated to be 0.1362. The dependence in the data can also be seen in Fig. 10, where the expected number of m clusters (with more than one observation) for different values of m and $u = 15mm$ are plotted. The expected number of these m clusters, assuming the observations are independent is denoted by 'o' and are consistently less that the observed number of m clusters that is denoted by '+'. The expected number of m clusters computed assuming the observations are actually correlated ($\rho = 0.1362$) is denoted by '*' and provides with an obvious improvement to the assumption of independence. We also provide with 95% exact confidence intervals for both cases. The observed values fall inside the confidence interval constructed assuming correlated data.

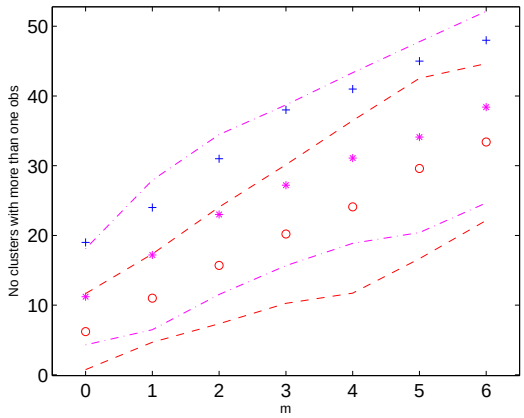

FIG. 10: Number of m-clusters with more than one observation. '+' denotes the observed and 'o' the theoretical number of m-clusters assuming that the observations are independent, while '*' denotes the number of m clusters using $\rho = 0.1362$. Line '–' denotes the 95% confidence interval for the theoretical number of m-clusters assuming independence,

while '-.' denotes the 95% confidence interval for the theoretical number of m-clusters assuming $\hat{\rho} = 0.1362$.

After the cluster size has been decided, in the case of the station in Lund $m = 0$, we turn to the problem of estimating the parameters ξ, σ and θ for the generalized Pareto model. The choice of the specific threshold ($u = 15mm$) was based on mean residual life plot. It is expected, see Coles (2001) that for the threshold u for which the generalized Pareto model provides a good approximation for the excesses above that level, the mean residual life plot i.e. the locus of the points

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (Y_{t(i)} - u) \right) : u < Y_t^{\max} \right\},$$

where $Y_{t(1)}, \ldots, Y_{t(n_u)}$ are the n_u observations that exceed u and $Y_t^{\max}$ is the largest observation of the process Y_t, should be approximately linear in u. Fig. 11 shows the mean residual life plot with approximate 95% confidence interval for the daily precipitation in Lund. The graph appears to curve from $u = 0$ until $u = 15$ and is approximately linear after that threshold. It is tempting to conclude that there is no stability until $u = 28$ after which there is approximate linearity which suggests $u = 28$. However, such threshold gives very few excesses for any meaningful inference (33 observations out of 16000). So we decided to work initially with the threshold set at $u = 15$.

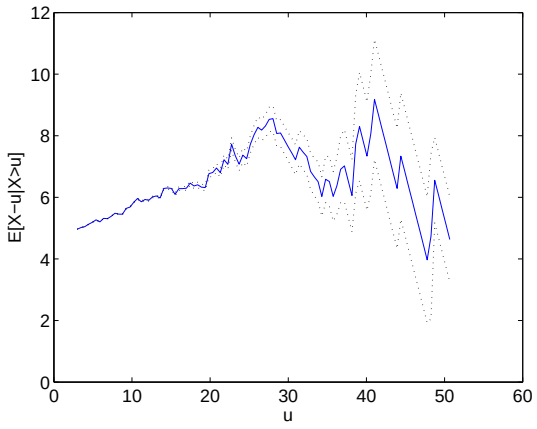

FIG. 11: Mean residual life plot of amount precipitation process from Lund, dotted lines give the 95% confidence interval.

Finally, the different diagnostic plots for the fit of the Generalized Pareto distribution are collected in Fig. 12. The data from the rest of the stations have produced similar plots

none of which gave any reason for concern about the quality of the fitted models. The parameters of the generalized Pareto model for the data from all the stations together with 95% confidence intervals are collected in Table 4. For three different stations, (i.e. Bolmen, Borås, and Hapamanda), the estimates of the shape parameter, ξ, are negative.

