PLISKA
STUDIA MATHEMATICA BULGARICA

ПЛИСКА
БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **14** (2003), 91–106

PLISKA
STUDIA MATHEMATICA
BULGARICA

DETECTING PRECIPITATION CLIMATE CHANGES: AN APPROACH BASED ON A STOCHASTIC DAILY PRECIPITATION MODEL

N. Neykov, P. Neytchev and W. Zucchini

Abstract. We consider development of daily precipitation models based on [3] for some sites in Bulgaria. The precipitation process is modelled as a two-state first-order nonstationary Markov model. Both the probability of rainfall occurrance and the rainfall intensity are allowed depend on the intensity on the preceeding day. To investigate the existence of long-term trend and of changes in the pattern of seasonal variation we use a synthesis of the methodology presented in [3] and the idea behind the classical running windows technique for data smoothing. The resulting time series of model parameters are used to quantify changes in the precipitation process over the territory of Bulgaria.

1. Introduction

We consider development of daily precipitation models for some sites in Bulgaria. The precipitation process is described as a two-state first-order Markov chain which has been found to be an adequate model in many different regions, e.g. [2], [6], [9], [10] and [11]. Finite Fourier series are used to approximate the seasonal cycle in the probability of rainfall occurrence and in the parameters of the intensity (amount when it rains) distribution. The resulting generalized linear model (glm), see [5], can be fitted using standard software. A good overview

2002 *Mathematics Subject Classification:* 62M10

Key words: Binary time series, climate change, gamma time series, generalized linear models, Markov chain, rainfall modeling

concerning daily precipitation modelling techniques is given in [10]. Following that methodology [11] developed a stochastic model to describe the process of daily precipitation for 133 sites in southwest Bulgaria for which suitable records with lengths at least 20 years were available. Their work provided the opportunity to assess the performance of the model for a large number of sites and, in particular, to check that the estimates were plausible when considered spatially. For example one would expect the estimates at neighboring sites to be similar unless the local topography varies substantially. A second advantage of having estimates on a dense network of sites is that selected properties of the process can be interpolated and mapped.

The methodology developed in [9] is designed to describe the daily rainfall process under the assumption that there have been no changes in the process, i.e that there are no long-term trend and that the seasonal pattern remains the same in each year. Furthermore their analysis is based on that the probability of rain occurring on a given day depends only whether it was wet or dry on the preceeding day. However, it is plausible that this probability also depends on how much it rained on the previous day. The dependency on the previous intensities simultaneously in the occurrence and intensity models was studied for the first time in [3]. Moreover, apart from seasonal– and temporal–dependence effects, some slowly-varying trend function (linear spline with unknown knots) over the years were considered in [1] and [3] using glms, which are able to accommodate the high variability present in the data.

In order to analyse not only the long-term changes (slowly-varying trend, temporal variation) but also the changes of the seasonal variation pattern [12] propose a synthesis between the methodology presented in [3] and the basic idea behind the classical running windows technique for data smoothing, that is to estimate the parameters of the model for each year using only the data from a symmetric "window" of neighbouring years. (The neighborhoods are necessarily asymmetric at the start and at the end of the rainfall record.) In this way the model is fitted for each year separately using only the $365m$ observations from it's corresponding window of length m years. Thus, for example, the data used to model the rainfall occurrances for a given year data comprise $365m$ Bernoulli (wet day or dry day)observations; the covariates are day of the year and the rainfall intensities on the previous day. Logistic regression is used to fit these Bernoulli (binary) data.

The aim of the paper is to demonstrate the statistical technique developed in [11] for the detection of precipitation climate changes over the territory of Bulgaria and to quantify these changes.

2. Description of the daily precipitation model

The daily rainfall totals can be considered as a time series of amounts, where the amount can also be zero. In our application a day was defined to be dry if less than 0.1mm was recorded. Let Y_t be a nonnegative random variable denoting the amount on day t, $t = 1, \ldots, n$ and let y_t be its observed value. The stochastic process Y_t is referred to as the *amount process*. The distribution of Y_t is assumed to be a mixture of a discrete component ($Y_t = 0$) and a continuous component ($Y_t > 0$). Let $f_t(y|X_t = x_t)$ denote the transition mixed density of Y_t, where X_t denotes a vector of covariates including Y_{t-1}, and possibly other explanatory variables. It is convenient to express the density of Y_t in an explicit form by the so-called *occurrence* and *intensity* processes (see, e.g., [6] and [9]).

