Bulg. J. Phys. **44** (2017) 162–188

Development of Stochastic Daily Weather Generator Conditional on Atmospheric Circulation. Part 1: Daily Precipitation Model*

N. Neykov, **P. Neytchev**
National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66 Tsarigradsko shose Blvd., 1784 Sofia, Bulgaria

Received *10 December 2016*

Abstract. We consider development of daily precipitation models for 31 sites in Bulgaria. The precipitation processes are modeled as a two-state first-order nonstationary Markov model with mixed transition density of a discrete component at zero and a continuous component describing non-zero amounts. Binary logistic regression is used to fit the occurrence data, and the intensity series is modeled by gamma distribution. Standard software for generalized linear models can be used to perform the computations. Detailed model validation is carried out on various aspects. The proposed model reproduces well the precipitation statistics for the observed and reserved data.

PACS codes: 92.60.-e, 02.50.Ga

1 Introduction

In Bulgaria, precipitation is the element of climate most influential in determining the variety and abundance of flora and fauna, land use, economic development and practically all aspects of human activity. There have, in consequence, been numerous studies of the climatic and agricultural regions of Bulgaria based largely on the areal distribution and seasonality of precipitation. Many of these studies have concentrated almost exclusively on the annual and monthly precipitation regime and the mapping of means as [1–3]. It is well known that annual and monthly means provide little or no information on many properties of the precipitation that are relevant to the wide variety of precipitation-related activities. For example, the risk and severity of storms, the risk, severity and duration of drought and the timing of precipitation within each year are all aspects of precipitation that are of importance to decision making. It is of course possible to make a special study of any particular property of daily precipitation.

*The work was presented at the 3rd National Congress of Physical Sciences, 28 September – 2 October, 2016, Sofia, Bulgaria.

For instance, [4–6] tabulated the annual and monthly extreme daily precipitation quantiles, the annual and monthly precipitation quantiles, monthly drought quantiles and other statistics at some 400 points in Bulgaria. However the variety of statistics that might be of interest to different decision makers is effectively infinite, which renders that approach problematic. On the other hand the relative frequencies based on the observed rainfall data are short as usual and so the corresponding estimates are far from smooth.

An alternative and more flexible approach is to model the total daily precipitation process itself and thereby encapsulate all the properties of daily total precipitation by means of a small number of model parameters. Until the advent of cheap fast computers this approach would have been fruitless because it is difficult or impossible to determine properties of interest purely analytically. For example it is doubtful that one could derive a formula for the probability of events such as "there will be at least 50 mm rainfall in November but not more than 5 mm on any one day". Computers have made it easy to evaluate the probability of any such event or sequence of events, regardless of complexity. Once calibrated, the model can be used to generate long artificial precipitation sequences (typically 1000–5000 sample paths) which preserve all the statistical properties of precipitation. These sequences can be used to estimate statistics relating to precipitation events in exactly the way one would do so if a long sequence of real rainfall data were available, i.e., parametric bootstrap in action for better assessment of precipitation events according to [7]. For example, suppose that we require an estimate of the probability that station Kneja will have less than 40 mm rainfall in June. This can be done by using the model to generate a 1000 sample year daily precipitation sequence at Kneja and counting the number of years in which this event occurred. Suppose that in 689 out of the 1000 years the June rainfall total was less than 40 mm. Then an estimate of the required probability is 689/1000 = 0.689.

In effect one estimates probabilities of this type by simply regarding the artificial precipitation sequence generated as a very long real precipitation record. One can do this because the model used to generate the sequences preserves the properties of real precipitation sequences, that is the entire probability distribution of daily, monthly and annual precipitation totals, as well as the correlation between precipitation totals on consecutive days, the seasonal distribution of wet and dry runs, and so on. The following examples are intended to illustrate the enormous range of questions that can be answered using the model. One can use the model to compute standard precipitation statistics but in addition to that it can be used to answer detailed specific questions relating to precipitation, including storm-events and drought-related events. Let us consider some of them.

Standard statistics: The distribution of annual precipitation, it's quantiles, mean, standard deviation, etc. The distribution of monthly (weekly, daily) precipitation for each month (week, day) and the associated properties.

Statistics relating to seasonality: The day (week, month, 60-day period, ...) of the year which has the highest (lowest) expected precipitation. The day (week, month, ...) of the year which has the highest (or lowest) probability of having at least, say, 20 mm precipitation.

Statistics and probabilities for specific events such as droughts: What is the probability of having a run of 90 consecutive dry days starting sometime in June or September? What is the least amount of precipitations that can be expected (for example with probability 0.9) between 1st of March and 30th of May? What is the probability to observe at most 20 mm precipitation between 1st of November and 31st of December with no 5-day run with less than 4 mm?

Storm-related statistics and questions: What is the distribution of the 1st-day (2nd-day, 5th-day, ...) storm for each day of the year? How is the frequency of storm distribution over the year? How much precipitation is attributable to storms?

One can answer any of these and similar questions by simply averaging over the generated sequence, that is treating the generated sequence as if it were a very long real precipitation record.

2 The Scope and Aims of the Study

The general objective of the paper is to statistically model the relationship between daily precipitation data at a network of rain gauges broadly covering Bulgaria and large-scale atmosphere circulation data over the Balkans peninsula window 20°W – 30°E, 40°N – 45°N for the period 1960–2000.

The specific objectives are: a) to develop at site two-states first-order non-stationary Markov stochastic daily precipitation amount model, conditional on the value of some synoptic atmospheric variables; b) to assess the fitted at site two-states first-order non-stationary Markov daily precipitation model by comparison of model simulated rainfall statistics to historical rainfall statistics.

2.1 Methodology

Technically, the methodology will involve fitting at-site precipitation occurrence (probability) and intensity (wet-day precipitation amounts) models conditional on some atmospheric variables, to formally test and quantify some suggested structures. We shall closely follow the methodology developed in [8, 9] and further extended in [10–12]. This methodology is based on the Generalized Linear Models (GLMs), see [7, 14].

2.2 Daily precipitation data

The modeling of the daily precipitation totals is carried out over the territory of Bulgaria, located in the Balkan peninsula. The data available for this study are daily precipitation records for the period 1960–2000 from a network of 31 stations of the National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences shown in Figure 1. Each record value represents the precipitation total over a 24 hour period ending at 6:00 GMT, i.e., 8:00 a.m. local time. The Wild's standard rain gauge measurement instrument is used in collecting the data which is mounted 1 m above the ground. In this study a wet day is defined to be a day with at least 0.1 mm of precipitation. Whenever a leap year occurs, the value observed on 29th of February is added to the value observed on March 1st. If March 1st has a missing value then it is replaced by the observed value by 29nt of February.

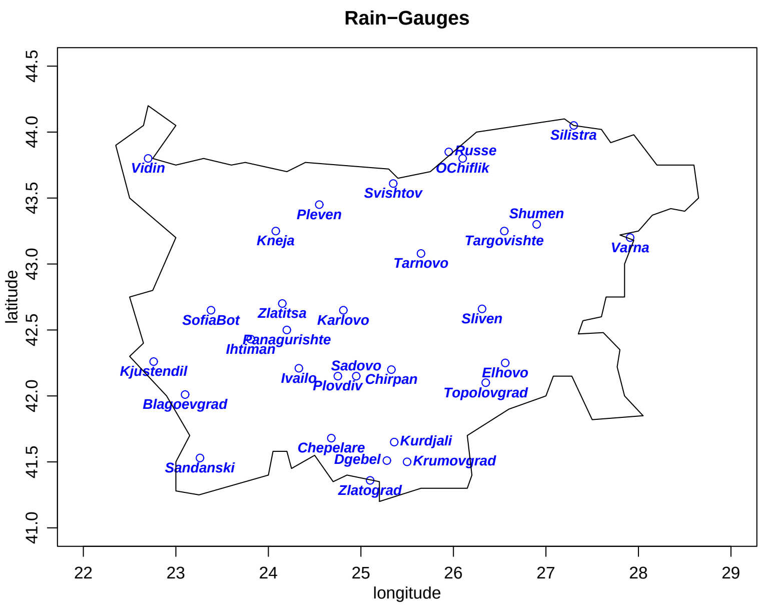

Figure 1. Map of rain stations over the territory of Bulgaria.