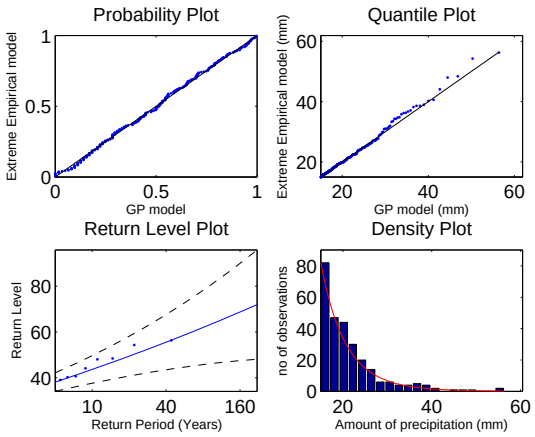

FIG. 12: Diagnostic plots for threshold excess model fitted to daily precipitation data from the station in Lund.

Table 5 shows θ for different values of m-clusters and threshold $u = 15$ for the data from Lund.

6 Evaluation

To verify the validity of the model, we have obtained distribution functions of the different precipitation indices as stipulated by the Expert Team and its predecessor, the CCl/CLIVAR Working Group (WG) on Climate Change Detection, see Peterson *et al.* (2001) and Karl *et al.* (1999). Sixteen of those indices are of relevance to this work, two regarding only the occurrence of precipitation process (CDD and CWD), another two regarding only the amount precipitation process (SDII and Prec90p) and the remaining twelve concerning both processes, see Table 6. Using the chain dependent model, we have obtained the distribution of each index based on $100,000$ simulations. This has been compared to the empirical distribution ('.-' line in Figs. 14 - 18). The agreement between the two distributions is more than satisfactory. Moreover, the empirical distribution falls always inside the 90% exact confidence intervals. The results have been presented for the weather station in Lund. The rest of the stations give similar results.

Station	$\hat{\sigma}$	CI for $\hat{\sigma}$	$\hat{\xi}$	CI for $\hat{\xi}$	$\hat{\theta}$	u (mm)	$\hat{\rho}$
Lund	5.91	(4.93, 7.03)	0.076	(-0.041, 0.236)	0.935	15	0.1362
Bolmen	6.44	(5.56, 7.41)	-0.0002	(-0.095, 0.116)	0.921	15	0.2008
Hanö	5.29	(3.044, 8.737)	0.458	(0.115, 1.05)	0.977	25	0.1649
Borås	7.63	(7.01, 8.28)	-0.011	(-0.067,0.053)	0.794	10	0.1982
Varberg	5.48	(4.687, 6.378)	0.106	(0.001, 0.236)	0.926	15	0.1206
Ungsberg	5.768	(4.622,7.115)	0.245	(0.089,0.445)	0.925	15	0.1843
Säffle	6.62	(5.96,7.329)	0.099	(0.027,0.183)	0.857	10	0.1809
Söderköping	6.259	(4.32,8.884)	0.297	(0.1, 0.649)	0.984	25	0.1678
Stockholm	5.597	(4.827,6.453)	0.135	(0.033,0.259)	0.903	10	0.1523
Malung	6.355	(5.676,7.095)	0.08	(0.004,0.17)	0.86	10	0.2280
Vattholma	4.964	(3.521,6.784)	0.334	(0.098,0.667)	0.984	20	0.1709
Myskelåsen	6.854	(5.962,7.844)	0.019	(-0.072,0.13)	0.849	10	0.2311
Härnösand	7.863	(7.053, 8.742)	0.087	(0.011, 0.175)	0.832	10	0.2068
Rösta	6.276	(5.453,7.19)	0.032	(-0.062, 0.145)	0.876	10	0.2116
Piteå	5.937	(4.429, 7.822)	0.19	(0.004, 0.456)	0.96	20	0.2010
Stensele	7.66	(6.098, 9.5)	0.041	(-0.11, 0.236)	0.915	15	0.2249
Haparanda	5.628	(4.405, 7.07)	-0.073	(-0.196, 0.125)	0.984	18	0.1871
Kvikkjokk	5.66	(5.01, 6.36)	0.04	(-0.04, 0.137)	0.864	10	0.2526
Pajala	5.033	(3.705, 6.728)	0.356	(0.153, 0.646)	0.966	18	0.2385
Karesuando	5.303	(4.117, 6.754)	0.12	(-0.037, 0.34)	0.922	15	0.2206

Table 4: Extremal parameters and their 95% confidence intervals for each weather station.

m	$\hat{\theta}$
0	0.9144
1	0.8836
2	0.8425
3	0.8322

Table 5: Values of the parameter θ for different choices of m clusters.