The *occurrence* process is defined as

$$J_t = \begin{cases} 0, & \text{if } \mathrm{Y_t} = 0 \\ 1, & \text{if } \mathrm{Y_t} > 0, \quad \text{where} \quad \text{t=1, \ldots,n.} \end{cases}$$

Denote by $\pi_w^t = \Pr(J_t = 1)$ and $\pi_t(x_t) = \Pr(J_t = 1|X_t = x_t)$ the unconditional and conditional Bernoulli distribution of the process J_t, where $X_t = (J_{t-1}, \ldots, J_{t-p}, Y_{t-1}, \ldots, Y_{t-p}, X_{1t}, \ldots, X_{k-p,t})^T$ is a vector of possible covariates. Interactions between the components of X_t may be included in the standard way. $X_t = (J_{t-1}, Y_{t-1}, \ldots, Y_{t-p}, X_{1t}, \ldots, X_{k-p,t})^T$, then

$$\Pr(J_t = j_t|X_t = x_t) = \begin{cases} 1 - \pi_t(j_{t-1}, y_{t-1}, \ldots, y_{t-p}, x_{1t}, \ldots, x_{kt}), & \text{if } \mathrm{j_t} = 0 \\ \pi_t(j_{t-1}, y_{t-1}, \ldots, y_{t-p}, x_{1t}, \ldots, x_{kt}), & \text{if } \mathrm{j_t} = 1. \end{cases}$$

We note that J_t is not strictly Markovian since we allow dependence of J_t on x_t, not just on j_{t-1}.

The *intensity* process is defined to be $W_t = Y_t$ when $Y_t > 0$ and *missing* otherwise. The intensity process W_t has a positively skewed continuous conditional distribution with density $q_t(w|X_t)$ for $w > 0$ and 0 otherwise. Common assumptions for the form of $q_t(w|X_t)$ are exponential, lognormal, Weibull or gamma.

The transition distribution of precipitation amount Y_t under the above assumptions is a mixture of the occurrence and intensity distribution

$$f_t(y|X_t = x_t) = (1 - \pi_t(x_t))\,\delta_0(y) + \pi_t(x_t) q_t(y|x_t),$$

where $\delta_0(y)$ is the Dirac delta function with zero support. Inference about Y_t can be done provided $\pi_t(x_t)$ and $q_t(w|x_t)$ are known.

In what follows we will restrict our attention to the special case in which the covariate vector is $X_t = (J_{t-1}, Y_{t-1})^T$.

2.1. Models for occurrence

Let $\pi^t_{w|d}(x_t) = \pi_t(j_{t-1} = 0) = Pr(J_t = 1|j_{t-1} = 0)$ and $\pi^t_{w|w}(x_t) = \pi_t(j_{t-1} = 1, Y_{t-1} = y_{t-1}) = Pr(J_t = 1|j_{t-1} = 1, Y_{t-1} = y_{t-1})$ be the two-state nonstationary Markov chain transition probabilities that a wet day follows a dry day, and a wet day follows a wet day respectively, conditional on the amount on previous day. Given the previous state of the occurrence process and using arguments based on the total probability we can express the unconditional probability for a wet state on day t as

$$\pi^t_w = \pi^{t-1}_w \pi^t_{w|w}(x_t) + (1 - \pi^{t-1}_w)\pi^t_{w|d}(x_t).$$

Under the plausible assumption that $\pi^t_w \approx \pi^{t-1}_w$ for any t, the unconditional probability of wet day is given by

$$\pi^t_w = \pi^t_{w|d}(x_t)/(\pi^t_{w|d}(x_t) + 1 - \pi^t_{w|w}(x_t)).$$

In order to model the transition probabilities $\pi^t_{w|w}(x_t)$ and $\pi^t_{w|d}(x_t)$, [9] and many others, have used the logit link function

$$\pi(x_t) = l(u(x_t)) = \exp(u(x_t))/\left(1 + \exp(u(x_t))\right).$$

The function $u(x_t)$ should be a periodic parametric function, approximately sinusoidal in shape, that links the covariates and the unknown parameters in order to account for various temporal and seasonal effects. So it should be composed of seasonal terms that repeat each year and represent a 'typical' year, and a remainder term that represents deviation from this regular pattern. We will use the following function

$$u(x_t) = \alpha_0 + \alpha_1\sqrt{y_{t-1}} + \alpha_2 \sin\left(2\pi tk/365\right) + \alpha_3 \cos\left(2\pi tk/365\right).$$

Of course some other smooth function, such as logarithm, cubic root, or power transformation could be used instead the square root.

The probabilities $\pi^t_{w|w}(x_t)$ and $\pi^t_{w|d}(x_t)$ are estimated using the same method (maximum likelihood) but with different data sets. In the case of $\pi^t_{w|w}(x_t)$, $t = 1, \ldots, n$ the state of the previous day is wet and so the data are regarded as Bernoulli observations, indexed by the day, year, state and the corresponding rainfall amount on the previous day. In the case of $\pi^t_{w|d}(x_t)$ the state of the previous day is dry and so the data are regarded as Bernoulli observations, indexed by the day, year and the state. In practice one finds that $\pi^t_{w|d}(x_t) \leq \pi^t_w \leq \pi^t_{w|w}(x_t)$, reflecting the persistent nature of daily rainfall occurrence.