2.3 Atmospheric data

In the development of the at-site precipitation occurrence and intensity models the following atmospheric variables are used: sea level pressure (slp), air temperature at surface, air temperature at 2 m, specific humidity at 2 m, u and v wind component at 10 m, precipitation rate and precipitable water, geopotential height at standard levels from 200–1000 (hgt. 200 – hgt. 1000) hPa (pressure lev-

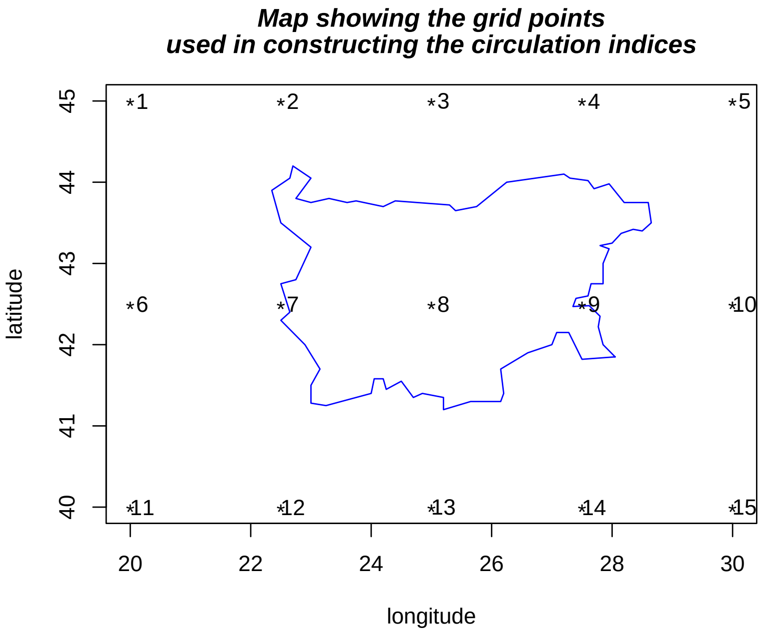

Figure 2. The grid points map used in constructing the circulation indices.

els), air temperature, relative and specific humidity at standard levels, and u and v wind component at the same pressure levels taken at those 15 grid points from the above NCEP-NCAR Reanalysis data within an window closely surrounding Bulgaria from 20°W – 30°E, 40°N – 45°N latitude-longitude grid for the same period as the precipitation data. In the development of at-site stochastic daily precipitation model we tested indirectly the abilities of these atmospheric variables via their widely accepted summaries (derivatives) computed at 15 nodes over the whole area shown in Figure 2. These derivatives are defined as function of the NCEP/NCAR reanalysis model data values taken at the nodes as follows:

```
ampl.t          = tmax.2m(8)-tmin.2m(8)
nwse.slp        = slp(15)-slp(1)
nesw.slp        = slp(11)-slp(5)
ew.slp          = slp(10)-slp(6)
ew.h700         = hgt.700(10)-hgt.700(6)
ew.t850         = air.850(10)-air.850(6)
laplas.slp      = slp(1)+slp(5)+slp(11)+slp(15)- 4*slp(8)
laplas.prate    = prate.sfc(1)+prate.sfc(5)+prate.sfc(11)
                   +prate.sfc(15)-4*prate.sfc(8)
laplas.prwtr    = prwtr.eatm(1)+prwtr.eatm(5)+prwtr.eatm(11)
                   +prwtr.eatm(15)-4*prwtr.eatm(8)
up.t700         = (air.850(8)-air.700(8))/(hgt.700(8)-hgt.850(8))
adv.u.t.850     = uwind.850(8)*(air.850(10)-air.850(6))
adv.v.t.850     = vwind.850(8)*(air.850(3)-air.850(13))
```

```
nwse.h.t.850    = (hgt.850(15)-hgt.850(1))*(air.850(11)-air.850(5))
nesw.h.t.850    = (hgt.850(11)-hgt.850(5))*(air.850(15)-air.850(1))
nwse.h.500.1000 = (hgt.700(15)-hgt.700(1))*([hgt.500(11)
                  -hgt.1000(11)]-[hgt.500(5)-hgt.1000(5)])
nesw.h.500.1000 = (hgt.700(11)-hgt.700(5))*([hgt.500(15)
                   -hgt.1000(15)]-[hgt.500(1)-hgt.1000(1)])
adv.u.r.700     = uwind.700(8)*(rhum.700(10)-rhum.700(6))
adv.v.r.700     = vwind.700(8)*(rhum.700(3)-rhum.700(13))
adv.u1.prate    = uwind.700(7)*(prate.sfc(8)-prate.sfc(6))
adv.u2.prate    = uwind.700(9)*(prate.sfc(10)-prate.sfc(8))
adv.v.prate     = (vwind.700(7)+vwind.700(9))*(prate.sfc(3)
                   -prate.sfc(13))
adv.u1.prwtr    = uwind.700(7)*(prwtr.eatm(8)-prwtr.eatm(6))
adv.u2.prwtr    = uwind.700(9)*(prwtr.eatm(10)-prwtr.eatm(8))
adv.v.prwtr     = (vwind.700(7)+vwind.700(9))*(prwtr.eatm(3)
                   -prwtr.eatm(13))
```

Here, `nesw.slp`, `nwse.slp`, `ew.h700` are the northeast-southwest, northwest-southeast and west-east gradients; `up.t700` is the vertical temperature gradient (a measure of the vertical instability of the air mass; convection precipitations in the warm season are associated with air mass instability); `ew.t850` the horizontal temperature west-east gradient at 850 hPa is a measure of the presence of the frontal zones and represent the dominant directions of the movement of the fronts over Bulgaria; `laplas.prate` and `laplas.prwtr` characterized the cloud systems `nwse.h.t.850` and `nesw.h.t.850` are the geostrophic advections of air.850 in southwest-northeast and northwest-southeast directions; `adv.u.t.850` and `adv.v.t.850` are the advections at level 850 hPa in westeast and south-north directions (measure of the heat transport); `adv.u.r.700` and `adv.u.r.700` are the advections of relative humidity (rhum) at 700 hPa in west-east and south-north directions; `nesw.h.500.1000` and `nwse.h.500.1000` are the termic advections in northeast-southwest and northwest-southeast directions; `adv.u1.prate` and `adv.u2.prate` `adv.u1.prwtr` and `adv.u2.prwtr` `adv.v.prwtr` are the wind components. All these atmospheric derivatives comprises the predictors, covariates, explanatory variables.

2.4 The GLMs and daily rainfall modelling

The sequences of rainfall values exhibit a number of distinctive features. In particular the distribution of daily precipitation depths varies seasonally, rainfall depths on consecutive days are not independently distributed, that is, the probability that a wet day will follow a wet day is higher than the probability that a wet day will follow a dry day, and finally the distribution of rainfall is partly discrete and partly continuous. Any useful model for the description of precipitation sequences must of course preserve all these properties. Several models have been proposed for simulating daily precipitation [9, 15, 16]. Most precipitation models are specified by a discrete occurrence process describing the sequence of wet and dry days, and a continuous distribution function for the amount of precipitation of days with rain. Fitting these models to data is based on GLMs [7]. The parameters of the model are allowed to vary seasonally. The fundamental idea is to predict a probability distribution for each day's rainfall at every site of interest, by relating the expectation of that distribution to the values of various predictors (covariates, explanatory variables). Possible predictors include the occurrence of precipitation on the previous day, previous day's precipitation amount, one or more derived measures from the available atmospheric data, inter and intra annual cycles, variables representing topographic and other location effects. Additionally, interaction terms between these covariates and the occurrence of precipitation on the previous day, allow different effects of covariates on whether it rained the previous day or not. Standard references in the rainfall framework are [8, 9]. The methodology developed in [9] is effective in describing typical rainfall patterns throughout the year, but it assumes the same seasonal pattern for each year and thus is not capable of highlighting trends or other effects not well-modeled by periodic seasonal patterns. Their analysis is based on the simplifying assumption that there is no temporal dependence, i.e., the probability of rainfall occurrence on day t depends only on the two states wet and dry, but not on the rainfall amount, on day $t-1$. However, it is plausible that the occurrence of rain on day t depends on whether or not it rained heavily on the previous day. The properties of such series, called generalized linear autoregressive models, were considered in [14]. Meanwhile, the GLM approach to the stochastic modeling of daily weather variables has been revisited, showing that it can be applied to essentially any variable, including temperature and wind speed, some slowly-varying trend function over the years, e.g., [2, 10, 12, 14, 17, 20, 22]. Examples are provided in which the covariates need not be restricted to simple deterministic functions, such as sine waves to account for annual cycles or trends to account for climate change, but can be geophysical variables such as an index of the North Atlantic Oscillation (NAO) or Arctic Oscillation (AO), e.g., [9, 12, 13].

2.5 Stochastic daily precipitation models

Let Y_t be the precipitation amount on day t and $\mathbf{Z}_t = (Z_{1t}, \ldots, Z_{kt})'$ is a vector of associated atmospheric variables or their derivatives for $t = 1, \ldots, T$. Day t is defined to be dry if $Y_t < c$, where c is a prespecified cutoff constant – we used the standard choice $c = 0.1$ mm – and as wet if $Y_t \geq c$. Observed values of the above quantities are denoted by lower case letters.