Index	Description	Formula
R10mm	Heavy precipitation days	$\sum 1_{\{Z_i>10\}}$
R20mm	Very heavy precipitation days	$\sum 1_{\{Z_i>20\}}$
RX1day	Highest 1 day precipitation amount	$\max_i Z_i$
RX5day	Highest 5 day precipitation amount	$\max_i \sum_{j=0}^{4} Z_{i+j}$
CDD	Max number of consecutive dry days	$\max\{j : \tau_j(X^i) = \mathbf{0}\}$
CWD	Max number of consecutive wet days	$\max\{j : w = \tau_j(X^i), w_k > 0, \forall k\}$
R75p	Moderate wet days	$\sum 1_{\{Z_i>q_{0.75}\}}$
R90p	Above moderate wet days	$\sum 1_{\{Z_i>q_{0.90}\}}$
R95p	Very wet days	$\sum 1_{\{Z_i>q_{0.95}\}}$
R95p	Extremely wet days	$\sum 1_{\{Z_i>q_{0.99}\}}$
R75pTOT	Precipitation fraction due to R75p	$\sum Z_i 1_{\{Z_i>q_{0.75}\}} / \sum Z_i$
R90pTOT	Precipitation fraction due to R90p	$\sum Z_i 1_{\{Z_i>q_{0.90}\}} / \sum Z_i$
R95pTOT	Precipitation fraction due to R95p	$\sum Z_i 1_{\{Z_i>q_{0.95}\}} / \sum Z_i$
R99pTOT	Precipitation fraction due to R99p	$\sum Z_i 1_{\{Z_i>q_{0.99}\}} / \sum Z_i$
SDII	Simple daily intensity index	$\sum Y_i / \sum 1_{\{Y_i>0\}}$
Prec90p	90%-quant. of thinned amount of precipitation	$F_Y^{-1}(0.9)$

Table 6: Weather Indices and their mathematical expressions. The quantiles $q_{(\cdot)}$ have been estimated using the observed data.

As we can see, Fig. 14 (top left), approximately during two years we expect to have about 17 days with precipitation more than $10mm$ and, Fig. 14 (top right), about 3 days with precipitation more than $20mm$. But then, see Fig. 14 (bottom left), the precipitation during each one of these three days will be quite a lot more than $20mm$. Fig. 14 (bottom right) tell us that the probability of having 5 consecutive days of really heavy precipitation in Lund is quite high.