2.2. The Intensity Models

The distribution of precipitation depths on wet days is positively skewed (i.e. smaller amounts occurr more frequently than larger amounts) and also exhibits seasonal variability. A simple and widely accepted technique to model this behaviour (see, e.g., [3], [9] and [11]) is to fit a single family of distributions whose parameters are allowed to vary smoothly over the year, usually by representing them as a Fourier series.

In this region the gamma distribution was found to be suitable ([11]); its probability density function given by

$$\gamma(z, \mu, \beta) = \begin{cases} \dfrac{(\beta/\mu)^{\beta} z^{\beta-1} \exp(-\beta z/\mu)}{\Gamma(\beta)} & z > 0 \\ 0 & z = 0, \end{cases}$$

where $\Gamma(\beta)$ is the gamma function, μ the mean and β the shape parameter. The following log link function was used for the intensity model

$$\log(\mu_t(x_t)) = \theta_0 + \theta_1\sqrt{y_{t-1}} + \theta_2 \sin(2\pi tk/365) + \theta_3 \cos(2\pi tk/365),$$

where $\theta_0, \ldots, \theta_3)^T$ are the unknown parameters. In place of the square root another smooth function, e.g, logarithm, a cubic root, or power transformation could be used.

2.3. The amplitude-phase interpretation

The sine-cosine representation used in the rainfall model specification is convenient for computational purposes, but for interpretating the parameters, or for comparing the parameters of different sites, the (equivalent) amplitude-phase representation is preferable. For example the phase parameters indicate the time of year of maximum probability of rain, or of maximum mean intensity; the amplitudes describe the maximum size of the seasonal change in mean intensity, or in the probability of rainfall. The intercepts represent the average rainfall intensity, or the average probability of rain over the year.

2.4. Computational aspect and model choice

As the Bernouli and Gamma distribution belong to the exponential family, estimates of $\pi^t_{w|w}(x_t)$, $\pi^t_{w|d}(x_t)$, $\mu_t(x_t)$ and β can be obtained using standard glm estimation methodology [5].

Our parameter estimates were computed by the method of fitting the expectation using a standard nonlinear least squares regression program which employs a modified Gauss-Newton algorithm in iteratively reweighted mode [4].

3. Applications

Many different aspects of the precipitation process are of interest in meteorological and hydrological applications, for example the monthly, seasonal and annual means, the distribution n-day extreme precipitation totals, the expected number of wet days, the expected length of dry spells, and so on.

An important feature of the proposed model is that it can be applied to quantify complex features of daily precipitation without special knowledge of the underlying statistical theory. Once the model has been calibrated at a given site one uses it to generate long sequences of artificial precipitation for that site. These sequences can be used to estimate any statistic, or probablity, relating to precipitation event of interest in exactly the way one would do so if a long sequence of real rainfall data were available. Furthemore, by using appropriate adjustments that are considered in the next sections, some properties of the process can be obtained approximately but directly from the model.

3.1. Seasonal adjustments

Because the transition probabilities of the amount process Y_t depend on the intensity of the previous day, the main findings and results of the well developed stationary theory of chain-dependent Markov process do not apply. One way to overcome this problem of dependence (approximately) is to replace the expressions $\pi^t_{w|w}(x_t)$ and $\mu_t(x_t)$ by their expectations with respect of the preceding day amount, namely by $\pi^t_{w|w}$ and μ_t, respectively. Alternatively we can make the approximation that the transition probabilities of occurrence $\pi^t_{w|d}$ and $\pi^t_{w|w}$, and the intensity mean μ_t, are roughly constant over a short T-day period of time, e.g. week, month or, in some cases, even season. Taking the expectation over a T-day period of time the corresponding transition probabilities $\pi_{w|w}$ and $\pi_{w|d}$, and the intensity μ of a stationary two-state first-order Markov chain can be determined using a fixed point algorithm such as the following:

$$
\begin{aligned}
\pi_{w|d} &= l\left(\bar{s}^d + \hat{\alpha}_0^d\right) \\
\pi_{w|w}^{(i)} &\approx l\left(\bar{s}^w + \hat{\alpha}_0^w + \hat{\alpha}_1^w \sqrt{\mu^{(i-1)}}\right) \\
\pi_w^{(i)} &\approx \frac{\pi_{w|d}}{\pi_{w|d} + 1 - \pi_{w|w}} \\
\mu^{(i)} &\approx \bar{\mu}^s \exp(\hat{\theta}_0)\left(1 - \pi_w + \exp(\hat{\theta}_1\sqrt{\mu^{(i-1)}})\pi_w^{(i)}\right),
\end{aligned}
$$

where,

$$\begin{aligned}\bar{s}^d &= \mathrm{E}_t\{\alpha_2^d \sin(2\pi t/365) + \alpha_3^d \cos(2\pi t/365)\}\\ \bar{s}^w &= \mathrm{E}_t\{\alpha_2^w \sin(2\pi t/365) + \alpha_3^w \cos(2\pi t/365)\}\\ \bar{\mu}^s &= \mathrm{E}_t\{\exp(\theta_2^\mu \sin(2\pi t/365) + \theta_3^\mu \cos(2\pi t/365))\}\end{aligned}$$

where the hats indicate parameter estimates of the daily rainfall model fitted using observations in the relevant (running) window; $l(.)$ is logit link function and i is the current iteration number. Starting with $\mu^{(0)}$ equal to the mean daily precipitation (approximately 2 mm) convergence is reached in about 3 iterations.