The sequence of wet and dry days is represented by the indicator function $J_t = I_{[y_t \geq c]}$ which takes on a value of 1 if day t is wet, and zero if day t is dry. Let $p_t(\mathbf{x}_t)$ represent the probability that day t is wet, conditional on the vector of covariates $\mathbf{x}_t = (j_{t-1}, \ldots, j_{t-p}, y_{t-1}, \ldots, y_{t-p}, z_{1t}, \ldots, z_{kt})'$. Interaction terms between the covariates can be considered as well. We define the daily precipitation intensity as $R_t = Y_t$ if $Y_t \geq c$, as $R_t =$ missing otherwise, and denote its probability density function, conditional on the atmospheric predictors, by $q(r_t|\mathbf{x}_t)$. This distribution is positively skewed because smaller intensities occur more frequently than larger intensities.

The daily precipitation series is modeled using a mixed distribution comprising a discrete component at zero (for dry days) and a continuous-valued right-skewed density (for wet days). As the wet and dry states are exclusive and exhaustive the resulting transition density distribution is given by

$$\begin{aligned} f_t(y_t|\mathbf{x}_t) &= (1 - p_t(\mathbf{x}_t))\, I_{[y_t<c]} + p_t(\mathbf{x}_t) q_t(r_t|\mathbf{x}_t) I_{[y_t \geq c]} \\ &= (1 - p_t(\mathbf{x}_t)) \left(1 - I_{[y_t \geq c]}\right) + p_t(\mathbf{x}_t) q_t(r_t|\mathbf{x}_t) I_{[y_t \geq c]} \\ &= (1 - p_t(\mathbf{x}_t))^{\left(1 - I_{[y_t \geq c]}\right)} \left(p_t(\mathbf{x}_t) q_t(r_t|\mathbf{x}_t)\right)^{I_{[y_t \geq c]}} \end{aligned} \tag{1}$$

In practice $q_t(r_t|\mathbf{x}_t)$ is taken to be gamma, log normal or some other continuous right-skewed distribution. If the interest is on extremes intensities then the generalized Pareto density can be used.

Assuming $p_t(\mathbf{x}_t)$ has no parameters in common with $q_t(r_t|\mathbf{x}_t)$ the likelihood for $(y_{t-p-1}, \ldots, y_n)$ can be factorized as follows:

$$\begin{aligned} L = \prod_{t=p+1}^{T} f_t(y_t|\mathbf{x}_t) &= \prod_{t=p+1}^{T} (1 - p_t(\mathbf{x}_t))^{\left(1 - I_{[y_t \geq c]}\right)} \left(p_t(\mathbf{x}_t) q_t(r_t|\mathbf{x}_t)\right)^{I_{[y_t \geq c]}} \\ = \prod_{t=p+1}^{T} (1 - p_t(\mathbf{x}_t))^{\left(1 - I_{[y_t \geq c]}\right)} &\left(p_t(\mathbf{x}_t)\right)^{I_{[y_t \geq c]}} \prod_{t=p+1, y_t > c} q_t(r_t|\mathbf{x}_t). \end{aligned} \tag{2}$$

Standard GLMs software can be used to estimate the unknown parameters due to this factorization of the likelihood. The first part is the likelihood of the binary time series and the second product is the likelihood of the intensity time series.

Because the Bernoulli and Gamma distributions (with fixed shape parameter) belong to the exponential family, the parameter estimates of the precipitation occurrence and intensity models for p_t and μ_t, respectively, may be obtained by using the theory for estimation for GLMs and/or GAMs, [7] and [19]. All model fitting in this paper was done with the free software environment for statistical computing and graphics R. We used the functions: `glm` with argument `family=binomial(link="logistic")` for occurrence and `family=Gamma(link="log")` for intensity, and `gamma.shape` of the library `MASS` for the maximum likelihood estimate of the shape parameter.

We note that on the base of the estimated parameters for the mean model μ_t of the intensity model the scale parameter of the gamma distribution is given by $\hat{s}_t = \hat{\mu}_t / \hat{\beta}_t$ where $\hat{\beta}_t$ is the MLE of the shape parameter.

Thus inference about Y_t can be done provided $p_t(\mathbf{x}_t)$ and $q_t(y|\mathbf{x}_t)$ are estimated.

The standard approach is to model the probabilities $p_t(x_t)$ within GLMs with the logistic distribution, i.e., logit link function

$$\begin{aligned}\text{logit}(p_t(\mathbf{x}_t)) &= \log\left(\frac{p_t\mathbf{x}_t}{1-p_t(\mathbf{x}_t)}\right) = u(\mathbf{x}_t)\\ &= \alpha_0 + \sum_{l=1}^{p}(\alpha_l j_{t-l} + g_l(y_{t-l})) + \sum_{l=1}^{k} g_{p+l}(z_{lt}) + g_{p+k+1}(t). \quad (3)\end{aligned}$$

The function $u(\mathbf{x}_t)$ should be periodic and approximately sinusoidal in shape in order to reflect the seasonal behaviour of rainfall occurrence, and a remainder term accounts for deviations from this pattern, i.e., the g_l for $l = 1, \ldots, p+k+1$ should be smooth functions. For instance $g_{p+k+1}(t)$ can be a finite Furrier series

$$\begin{aligned}\sum_{l=1}^{n_1}[\alpha_{2l-1}s_t(l) + \alpha_{2l}c_t(l)] &+ \sum_{l=1}^{n_2}[\beta_{2l-1}s_t(l) + \beta_{2l}c_t(l)]\, j_{t-1}\\ &+ \sum_{l=1}^{n_3}[\delta_{2l-1}s_t(l) + \delta_{2l}c_t(l)]\, j_{t-1}\log(y_{t-1}), \quad (4)\end{aligned}$$

where $\mathrm{c}_t(l) = \cos(2\pi tl/365.25)$ and $s_t(l) = \sin(2\pi tl/365.25)$ for $t = 1, \ldots, T$, the terms involving α's describe seasonal changes in rainfall probability following a dry day, the terms involving β's describe the difference in patterns over the year between days following a wet day and days following a dry day, the terms involving δ's describe a differing effect of previous day's intensity throughout the season and the orders n_1, n_2 and n_3 are usually taken to be 1 or 2. As $j_{t-1} = 0$ when $y_{t-1} = 0$ then for convenience let $j_{t-1}\log(y_{t-1}) = 0$. Instead of logarithmic transformation of the intensity y_{t-1} another one such as cube root might be used in order to enter it into the model in a more symmetric way. A simple link function, consisting of a seasonal cycle, lagged occurrence and NAO effects, is

$$\begin{aligned}\text{logit}(p_t(\mathbf{x}_t)) = \alpha_0 &+ \alpha_1 j_{t-1} + \alpha_2 c_t(1) + \alpha_3 s_t(1) + \alpha_4 NAO_{t-1}\\ &+ [\beta_2 c_t(1) + \beta_3 s_t(1) + \beta_4 NAO_{t-1}] j_{t-1}, \quad (5)\end{aligned}$$

with a covariate vector

$$\mathbf{x}_t = (1, j_{t-1}, c_t(1), s_t(1), NAO_{t-1}, c_t(1)j_{t-1}, s_t(1)j_{t-1}, NAO_{t-1}j_{t-1})'.$$

Due to the lagged occurrence and related interactions included in the logit link function, the conditional two-states non-stationary transition probabilities of a wet day following a dry day $p_{01}(t) = p_t(\mathbf{x}_t)$ for $j_{t-1} = 0$ and a wet day following a wet day $p_{11}(t) = p_t(\mathbf{x}_t)$ for $j_{t-1} = 1$ are allowed different cyclic behavior in the model. In this way the parameter estimates of these probabilities can be computed from $p_t(\mathbf{x}_t)$ in one run instead formulating two separate models and respective data set. Moreover, based on the total probability formula one can get the following relationship between the unconditional of previous states probability $\pi(t) = Pr(I_t = 1|z_t)$ and the above two transition probabilities: $\pi(t) = \pi(t-1)p_{11}(t) + (1-\pi(t-1))p_{01}(t)$. This representation is very useful in simulation of artificial rainfall sequences because of the recurrence relationship. Indeed, under the plausible assumption $\pi(t) \approx \pi(t-1)$ for any t we get $\pi(t) \approx p_{01}(t)/(p_{01}(t) + 1 - p_{11}(t))$ [9, 16].

Common assumptions for the form of $q_t(r_t|\mathbf{x}_t)$ are exponential, lognormal, gamma or Weibull or other right-skewed continuous distributions. In this study we use the gamma probability density function given by

$$\gamma(r, \mu, \beta) = \begin{cases} \dfrac{(\beta/\mu)^{\beta} r^{\beta} \exp(-\beta r/\mu)}{\Gamma(\beta)} & r > 0 \\ 0 & r = 0, \end{cases}, \tag{6}$$

where $\Gamma(\beta)$ is the gamma function, μ is the mean and β is the shape parameter ($1/\sqrt{\beta}$ is the coefficient of variation of the distribution).