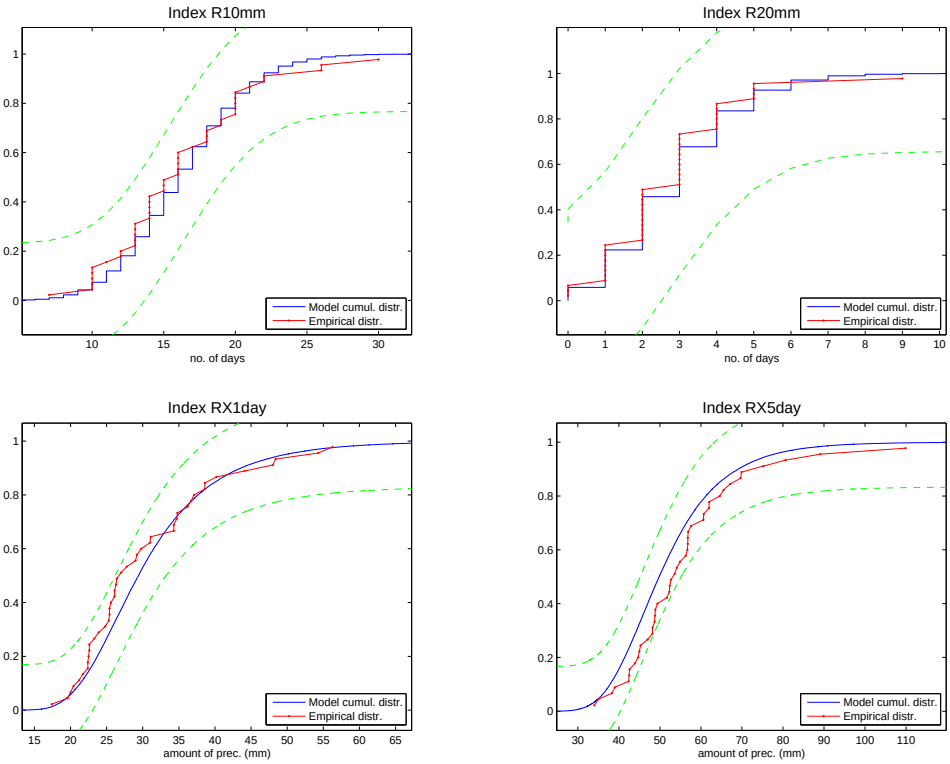

FIG. 14: Plots of R10mm (top left), R20mm (top right), RX1day (bottom left) and RX5day (bottom right). theoretical distribution '-' and empirical distribution '.-'.

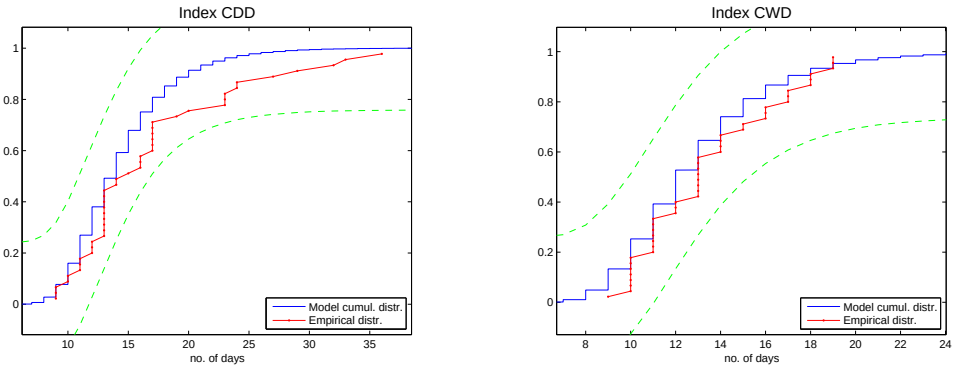

FIG. 15: Plot of maximum number of consecutive dry days (left), and maximum number of consecutive wet days (right).

As we notice in Fig. 15 (left), once every two years we should expect to have a dry spell with length more than two weeks, and a wet spell of approximately 12 days.

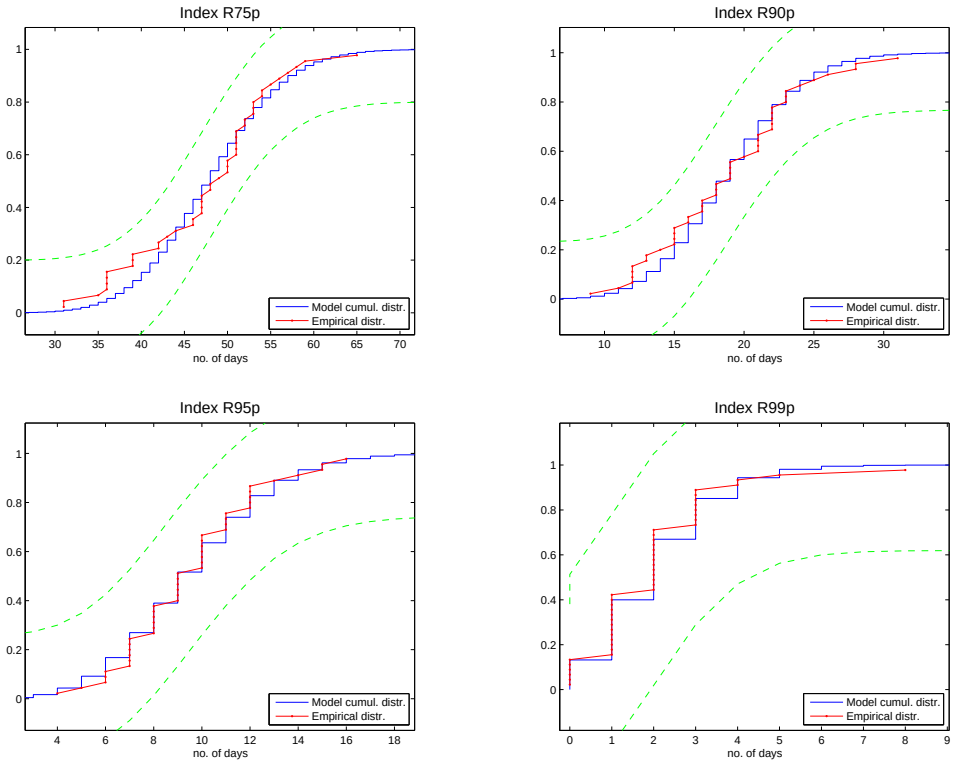

FIG. 16: Plot of the probability of number of moderate wet days (top left), above moderate wet days (top right), very wet days (bottom left) and extremely wet days (bottom right).