The annual period of time can be treated similarly because the expectations $\bar{s}^w$, $\bar{s}^d$ and $\bar{\mu}^s$ equal to zero over an year and thus the above technique reduces to the seasonal adjustment.

For a given T-day period of time this procedure is repeated for each year in the rainfall record; the output of the running windows technique comprises four time series of the estimates π_w, $\pi_{w|w}$, $\pi_{w|d}$ and μ which can be used for further analysis.

3.2. Monthly, seasonal, annual and extreme statistics

Having the quantities outlined in the previous subsection, the methodology concerning the monthly, seasonal, annual and maximum precipitation developed in [6] and [7] can now be applied.

Recall that the number of wet days $N(T) = \sum_{j=1}^{T} J_j$ and the total precipitation amount $S(T) = \sum_{j=1}^{N(T)} W_j$ in a T-day time period are random variables. Therefore, the expected number of rainy days, the total precipitation amount and their variances over a T-day time period are given by

$$\mathrm{E}\,[N(T)] = T\pi_w \quad \text{and} \quad \mathrm{Var}\,[N(T)] \approx T\pi_w(1-\pi_w)\,[(1+\rho)/(1-\rho)]$$

$$\mathrm{E}\,[S(T)] = T\pi_w\mu \quad \text{and} \quad \mathrm{Var}\,[S(T)] \approx T\pi_w \left(\sigma^2 + (1-\pi_w)\,[(1+\rho)/(1-\rho)]\,\mu^2\right),$$

where $\rho = Corr(J_t, J_{t+1}) = \pi_{w|w} - \pi_{w|d}$ is the first-order auto-correlation coefficient and $\sigma^2 = Var(W_t) = Var(Y_t|J_t = 1)$, see [7].

4. Assessment of the effects of climate change

In order to assess the effect of climate change on the precipitation the running windows technique is applied, i.e., the time series of estimates are produced for each gauge. The series comprise intercepts, amplitudes, phases for the model parameters and also the deviances. For the purpose of exploring this series, an appropriate graphical technique based on the R environment, was developed.

The plots in Figures 1 - 3 show the time series of the model parameter estimates at the Plovdiv gauge. They are are based on a 5-year running windows. The continuous lines shown are the lowess smoother of the data presented in the subplots.

A formal goodness-of-fit test on the fitted model can be carried out using the fact that (under the hypothesis) the relevant deviance for the linear logistic model for each running window is approximately equal to its degrees of freedom, i.e., the ratios of the deviances to their degrees of freedom are approximately equal to one.

The plots exhibit long-term trends as well as changes of the seasonal variation patterns. However, the visual inspection alone is not sufficient for identifying changes in the parameter estimates series. Thus a modification of the Mann-Whitney test [8] was used to detect the years of changes. In all the subplots the sub-intervals means determined by the change-point technique are marked. A significant level of $\alpha = 0.05\%$ was used. The results of these tests provide evidence that change-points did indeed occur in the period considered.

Figure 4 refers to May precipitation characteristics at Plovdiv. The plots are based on a 5-year running window. The top left-hand window in each case is a scatterplot of the observed rainfall amounts against the corresponding expected amount under the model. The term "relative error" in the legend is the average ratio of the observed and to expected rainfall amounts. The subplots entitled "Rainfall Probability", "Rainfall Intensity" and "Expected Amount" correspond to the estimates of the basic rainfall model elements π_w, μ and $\pi_w \mu$, respectively, and their values are marked as by small circles.

Figure 5 is analogous to Figure 4 but refers to annual precipitation. Correspondingly the headings of the plots are prefixed with the words "seasonally adjusted". The observed rainfall data amounts in the subplots "Observed and Expected Amount" are marked as small circles.

Figures 1 to 5 illustrate the nature of the changes in the rainfall process over the period of observation. Interesting (Figure 5) is the indication that the seasonaly adjusted rainfall intensity decreased in the 1930's (and possibly again in around 1990) whereas the seasonaly adjusted probability increased in the 1930's but there is a hint that it is gradually decreasing again. Until recent times these two (opposing) features compensated each other so that the seasonally adjusted expected amount was approximately constant until about 1980.