It is plausible to assume that the intensity on day t depends on whether or not it rained heavily on the previous day. In other words, R_t is considered to be gamma distributed with conditional mean μ_t on day t. A standard approach in GLMs modeling is to link the parameters of the gamma distribution to covariates as follows:

$$\log \mu_t = \log \mu(\tilde{\mathbf{x}}_t) = \tilde{\mathbf{x}}_t^T \theta, \tag{7}$$

where θ is vector of unknown parameters and the covariate vector $\tilde{\mathbf{x}}_t$ is related by $\mathbf{x}_t$. The log-link function is used to ensure positiveness of μ_t in maximization of the intensity likelihood. A general class of such a log-link function can be defined by

$$\log \mu(\tilde{\mathbf{x}}_t) = \tilde{u}(\tilde{\mathbf{x}}_t) = \theta_0 + \sum_{l=1}^{p} (\theta_p j_{t-l} + h_l(y_{t-l})) + \sum_{l=1}^{k} h_{p+l}(z_{lt}) + h_{p+k+1}(t), \tag{8}$$

where the function $\tilde{u}(\tilde{\mathbf{x}}_t)$ has to be similar to $u(\mathbf{x}_t)$, i.e., $h_{p+1}, \ldots, h_{p+k+1}$ have to be smooth functions such as $g_{p+1}, \ldots, g_{p+k+1}$. A simple log-link function, consisting of a seasonal cycle, lagged occurrence and intensity and NAO effects, is

$$\log \mu(\tilde{\mathbf{x}}_t) = \theta_0 + (\theta_1 + \theta_2 \log(y_{t-1}))j_{t-1} + \theta_3 s_t(1) + \theta_4 c_t(1) + \theta_5 NAO_{t-1}. \tag{9}$$

For the days with wet or dry previous day, i.e. $j_{t-1} = 1$ or $j_{t-1} = 0$, we get the two components $\mu_t^{w|w} = \mu^{w|w}(\tilde{\mathbf{x}}_t)$ and $\mu_t^{w|d} = \mu^{w|d}(\tilde{\mathbf{x}}_t)$ of $\mu_t = \mu(\tilde{\mathbf{x}}_t)$.

Then the analog of the unconditional of previous day probability $\pi(t)$ is the unconditional of previous day mean intensity $\mu(t)$ defined by

$$\mu_t = \pi(t-1)\mu_t^{w|w} + (1-\pi(t-1))\mu_t^{w|d}. \quad (10)$$

3 Results: Model Inference and Validation

3.1 Comparing models

One important question to answer is whether the larger model, F with $p+q$ predictors provides an improvement in fit over the smaller nested model, R with p predictors. Within the GLMs framework the ML estimation of nested models uses a simple procedure called the deviance statistics defined as $D = 2\{l_F(\hat{\theta}_{p+q}) - l_R(\tilde{\theta}_p)\}$, where $l_F()$ and $l_R()$ are the maximised negative log-likelihood for models F and R, respectively. D has a $\chi^2(q)$ asymptotic distribution, with q degrees of freedom; this is the likelihood ratio test. A test of validity of model R relative to F at α(=0.05 or 5%) level of significance is to reject R in favor of F if $D > \chi^2(q, 1-\alpha)$ (the $1-\alpha$ quantile of the $\chi^2(q)$). This suggest that model F explains substantially more of the variability in the data than model R.

The Bayesian information criterion (BIC) defined as $BIC = -2\log(l(\hat{\theta})) - k\log(T)$ is used to compare non-nested models, where $l(\hat{\theta})$ is the maximized log-likelihood, k is the number of model parameters, and T is the number of days of data. The preferred model will be the one minimizing BIC.

The Wald test, the ratio $\hat{b}/se(\hat{b})$, is used to test the hypothesis that $\hat{b} = 0$ versus $\hat{b} \neq 0$ where $\hat{b}$ is the estimate of the unknown parameter b and $se(\hat{b})$ is its estimated standard error. However, due to lack of robustness of the Wald test the inference have to relay on the analysis of the deviances, more precisely, on their successive differences in order to assess which of these differences are significant and which are not in building the model.

3.2 Model inference

The parameter estimates of the occurrence and intensity models were fitted to the data for any one sites. Due to space limitations only selected outputs about Kneja station will be presented.

Let us consider at first the output presented in Table 1 of the occurrence model with linear predictors – the intercept, a harmonic, previous day's occurrence and an interaction between them. The estimated null deviance of this model is 14203.1 on 11323 degrees of freedom with a BIC=14212.33 according to Table 2. It is seen that deviance decreases to 14060.7 on 11320 degrees of freedom when introducing the seasonal cycle based on one harmonic:

Table 1. Station Kneja: precipitation occurrence model parameter estimates based on seasonal cycles and previous day wet-dry status for the 1960-1990 period. Signif. codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1.

Coefficients:	Estimate	Std.Error	z value	Pr($>\|z\|$)	Signif.
(Intercept)	-1.22755	0.02741	-44.787	<2e-16	***
sin	0.31329	0.03841	8.156	3.47e-16	***
cos	-0.03385	0.03909	-0.866	0.38652	
J_{t-1}	1.28482	0.04352	29.519	<2e-16	***
$J_{t-1}\times$sin	-0.12123	0.06230	-1.946	0.05168	.
$J_{t-1}\times$cos	0.19655	0.06051	3.248	0.00116	**

sin= $\sin(2\pi t/365)$ and cos= $\cos(2\pi t/365)$. Because the sequences of wet and dry days exhibits moderate persistence, the inclusion of wet-dry status of the previous day improve the reproduction of the persistence of daily rainfall occurrence. Indeed, the residual deviance significantly decreases to 13171.1. Thus for this model the deviance reduced to 13156.4 on 11317 degrees of freedom while BIC=13212.40. Let us call this model occurrence seasonal model. Therefore we can conclude that the occurrence seasonal model is an improvement against the null model – the model with just a constant term, the intercept, which means no relation between the predictors and the response. Thus the Null deviance correspond to the deviance for the model with an intercept term only, while the Residual deviance is the deviance of the fitted model.

The next step is to extend this model by incorporating some atmospheric variable or derived indices as predictors. Because the correlation between the daily precipitation and some of the atmospheric variables or their derivatives in observed data is generally weak, it is quite difficult to decide which ones might be considered as potential predictor variables (covariates) from statistical view point. Thus we add all atmospheric variables to the seasonal model as predictors and get a full precipitation occurrence model. Our goal is to develop a stochastic daily precipitation model including a reasonable subset of meaningful

Table 2. Station Kneja: analysis of deviances table of the precipitation occurrence model based on seasonal cycles and previous day wet-dry status.

	Df	Deviance Resid.	Df	Resid. Dev	Pr($>\|z\|$)
NULL			11322	14203.1	
sin	1	137.8	11321	14065.3	8.008e-32
cos	1	4.6	11320	14060.7	3.271e-02
J_{t-1}	1	889.6	11319	13171.1	1.781e-195
$J_{t-1}\times$sin	1	4.1	11318	13167.0	4.211e-02
$J_{t-1}\times$cos	1	10.6	11317	13156.4	1.154e-03

atmospheric derivatives and their interactions with wet-dry state of the previous day as predictors. An efficient way to derive such a parsimonious model is to use a logistic regression estimation procedure in a step-wise manner. Indeed, the step-wise procedure stepAIC from MASS library turns out to decreases significantly the residual deviance to 10023.142 on 11303 degrees of freedom with the smallest BIC=10209.83.

The parameter estimates of the final model including the most influential derivatives are given in Table 3. In this way the influence of the atmospheric derivatives on the precipitation occurrence probability is not modeled as the same for both previous wet and dry days because J_{t-1} and its interactions are highly statistically significant. Therefore the modeled effects on the probability of a transition from a wet day to a wet day, $p_{11}(t)$, can be 'predicted' as function of the time of the year and of the values of the atmospheric derivatives by setting the previous day's precipitation occurrence to $J_{t-1} = 1$ in the fitted model p_t. Similarly, the effects on $p_{01}(t)$ can be 'predicted' by setting the previous day's precipitation occurrence to $J_{t-1} = 0$ in p_t. Using the values of $p_{11}(t)$ and $p_{01}(t)$ together with a starting value of 0.3 (approximately 30% of the days are wet), we predict the unconditional of previous day probability of precipitation $\pi(t)$ and the first-order autocorrelation coefficient for the entire 1960–2000 period.

Table 3. Station Kneja: precipitation occurrence model parameter estimates for the period 1960–1990. Signif. codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1.