In Fig. 16 (top left), we see that every two years in Lund, we expect to have almost fifty moderately wet days (top right), almost 18 above moderately wet days (top right), almost 8 very wet days (bottom left) and almost 2 extremely wet days (bottom right).

In Fig. 17 (top left), we see that during the fifty moderately wet days that we expect over a period of two years in Lund we will have about 70% of the total amount of precipitation. Similarly, during the 18 above moderate wet days we expect on average a little more than 40% of the total precipitation amount (top right), for the 8 very wet days about 25% of the total amount (bottom left) and for the 2 extremely wet days about 10% (bottom right) of the total amount.

In Fig. 18 (left), we see that the average amount of precipitation per day of precipitation is 3.5mm and also every year on average only 1 out of the 10 precipitation days the downfall exceeds 9.5mm.

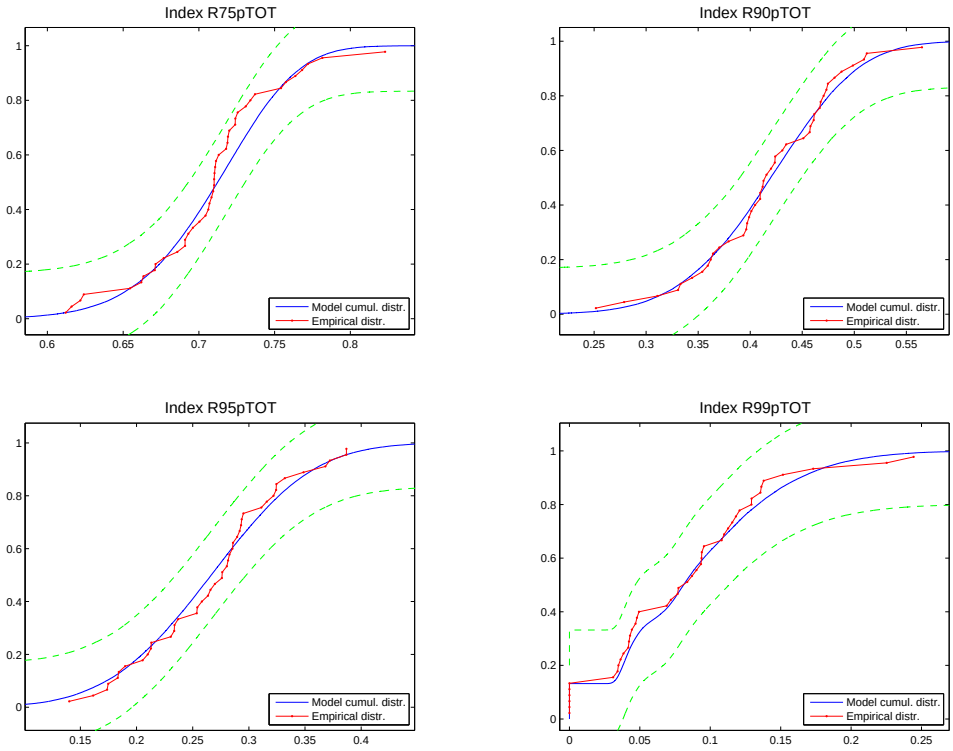

FIG. 17: Percentage of precipitation during the moderately wet days (top left), the above moderate wet days (top right), the very wet days (bottom left) and the extremely wet days (bottom right).