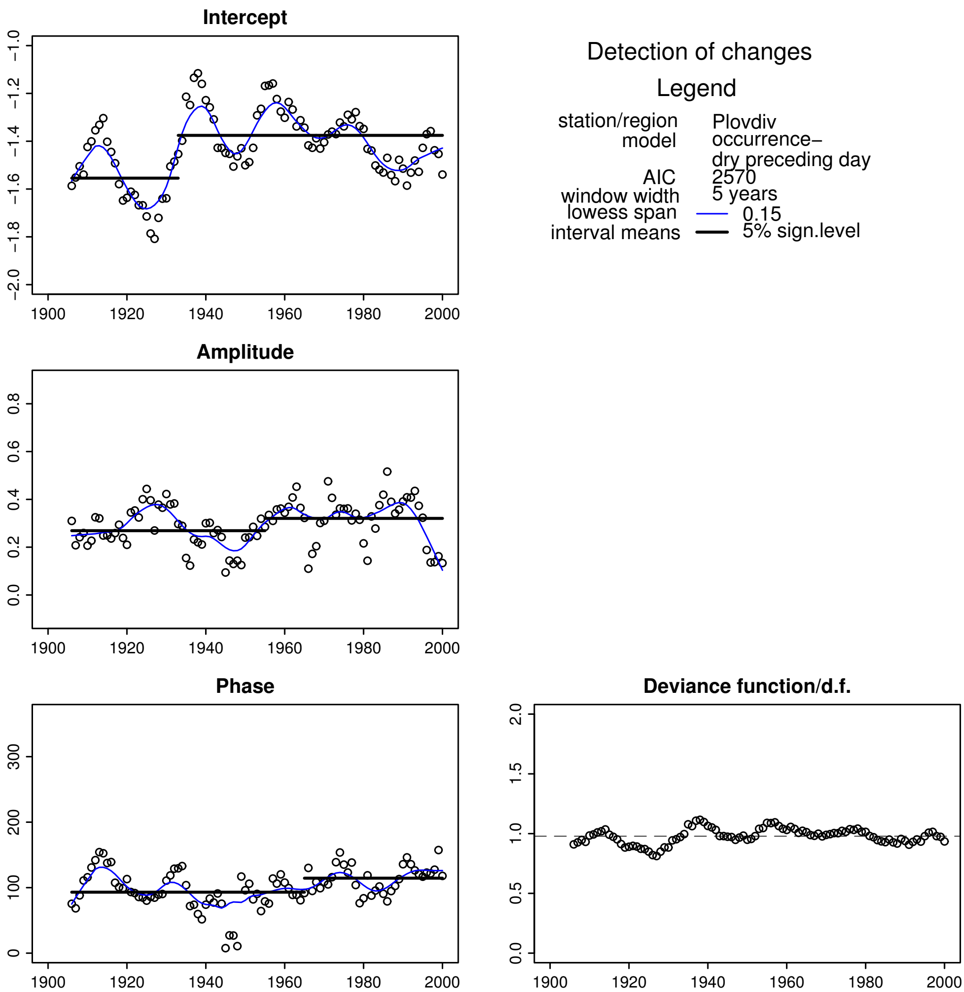

Figure 1: $\pi^t_{w|d}$ model estimates for gauge Plovdiv — 5-years window. The means of the sub-interval, determined by the change points are marked.