Coefficients:	Estimate	Std.Error	z value	Pr(> $\|z\|$)	Signif.
(Intercept)	1.115e+00	1.131e-01	9.860	< 2e-16	***
`sin`	2.061e-01	3.743e-02	5.506	3.67e-08	***
`cos`	-5.025e-01	5.299e-02	-9.482	<2e-16	***
`adv.u.t.850`	-1.783e-02	4.918e-03	-3.624	0.000290	***
`adv.v.t.850`	-2.579e-02	3.518e-03	-7.330	2.31e-13	***
`nwse.h.t.850`	-9.035e-04	2.630e-04	-3.436	0.000590	***
`nesw.h.t.850`	1.669e-03	2.841e-04	5.873	4.27e-09	***
`adv.v.r.700`	1.921e-03	2.237e-04	8.587	< 2e-16	***
`adv.u1.prwtr`	9.827e-03	1.352e-03	7.267	3.69e-13	***
`adv.u2.prwtr`	1.380e-02	1.213e-03	11.374	<2e-16	***
`nesw.slp`	-1.794e-03	1.010e-04	-17.767	<2e-16	***
`ew.slp`	-1.187e-03	2.464e-04	-4.817	1.46e-06	***
`laplas.slp`	4.142e-04	5.453e-05	7.596	3.06e-14	***
`ew.h700`	-1.581e-02	2.988e-03	-5.291	1.22e-07	***
`ew.t850`	1.725e-01	2.856e-02	6.040	1.54e-09	***
`ampl.t`	-2.624e-01	1.048e-02	-25.050	<2e-16	***
J_{t-1}	5.427e-01	1.654e-01	3.282	0.001032	**
$J_{t-1}\times$`adv.u1.prwtr`	-1.052e-02	1.971e-03	-5.336	9.50e-08	***
$J_{t-1}\times$ `ampl.t`	5.321e-02	1.623e-02	3.278	0.001044	**

Table 4. Station Kneja: precipitation intensity model parameter estimates based on seasonal cycles and previous day wet-dry status for the period 1960–1990. Signif. codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1.

Coefficients:	Estimate	Std.Error	z value	Pr(> $\|z\|$)	Signif.
(Intercept)	0.523940	0.016727	31.322	< 2e-16	***
`sin`	-0.082261	0.024207	-3.398	0.000686	***
`cos`	-0.185169	0.022997	-8.052	1.11e-15	***
R_{t-1}	0.177363	0.023345	7.597	3.88e-14	***
$J_{t-1}\times$ `sin`	-0.007708	0.033761	-0.228	0.819425	
$J_{t-1}\times$ `cos`	0.040014	0.031938	1.253	0.210338	

Now we consider the seasonal cycle as covariate of the intensity model. The parameter estimates and the deviance analysis are given in Tables 4 and 5. The estimated null deviance equals 1745.47 on 3432 degrees of freedom whereas the residual deviance is 1667.9 on 3427 degrees of freedom. As the reduction is statistically significant due to the harmonic we can conclude that there is a strong seasonal improvement against the model based on the intercept. From this table it is seen that the previous day intensity R_{t-1} is also significant predictor. Further, including various atmospheric derivatives and interaction terms between them decreases significantly the residual deviance to 1439.4 on 3417 degrees of freedom.

The most influential being `ew.slp`, `nesw.slp`, `laplas.slp`, R_{t-1} rainfall amount on the previous days according to the residual deviance table. Most of the interaction terms included in the model are not statistically significant except seasonality and `ampl.t` the lower values of which indicate that the previous day is being wet. The parameter estimates of this complex intensity model are presented in Table 6. The estimated shape parameter of the gamma distribution is $\hat{\beta} = 0.3937768$. The assumption that the estimated shape parameter of the gamma distribution is constant is restrictive, since fitting separate models for precipitation intensity in summer and winter leads to two distinct estimates

Table 5. Station Kneja: intensity model analysis of deviances

	Df	Deviance Resid.	Df	Resid. Dev	Pr(> $\|z\|$)
`NULL`			3432	1745.47	
`sin`	1	7.51	3431	1737.96	4.510e-05
`cos`	1	42.83	3430	1695.14	2.013e-22
R_{t-1}	1	26.53	3429	1668.60	1.756e-14
$J_{t-1}\times$ `sin`	1	0.05	3428	1668.56	0.75
$J_{t-1}\times$ `cos`	1	0.70	3427	1667.86	0.21

Table 6. Station Kneja: precipitation intensity model parameter estimates for the period 1960–1990. Signif. codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1.

Coefficients:	Estimate	Std.Error	z-value	Pr(> $\|z\|$)	Signif.
(Intercept)	9.247e-01	7.984e-02	11.581	< 2e-16	***
`sin`	-6.055e-02	1.618e-02	-3.743	0.000185	***
`cos`	-2.632e-01	2.034e-02	-12.938	< 2e-16	***
R_{t-1}	6.195e-03	1.939e-03	3.195	0.001409	**
`nesw.h.500.1000`	9.398e-06	2.493e-06	3.770	0.000166	***
`adv.u1.prwtr`	-1.735e-03	4.726e-04	-3.671	0.000245	***
`adv.u2.prwtr`	1.885e-03	4.653e-04	4.052	5.20e-05	***
`nesw.slp`	-3.352e-04	3.912e-05	-8.569	< 2e-16	***
`ew.slp`	-5.308e-04	4.681e-05	-11.341	< 2e-16	***
`laplas.slp`	1.800e-04	2.014e-05	8.937	< 2e-16	***
`up.t700`	-3.834e+01	1.149e+01	-3.337	0.000856	***
`ampl.t`	-3.577e-02	3.818e-03	-9.370	< 2e-16	***
J_{t-1}	1.008e-01	2.495e-02	4.041	5.43e-05	***
`JtA850`	-3.781e-03	9.371e-04	-4.035	5.59e-05	***
`JtAP`	1.246e-03	3.740e-04	3.332	0.000871	***
`ew.t850`	1.369e-02	3.845e-03	3.560	0.000376	***

`JtA850`:=$J_{t-1}\times$ `adv.v.t.850`; `JtAP`:=$J_{t-1}\times$ `adv.v.prwtr`

of the shape. However, it is far from trivial to relax this assumption within the GLMs framework.

3.3 Model assessments

In this paragraph an assessment of the model is made based on the predicted values of the $p_{11}(t)$ and $p_{01}(t)$ (the p_t components), $\mu_t^{w|w}$ and $\mu_t^{w|d}$ (the μ_t components), $\pi(t)$ and $\mu(t)$ as a function of each day t and the corresponding atmospheric covariates included in the occurrence and intensity models for the period 1960-1990. Some explanation concerning the plots on the figures that follows are given. The reason is that it is not possible technically to represent a meaningful plot based on all days of the considered period. Thus the tiny circles in all these plots represent the averaged empirical probabilities (relative frequencies of rain on each day of the year) or mean intensities. The dashed lines in all the plots are their weekly and monthly smoothed values. Neither the predicted values of the above functions nor their averaged values are presented on the plots. Instead of their weekly and monthly smoothed values thereof are shown as solid curves. The dotted horizontal lines represent the empirical annual daily occurrence conditional or unconditional of the previous day probability or mean intensity for the period 1960-1990 whereas the dotted curves are based only on the corresponding sine and cosine models. For instance, the probability

$\pi(t)$ value is calculated for each day t of the years and therefore we get 31 values for any day of the year. Then the corresponding averaged daily values are calculated and finally the weekly or monthly smoothed values of these averages are presented thereof as solid lines. The plots concerning $p_{11}(t)$ and $p_{01}(t)$ are based on different relative frequencies of any day of the years in contrast with the $\pi(t)$. This is because the number of cases corresponding to the combination "wet previous day – wet on current day" for any day in the year varies between 0 and 31. Thus the empirical quantities given by the tiny circles in the upper plots of Figures 3 and 5 are based on a less observations number per day in comparison with those in the lower plots.