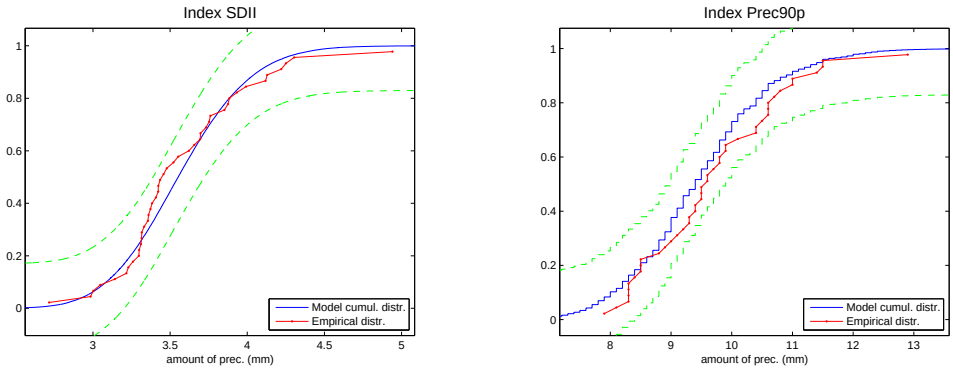

FIG. 18: Plot of the average amount of precipitation per day of precipitation (left) and the 90% quantile of the amount of precipitation of the thinned precipitation process (right).

7 Conclusions

In this paper, we have modelled the temporal variability of the precipitation in Sweden. The different weather stations have been assumed as not having any spatial dependence.

It is among our future research plans to try to model also the spatial variability of the precipitation in the different weather stations in Sweden. Some interesting conclusions can be drawn.

We have used a chain dependent model for the precipitation. That consists of a component for the occurrence of precipitation and a component for the amount of precipitation. For the first component, we have used high order Markov chains with two states. We have shown that the 1-Markov chain model that has been used extensively, is an inadequate model for most of the Swedish stations. For example, when the distribution of the long dry spell is of interest, the 1-Markov chains underestimates the length of the long dry spell in some cases up to half a day.

For the amount of precipitation process, we have used a copula to describe the temporal dependence structure between successive days, which in reality is a Gaussian process with transformed marginals. Then, the cumulative distribution has been modelled in two steps. First using the empirical distribution for the amounts of precipitation that are less than a given threshold and, then using a generalised Pareto distribution to model the excesses above the threshold. Such models have the advantage that they provide with the mathematical platform that allows computation of such quantities as return periods.

Finally, the distributions of different weather indices have been computed using Monte Carlo Markov Chain techniques, and been compared to the empirical distributions obtained from the data. The agreement between the two distributions has been really good, which supports the choice of the models.

References

[1] Akaike, H. (1974). A new look at statistical model identification, *IEEE Trans. Auto. Contol, AC*, **19**, pp.716-722

[2] Benjamin, J.R. and Cornell, C.A., (1970). *Probability, Statistics and Decision for Civil Engineers*, McGraw-Hill, Inc., New York, 685 pp.

[3] Bruhn, J.A., Fry, W.E. and Fick, G.W., (1980). Simulation of daily weather data using theoretical probability distributions. *J. Appl. Meteorol.* **19**, pp. 1029-1036.

[4] Castellvi, F. and Stockle, C.O., (2001). Comparing a locally-calibrated versus a generalised temperature weather generation. *Trans. ASAE* **44** 5, pp. 1143-1148.

[5] Chin, E. H., (1977). Modelling daily precipitation process with Markov chain, *Wat. Resources Res.*, **13**, 949-956.

[6] Coles, S., (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

[7] Cox, D.R. and Isham, V., (1988). A simple spatial-temporal model for rainfall (with discussion). *Proc. R. Soc. Lond. A*, **415**, 317-328.

[8] Cox, D.R. and Isham, V., (1994). Stochastic models of precipitation. In *Statistics for the Environment 2: Water Related Issues* (eds V. Barnett and K.F. Turkman), ch. 1, pp. 3-18. Chichester: Wiley.

[9] Crovella, M. and Taqqu, M., (1999). Estimating the heavy tail index from scaling properties, *Methodology and Computing in Applied Probability* **1**, 55-79.

[10] Dalevi, D., Pubhashi, D. and Hermansson, M., (2006). A New Order Estimator for Fixed and Variable Length Markov Models with Applications to DNA Sequence Similarity, *Stat. Appl. Genet. Mol. Biol.*, **5**, Article 8.

[11] Davison, A.C. and Smith, R.L., (1990). Models for exceedances over high thresholds, *J. Roy. Statist. Soc. B*, **52**, pp. 393-442.

[12] Frigessi, A., Haug, O., Rue, H., (2002). A Dynamic Mixture Model for Unsupervised Tail Estimation without Threshold Selection, *Extremes*, **5**, pp.219-235.