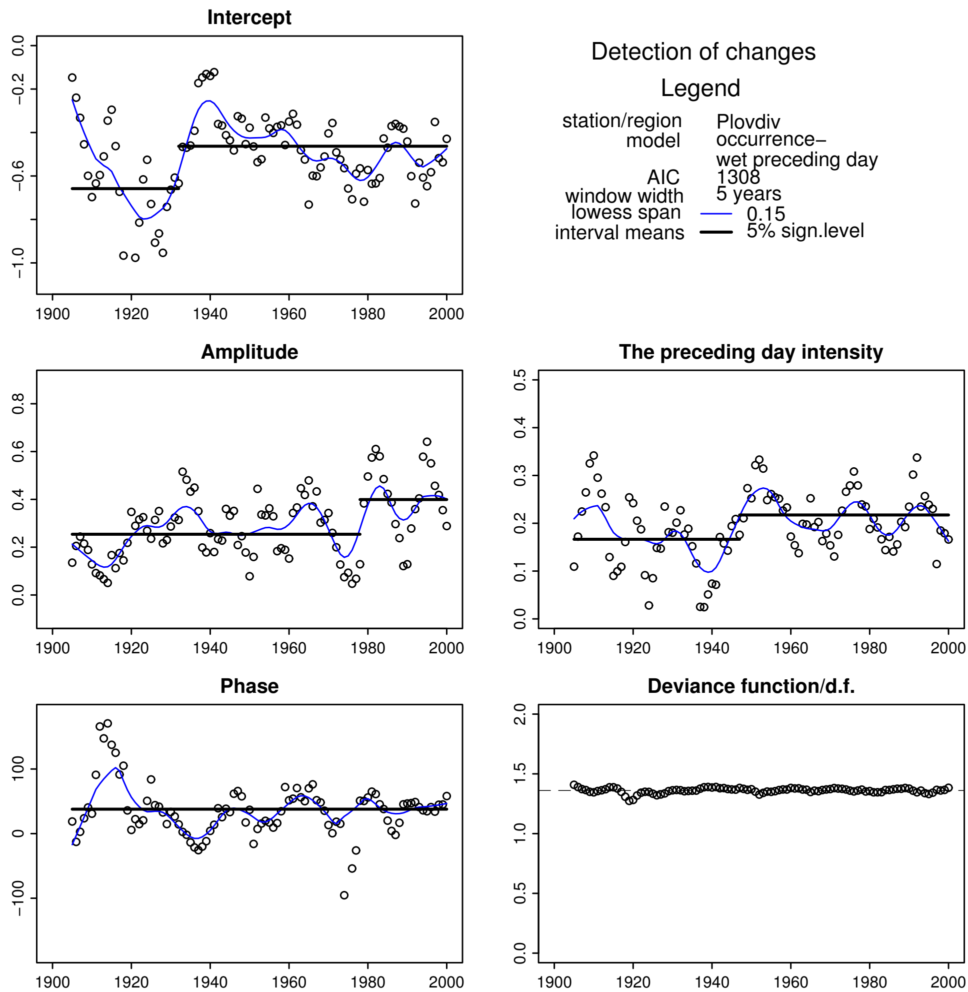

Figure 2: $\pi^t_{w|w}$ model estimates for gauge Plovdiv — 5-years window. The means of the sub-interval, determined by the change points are marked.

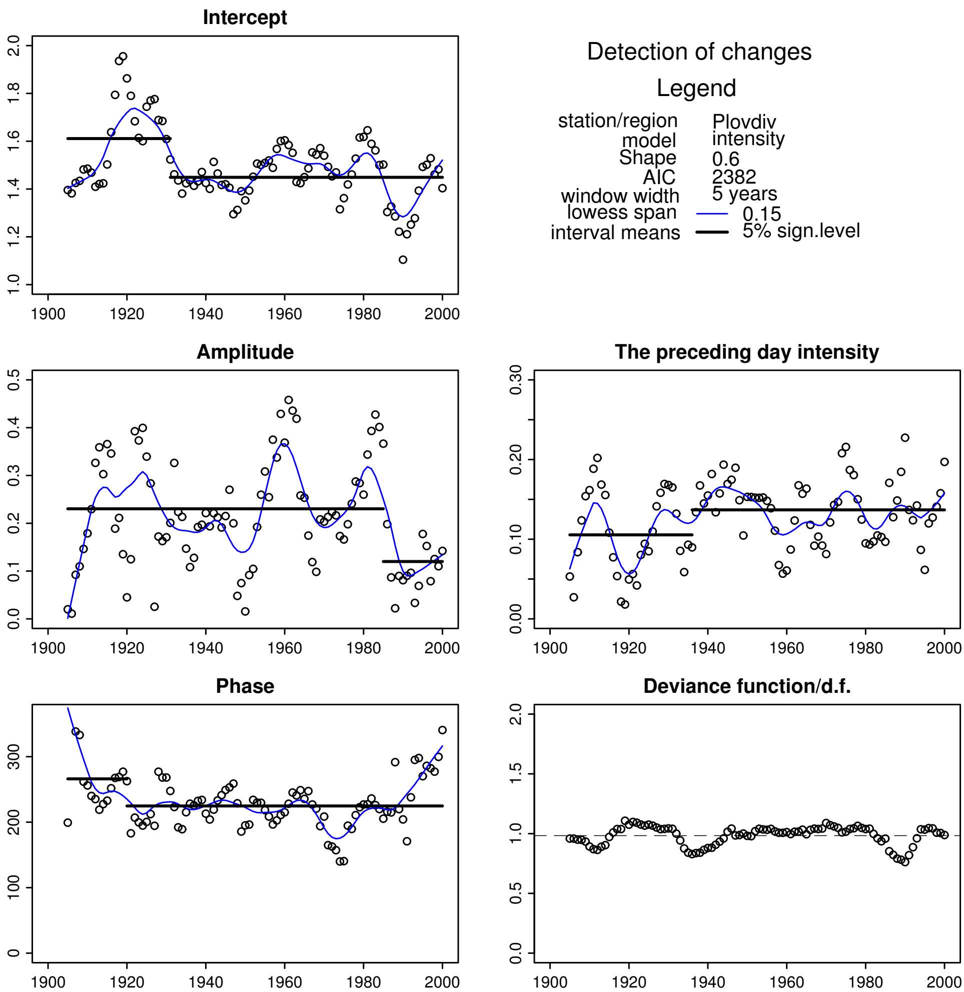

Figure 3: μ_t model estimates for gauge Plovdiv — 5-years window. The means of the sub-interval, determined by the change points are marked.