Thus the plot in Figure 3 characterizes the transition probabilities estimates of

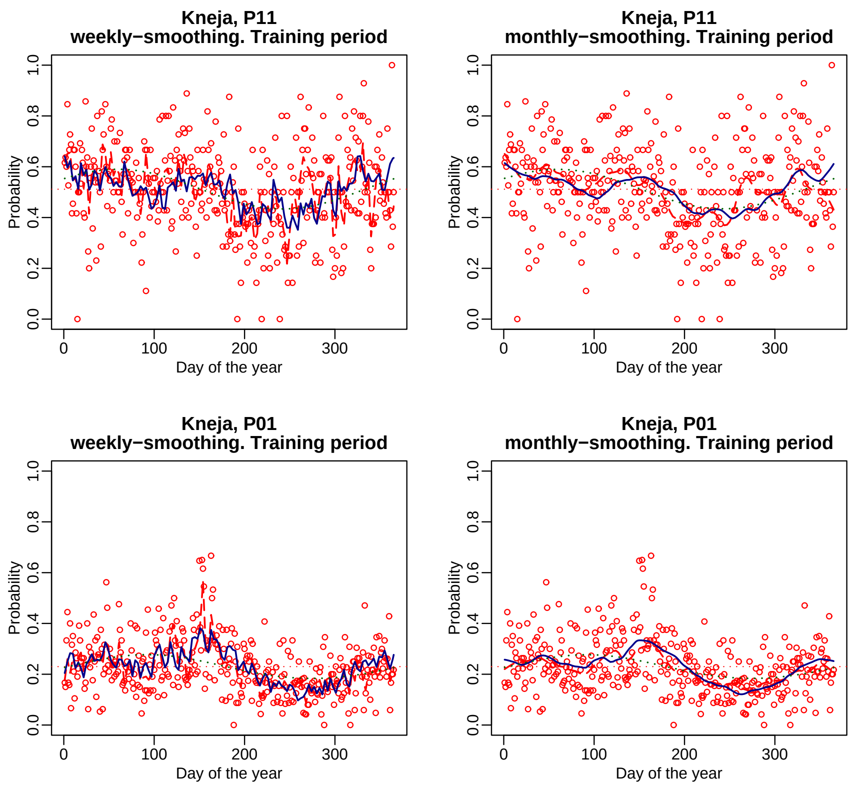

Figure 3. Modeled transition probabilities $p_{01}(t)$ and $p_{11}(t)$ based on atmospheric derivatives. Solid lines: weekly and monthly smoothed values. Tiny circles – empirical probabilities (relative frequencies of rain on each day of the year for the period 1960-1990). Dashed lines: weekly and monthly smoothed empirical values. Dotted lines: transition probabilities values based on two sine and cosine model. Dotted horizontal lines: empirical annual daily occurrence conditional probability for the period 1960–1990.

$p_{01}(t)$ and $p_{11}(t)$ (the p_t components) for station Kneja. One can see the different seasonal cycles as well as the different scales of $p_{11}(t)$ and $p_{01}(t)$. The transition probability $p_{11}(t)$ has a more pronounced dependency on the season, the chances of rain after a wet day being smaller in summer than in remaining part of the year. On the other hand $p_{01}(t)$ does not oscillate as $p_{11}(t)$ of the day of the year, the chances of rain after an already dry day being higher in the spring and cold-half year. The empirical transition probabilities and more so their smoothed versions substantiate our modeling strategy since magnitude and shape throughout the year of these very crude, empirical estimators of the transition probabilities agree very well with their modeled counterparts.

The results presented in the plots of Figures 4 concern the unconditional of the

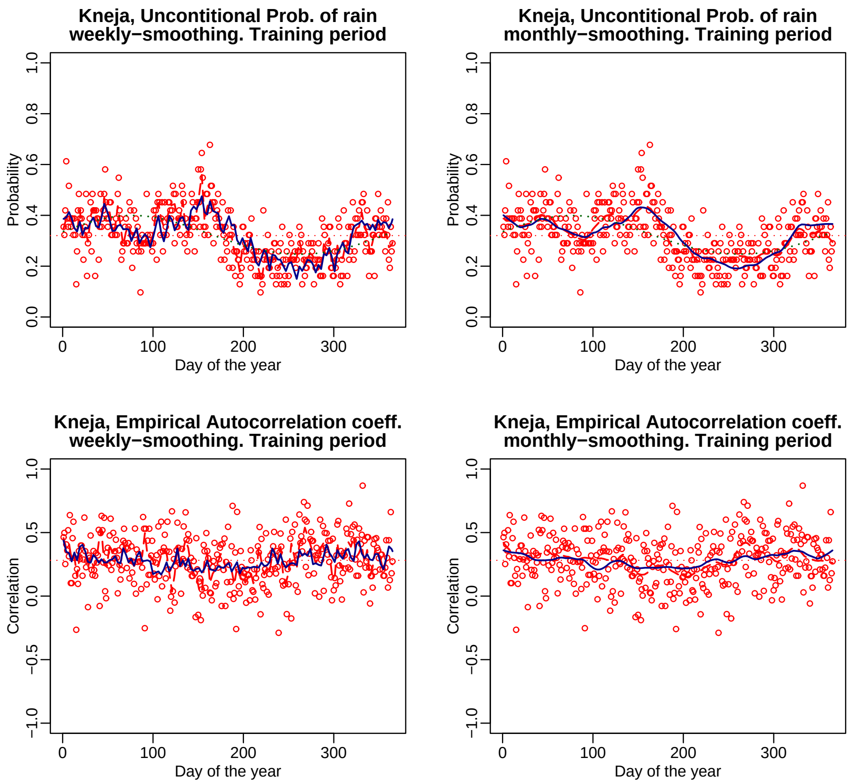

Figure 4. Modeled unconditional of previous day state probability of precipitation and first-order autocorrelation coefficient. Tiny circles – empirical probabilities (relative frequencies of rain on each day of the year for the period 1960-1990). Solid lines: weekly and monthly smoothed values. Dashed lines: weekly and monthly smoothed empirical values. Dotted lines: probability values based on sine and cosine model. Dotted horizontal lines: empirical annual daily occurrence probability for the period 1960–1990.

previous day probabilities $\pi(t)$ for precipitation occurrence. Again, empirical probabilities (relative frequencies of rain on each day of the year) and empirical autocorrelation coefficients (Pearson's correlation coefficient between occurrence on consecutive days on each day of the year) and smoothed versions thereof are shown together with the model curves. It is seen from the plots of these figures that rainfall is again, as expected, less probable in summer than in spring and winter. The first-order autocorrelation is stronger in winter than in summer, meaning that in cold-half year the occurrence of rain more strongly depends on the previous day's occurrence, or from the other viewpoint, in summer it can rain almost independently from the previous day being rainy or not. In both cases, the empirical values and their smoothed versions again indicate the validity of our modeling strategy.

In the plots of Figures 5 as a function of the day of the year are given the means

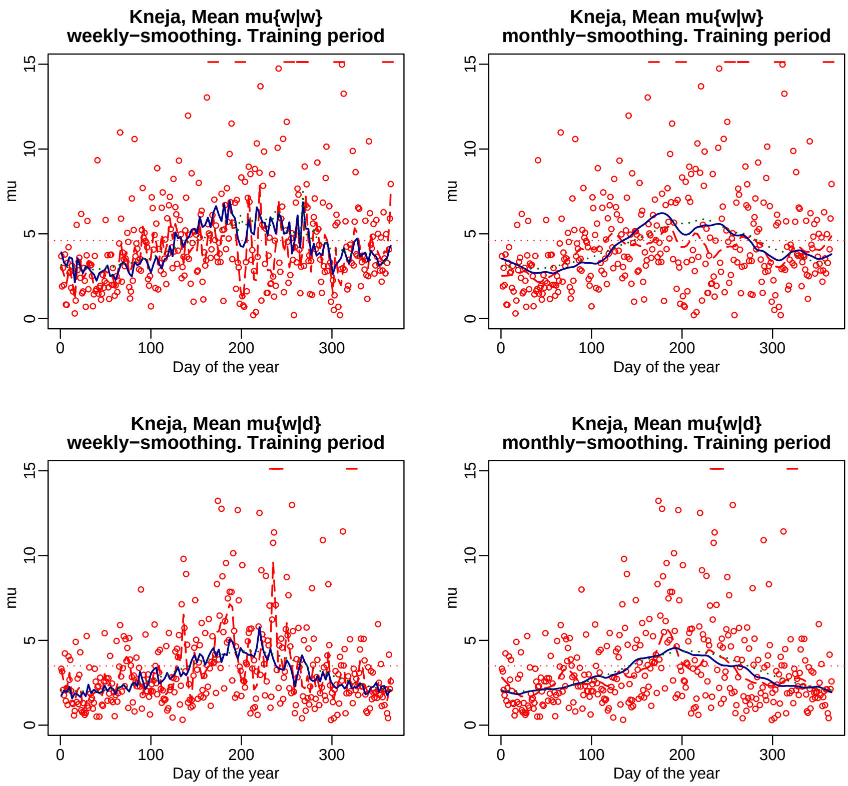

Figure 5. Modeled mean of precipitation intensity, conditional on the previous day state. Tiny circles – empirical means calculated for each day of the year for the period 1960-1990. Solid lines: weekly and monthly smoothed values. Dashed lines: weekly and monthly smoothed empirical values. Dotted horizontal line: average annual daily intensity.

of precipitation intensities, $\mu_t^{w|w}$ and $\mu_t^{w|d}$ (the μ_t components), conditional on the previous day state for stations Kneja. As expected, mean intensity is lower, and intensity is less variable during the winter months. The empirical means calculated for each day of the year from the 1960-1990 period of data and their smoothed version indicate a good agreement of our model with the data, although the fits are not perfect concerning the extreme precipitation. This is a well known deficiency of the gamma distribution. A potential candidate for the intensity model that could be used as regards of extreme rainfalls might be a mixture of two exponential distributions or a hybrid distribution between gamma and generalized Pareto.