[13] Gabriel, K.R. and Neumann, J., (1962). A Markov chain model for daily rainfall occurrences at Tel Aviv. *Quart.J.royal Met.Soc.* **88**, 90-95.

[14] Geng, S., Frits, W.T., de Vries, P. and Supit, I., (1986). A simple method for generating daily raifall data. *Agric. For. Meteorol.* **36**, pp. 363-376.

[15] Guttorp, P. (1995). *Stochastic Modelling of Scientific Data*,Chapman & Hall, London Chapter 2

[16] Hutchinson, M.F., (1995). Stochastic space-time weather models from ground-based data. *Agric. For. Meteorol.*, **73**, 237-264.

[17] Hutchinson, T.P. and Lai, C.D., (1990). *Continuous Bivariate Distributions, Emphasising Applications*. Sydney, Australia: Rumsby.

[18] Jimoh, O.D. and Webster, P., (1996). the optimum order of a Markov chain model for daily rainfall in Nigeria. *Journal of Hydrology*, 185, 45-69.

[19] Joe, H., (1997). *Multivariate Models and Dependence Concepts*. London: Chapman

& Hall

[20] Karl, T.R., Nicholls, N. and Ghazi, A., (1999). CLIVAR/GCOS/WMO workshop on indices and indicators for climate extremes: Workshop summary, *Climatic Change.* Vol. 32, pp. 3-7.

[21] Lana, X. and Burgueno, A., (1998). Daily dry-wet behaviour in Catalonia (NE Spain) from the viewpoint of Markov chains, *Int. J. Climatol.* **18**, 793-815.

[22] Leadbetter, M.R., Lindgren, G., and Rootzen, H., (1983). *Extremes and Related Properties of Random Sequences and Series.* Springer Verlag, New York.

[23] LeCam, L., (1961). A stochastic description of precipitation *Proc.4th Berkeley Symp.*, pp.165-186.

[24] Lennartsson, J., and Shu, M., (2005). *Copula Dependence Structure on Real Stock Markets*, Masters thesis, Chalmers University of Technology, 2005-01.

[25] Liao, Y., Zhang, Q. and Chen, D., (2004). Stochastic modeling of daily precipitation in China. *Journal of Geographical Sciences,* 14(4), 417-426.

[26] Mellor, D., (1996). The modified turning bands (mtb) model for space-time rainfall:i, model definition and properties. *J. Hydrol.*, **175** 113-127.

[27] McNeil, A.J., (1996). Estimating the tails of loss severity distributions using extreme value theory, *Technical report*, Department Mathematik, ETH Zentrum, Zurich.

[28] Nelsen, R. B., (2006). *An Introduction to Copulas* 2nd edition. New York: Springer.

[29] Norris, J.R., (2005). *Markov chains*, Cambridge University Press.

[30] Peterson, T.C., Folland, C., Gruza, G., Hogg, W., Mokssit, A. and Plummer, N., (2001). Report on the Activities of the Working Group on Climate Change Detection and Related Rapporteurs 1998-2001. *World Meteorological Organisation,* WCDMP-47, WMO-TD 1071.

[31] Racsko, P., Szeidl, L. and Semenov, M., (1991). A serial approach to local stochastic weather models. *Ecol. Model.* **57**, pp. 27-41.

[32] Resnick, S.I., (1997). Heavy tail modeling and teletraffic data, *The Annals of Statistics* **25**5, 1805-1869.

[33] Rootzen, H. and Tajvidi, N., (1997). Extreme value statistics and wind storm losses: A case study, *Scandinavian Actuarial Journal* **1**, 70-94.

[34] Rodríguez-Iturbe, I., Cox, D. and Isham, V., (1987). Some models for rainfall based

on stochastic point processes. *Proc. R. Soc. Lond., A* **410**, 269-288.

[35] Rodríguez-Iturbe, I., Cox, D. and Isham, V., (1988). A point process model for rainfall: further developments. *Proc. R. Soc. Lond., A* **417**, 283-298.

[36] Schwarz, G., (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, pp. 461-464.

[37] Selker, J.S. and Haith, D.A., (1990). Development and testing of simple parameter precipitation distributions. *Water Resour. Res.* **26** 11, pp. 2733-2740.

[38] Smith, R. L. and Robinson, P.J., (1997). A Bayesian approach to the modelling of spatial-temporal precipitation data. *Lect. Notes Statist.*, 237-269.