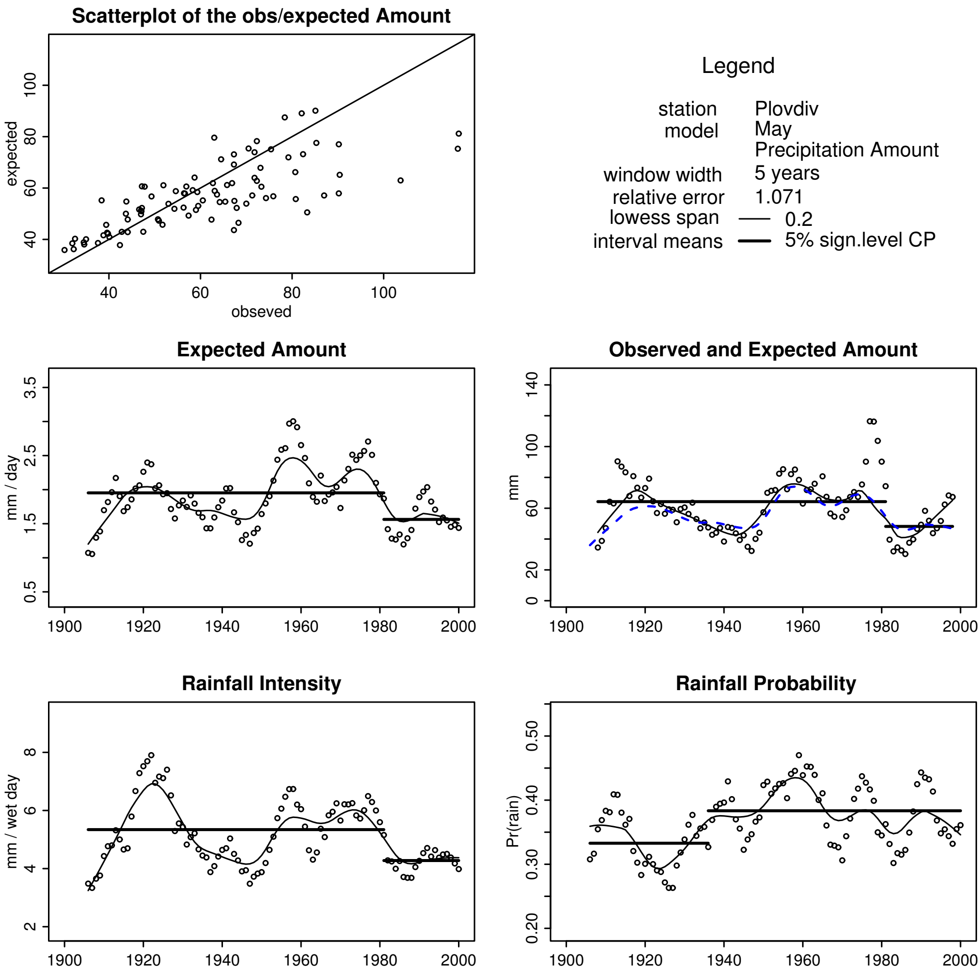

Figure 4: Monthly (May) model precipitation total for gauge Plovdiv

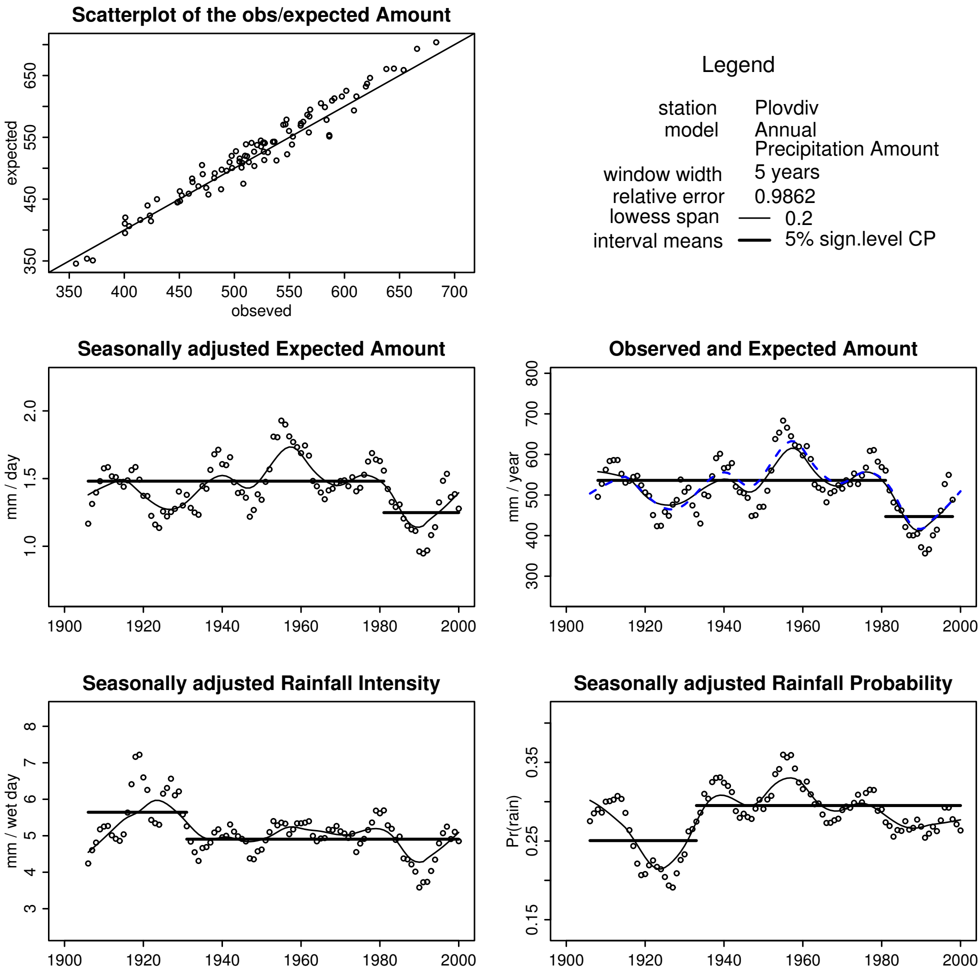

Figure 5: Annual model precipitation total for gauge Plovdiv

5. Conclusions and prospects for future research

Stochastic daily rainfall models have a long history and have been sucessfully applied in many parts of the world. The main advantage that they offer is that, via simulation, they can be used to estimate *any aspect* of the daily precipation process, no matter how complex. In most applications they have been fitted under the assumption that the process has not changed, i.e. that there were no trends and that the seasonal pattern has remained unchanged and (if one wishes to base decisions on the model) that it will remain unchanged. However it is becoming increasingly apparent that such an assumption cannot be taken for granted. Our analysis provides a case in point. The technique outlined in this paper is to fit a daily model in a running window of observations. This enables one not only to detect the existence changes but also to quantify such changes in terms of the model parameters, and hence to arbitrary properties of the process.

Of course such a model does not enable us to forecast the future behaviour of the process. One possibile way doing that would be to identify an appropriately strong relationship between the model parameters and some indicator for which long-term forecasts can be made, perhaps global temperature. Such a relationship has not yet been identified. Nevertheless for the present one can use of the model as a mechanism to generate "scenarios". Under the stationarity assumption very dry (or very wet) years are regarded as extreme observations from the model. If one relaxes the assumption of stationarity then such years can be regarded as "normal" years in a dry (or wet) period. One can use the estimates of the parameters from such periods in the historical record to assess the properties of the process under a dry (or a wet) scenario.

Acknowledgements