The plots presented in Figures 6 are about the means of precipitation intensity,

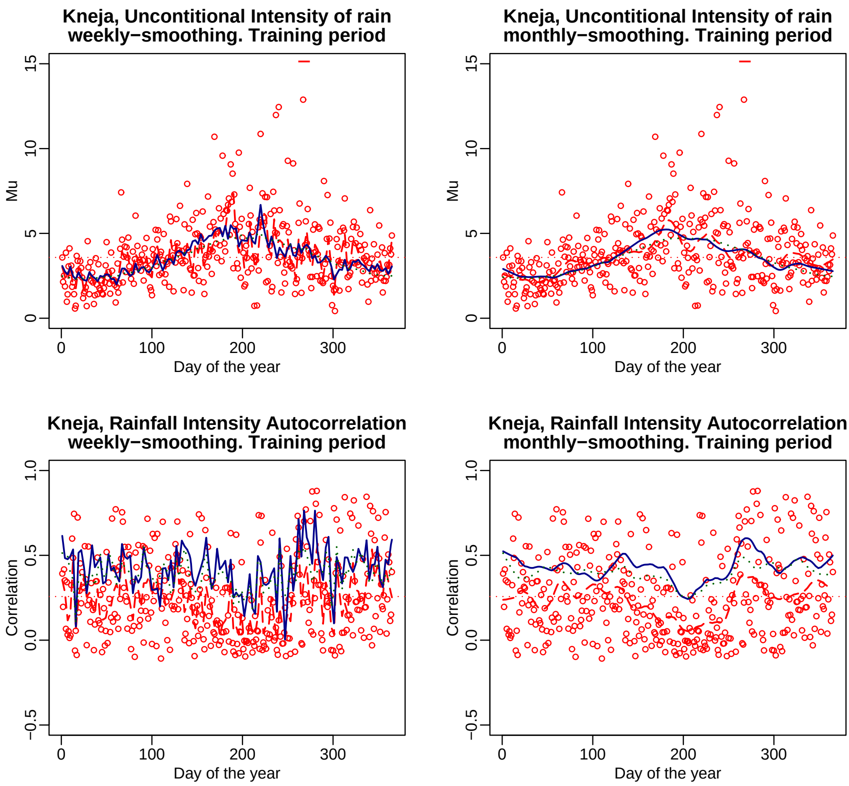

Figure 6. Modeled mean of precipitation intensity, unconditional on the previous day state. Tiny circles – empirical means calculated for each day of the year for the period 1960-1900. Solid lines: weekly and monthly smoothed values. Dashed lines: weekly and monthly smoothed empirical values. Dotted lines: unconditional probability values based on seasonal cycle. Dotted horizontal and curved lines, respectively: average annual daily intensity and based on two sine and cosine model.

μ_t, unconditional on the previous day state for stations Kneja as a function of the day of the year. The tiny circles represent the empirical means calculated for each day of the year for the period 1960-1900. The solid and dashed lines represent weekly and monthly smoothed modeled and empirical values, respectively. The solid lines represent the weekly and monthly smoothed values based on the intensity model mean $\mu(t)$ values calculated for each t of the year for the period 1960-1900. Dotted lines represent unconditional probability values based on seasonal cycle whereas the dotted horizontal and curved lines, respectively represent average annual daily intensity and based on two harmonics model. It is seen from the plots in these figures that although the rainfall is less probable in summer than in spring and winter the expected amounts in summer are much larger. The first-order autocorrelation is stronger in winter than in summer, meaning that in the cool season the mean precipitation intensity more strongly depends on the previous day's intensity, or from the other viewpoint, in summer the precipitation intensity of a wet day is almost independent from the previous day intensity. In both cases, the empirical values and their smoothed versions again indicate the validity of our modeling strategy.

Finally we note that results of a similar standard were obtained for the remaining 30 sites.

3.4 Model validation

This section provides validation of our stochastic precipitation at-site model. We generate a large number of artificial precipitation series of the same length as the historical (1960-1990) and reserved (1991-2000) data series. Various types of statistics are calculated for both of these data series and are compared. To calculate $\pi(t)$ we use the estimated $p_{01}(t)$ and $p_{11}(t)$, the generated value of J_{t-1} at the previous step and the current atmospheric variables values at day t and generate the wet-dry state J_t for $t = 1, \ldots, T$. The initial state J_0 is randomly selected, e.g., from the binary time series comprised of all days dated 31 December if the generation sequence started on 1 January. We remind that for day t a random number $uran \in [0, 1]$ uniformly distributed is generated and compared with $\pi(t)$. If $uran \leq \pi(t)$ then $J_t = 1$ (wet day) otherwise $J_t = 0$ (dry day) is generated. We generate precipitation intensity on the days for which precipitation occurrence ($J_t = 1$) has been generated using the estimated coefficients and the atmospheric variables entered in the occurrence and the intensity models.

Figure 7 presents conditional simulations, the range in simulated annual number of wet-day frequencies and total rainfall amounts for the historical 1960-1990 and reserved 1991-2000 periods for station Kneja. The model does well at capturing the interannual variability. The plots in Figure 8 gave an impression about the quality of the fits based on the observed versus model generated precipitation intensities (in log scale) for the training 1960-1990 and reserved 1991-2000 periods.

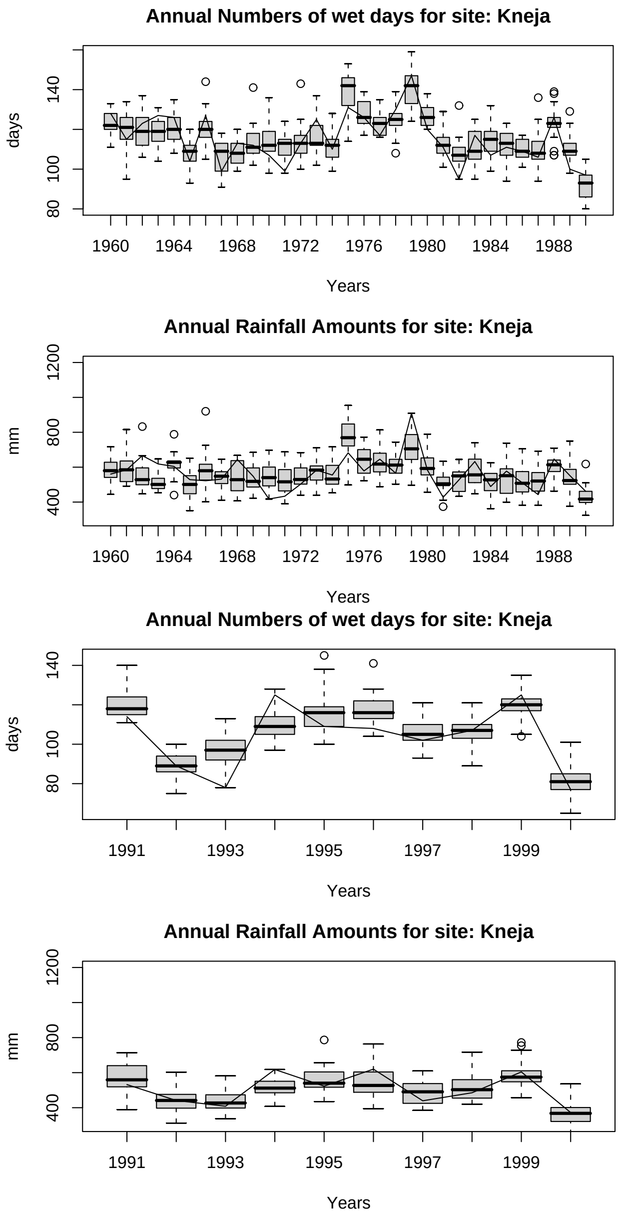

Figure 7. Observed (solid line) versus model-based (box-plots) for the historical 1960-1990 (upper two plots) and reserved 1991-2000 (bottom two plots) annual wet-day frequencies and amounts for Kneja. The edges of the box represent the 25 and 75 percentiles produced by the at-site stochastic precipitation model.

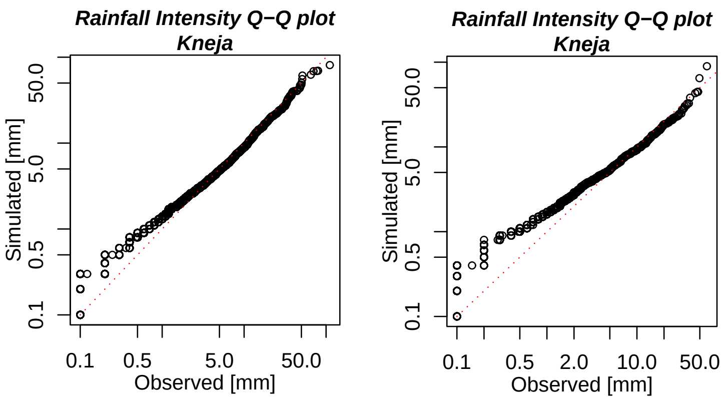

Figure 8. Quantile-quantile plots of observed versus model-based amounts (in log scale) at Kneja station for the training 1960-1990 (left) and reserved 1991-2000 (right) periods.