[39] Srikanthan, R. and McMAhon, T.A., (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrol. Earth Syst. Sci.* **5** 4, pp. 653-670.

[40] Stern, R.D. and Coe, R. (1984)., A Model fitting Analysis of Daily Rainfall Data, *J.R.Statist.Soc. A*, **147**, Part1, pp.1-34.

[41] Stidd, C.K., (1973). Estimating the precipitation climate. *Wat. Resour. Res.*, **9** 1235-1241.

[42] Trivedi, P. K. and Zimmer, D.M., (2007). Copula Modelling: An Introduction for Practitioners, *Foundations and Trends in Econometrics*, Vol. 1, No 1, 1-111.

[43] Waymire, E., Gupta, V. K., (1981). The mathematical structure of rainfall representations: 3,, Some applications of the point process theory to rainfall processes. *Wat. Resour. Res.***17**, 1287-1294.

[44] Waymire, E., Gupta, V. K. and Rodríguez-Iturbe, I., (1984). Spectral theory of rainfall intensity at the meso-β scale. *Wat. Resour. Res.***20**, 1453-1465.

[45] Woolhiser, D.A., (1992). Modelling daily precipitation-progress and problems. In: *A. Walden and P. Guttorp (Editors), Statistics in the Environmental and Earth Sciences.* Edward Arnold, London, pp.71-89.

8 Appendix

8.1 Review of Mathematical Order Estimators

Let X_t denote a k-Markov chain that is defined on a state space S and x_1^n its realisation. Let also $P_{\mathrm{ML}(k)}(x_1^n)$ be the kth order maximum likelihood, i.e.

$$P_{\mathrm{ML}(k)}(x_1^n) = \max P(X_1^k)\Pi_{i=k+1}^n P(X_i = x_i|\tau_k(X^{i-1}) = \tau_k(x^{i-1})).$$

Tong (1975) reported that the Akaike Information Criterion (AIC) order estimator, could be used as an objective technique for determining the optimum order k of the chain, see also Akaike [**?**]. The optimium order k is the order that has the minimum loss function:

$$\hat{k}_{\mathrm{AIC}}(x_1^n) = \mathrm{argmin}_k(-\log\ P_{\mathrm{ML}(k)}(x_1^n) + |S|^k).$$

Schwartz (1978) presented an alternative technique the Bayesian Information Criterion (BIC) order estimator whose consistency was established under general conditions [**?**] was only recently established. The optimum order, k is the order that minimises the loss function which now is given by:

$$\hat{k}_{\mathrm{BIC}}(x_1^n) = \mathrm{argmin}_k(-\log\ P_{\mathrm{ML}}(x_1^n) + \frac{|S|^k(|S|-1)}{2}\log(n)).$$

Dalevi *et al.* (?) showed using experimental results that the BIC order estimator tends to under-estimate the order as k gets larger for moderate data sizes.

Finally, the Maximal Fluctuation Criterion (MFC) contrary to the AIC and BIC order estimators, was specifically designed for multiple step Markov chains. Let for any realisation $x \in S^n$ of the k-Markov chain, $N_x(w) = |\{i \in [1,n] : \tau_l(x^i) = w, w \in S^l\}|$ denote the number of times w occurs in x. The *Peres-Shields Fluctuation* function is defined as

$$\Delta_k(v) = \max_{s\in S}|N_x(vs) - \frac{N_x(\tau_k(v)s)}{N_x(\tau_k(v))}N_x(v)|.$$

When the order of the Markov chain is k or less, this fluctuation is small. Therefore, the Maximal Fluctuation Criterion (MFC) order estimator is defined as

$$\hat{k}_{\mathrm{MFC}}(x_1^n) = \min\{k \geq 0 : \max_{k<|v|<\log\log(n)} \Delta_k(v) < n^{3/4}\}.$$

In practice the function $\log\log(\cdot)$ is substituted by any function that grows slower than $\log(\cdot)$. Dalevi *et al.* (?) suggested the Generalized Maximum Fluctuation Criterion

(GMFC) order estimator, which is closely related to the Maximal Fluctuation Criterion (MFC) order estimator,

$$\hat{k}_{\mathrm{GMFC}}(x_1^n) = \mathrm{argmax}_k \frac{\max_{k-1<|v|<f(n)} \Delta_{k-1}(v)}{\max_{k<|v|<f(n)} \Delta_k(v)},$$

where $f(n)$ is any function that satisfies the same conditions as for the GMF order estimator.