The work of N. Neykov and P. Neytchev on the 'Development of a daily precipitation model for south-west Bulgaria' and 'An assessment of the effects of climate change on precipitation in Bulgaria' was fully supported by Deutsche Forschungsgemeinschaft within the framework of cooperation between the Deutsche Forschungsgemeinschaft and the Bulgarian Academy of Sciences while they were Visiting Fellows of the Institute of Statistics and Econometrics, University of Göttingen during the periods May - June 2000, and 12 September - 12 December 2001.

REFERENCES

[1] R. E. Chandler, H. S. Wheater. Analysis of rainfall variability using Generalized Linear Models: A case study from the West of Ireland. *Water Resources Res.*, **38** (2002),

[2] R. Coe, R.D. Stern. Fitting models to daily rainfall data. *J. of Appl. Meteorology*, **21** (1982), 1024–1031.

[3] G. K. Grunwald, R. J. Jones. Markov models for time series with mixed distribution. *Environmetrics*, **11** (2000), 327–339.

[4] R. I. Jennrich. An introduction to computational statistics. Prentice Hall, New Jursey, 1995.

[5] P. McCullagh, J. A. Nelder. Generalized linear models. Chapman and Hall, London, 1989.

[6] R. W. Katz. An application of chain-dependent process. *J. of Appl. Probability*, **14** (1977), 598–603

[7] R.W. Katz, M. B. Parlange. Conditioning on indices of atmospheric circulation. *Meteorol. Appl.*, **5** (1998), 75–87.

[8] A. N. Pettitt. A nonparametric approach to the change point problem. *Appl. Statist.*, **28** (1979), 126–135.

[9] R. D. Stern, R. Coe. A model fitting analysis of daily rainfall data. *J. of the Roy. Statist. Soc., A*, **147** (1984), 1–34.

[10] D. A. Woolhiser. Modeling daily precipitation - progress and problems. In *Statistics in the environmental and earth sciences*, Walden, A.T. and Guttorp, P. (eds.), Halsted Press, 71–89.

[11] W.Zucchini, N. Neykov, P. Neytchev. Development of a daily precipitation model for south-west Bulgaria. Research Report conducted within the framework of cooperation between the Deutsche Forschungsgemeinschaft and the Bulgarian Academy of Sciences, Göttingen, (2001a).

[12] W. Zucchini, N. Neykov, P. Neytchev. An assessment of the effects of climate change on precipitation in Bulgaria. Research Report conducted within the framework of cooperation between the Deutsche Forschungsgemeinschaft and the Bulgarian Academy of Sciences, Göttingen, (2001b).

N. Neykov, P. Neytchev
National Institute of Hydrology and Meteorology
Bulgarian Academy of Sciences
66 Tsarigradsko chaussee
Sofia 1784, Bulgaria
e-mail: `neyko.neykov@meteo.bg`