A more complex issue for precipitation is the temporal correlation in the data like spells of consecutive dry or wet days presented in the plots of Figure 9. The distribution of wet spells , which is often important in hydrological applications, seems to be reproduced quite well at most stations according to the plots in this figure.

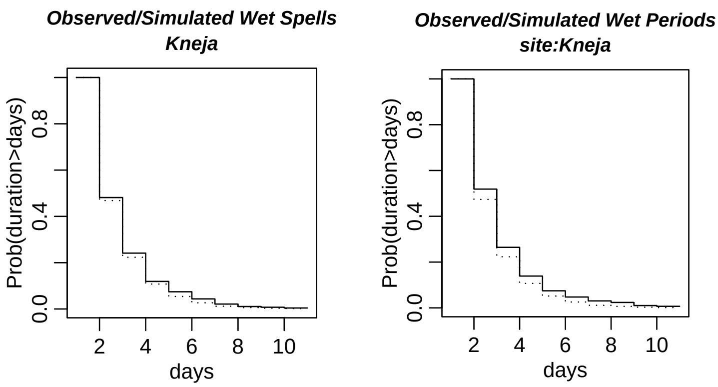

Figure 9. Observed and model-based wet spells distribution at Kneja station for the training 1960-1990 (left) and reserved 1991-2000 (right) periods.

The plots presented in Figure 10 are about the simulated and observed intensity histograms (upper plots) and observed versus simulated log-odds ratios of occurrences for all pairs of stations (lower left) and correlations of intensities (lower right) plots for all station pairs about the training 1960-1990 period. The comparison shows that the model generated rainfall data adequately represents the historical data. The log-odds ratio is a common measure of association for bi-

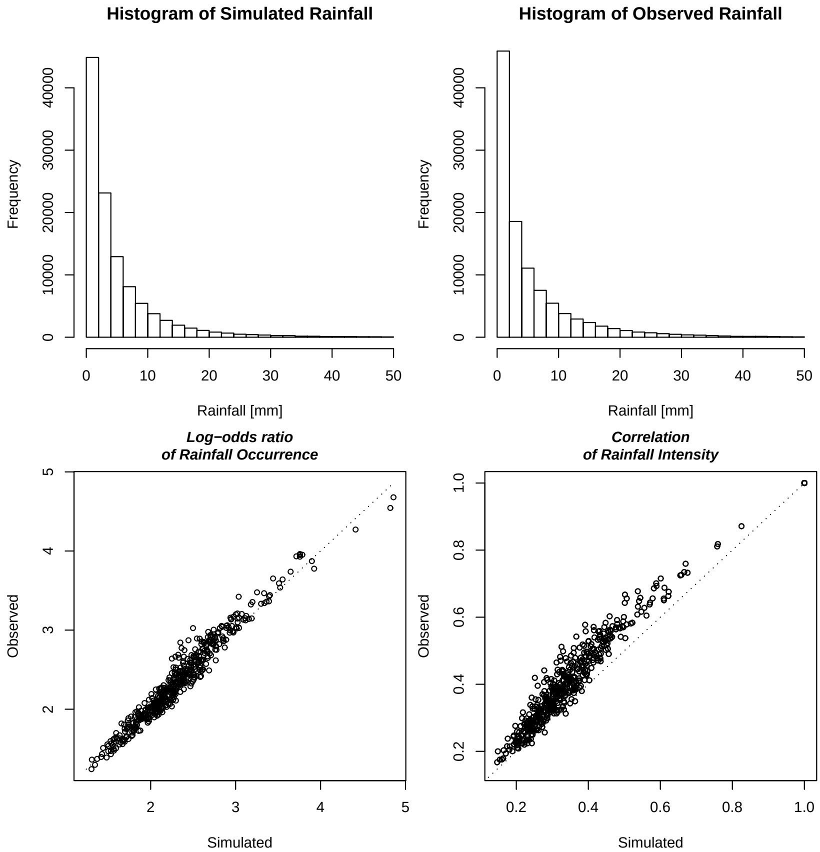

Figure 10. Simulated and observed intensity histograms (upper plots) and observed versus simulated log odds ratios of occurrences for all pairs of stations (lower left) and correlations of intensities (lower right) plots for all station pairs about the training period 1960–1990.

nary variables to reflect the spatial correlation between occurrences at each pair of i and j stations. It is defined as $\log[(n_{11}n_{00})/(n_{10}n_{01})]$, where we denote n_{11} as the number of days when precipitation occurs at both stations i and j, n_{00} as the number of days when precipitation occurs neither at stations i nor at stations j, n_{10} as the number of days when precipitation occurs at stations i but not at stations j, and n_{01} as the number of days when precipitation occurs at stations j but not at stations i. The log-odds ratio takes on values in ($-\infty$, +∞), with large negative numbers indicating strong negative association, large positive numbers corresponding to strong positive association and values close to 0 reflecting a weak association. It is seen that the simulated precipitation probabilities are reproduced well, while the intensity correlations are less adequately.

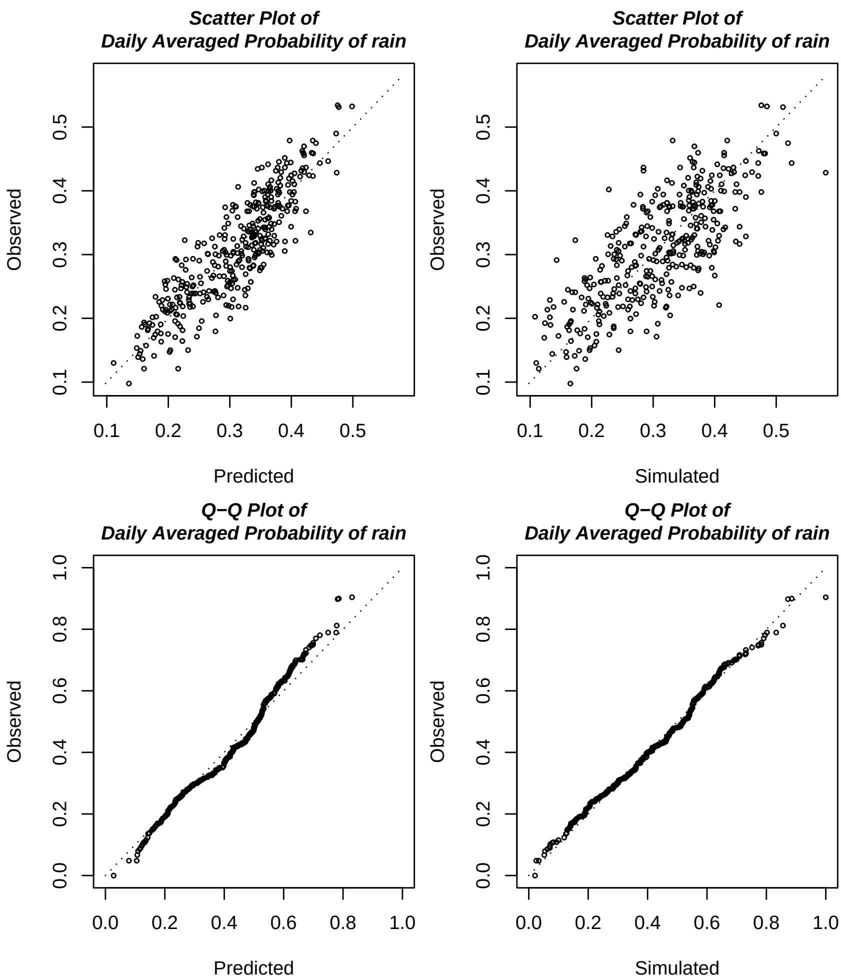

Figure 11. Comparison based on the scatter and Q-Q plots of observed versus model-predicted ($\pi(t)$), and observed versus simulated precipitation probabilities for the training period 1960–1990.

Some indications of how well the model fits the data can be obtained from the comparison between observed and model-based precipitation probabilities $\pi(t)$ and between observed and model-based log-odds for the historical period. For instance, in the plots of Figures 11-12 are given the scatter and Q-Q plots of observed versus model-predicted precipitation probability $\pi(t)$ and observed versus simulated based on $\pi(t)$ for the training 1960-1990 and reserved 1991-2000 periods. The plots title “Daily averaged probability of rain” must be interpret as follow: for each day t of the year the empirical, predicted and simulated $\pi(t)$

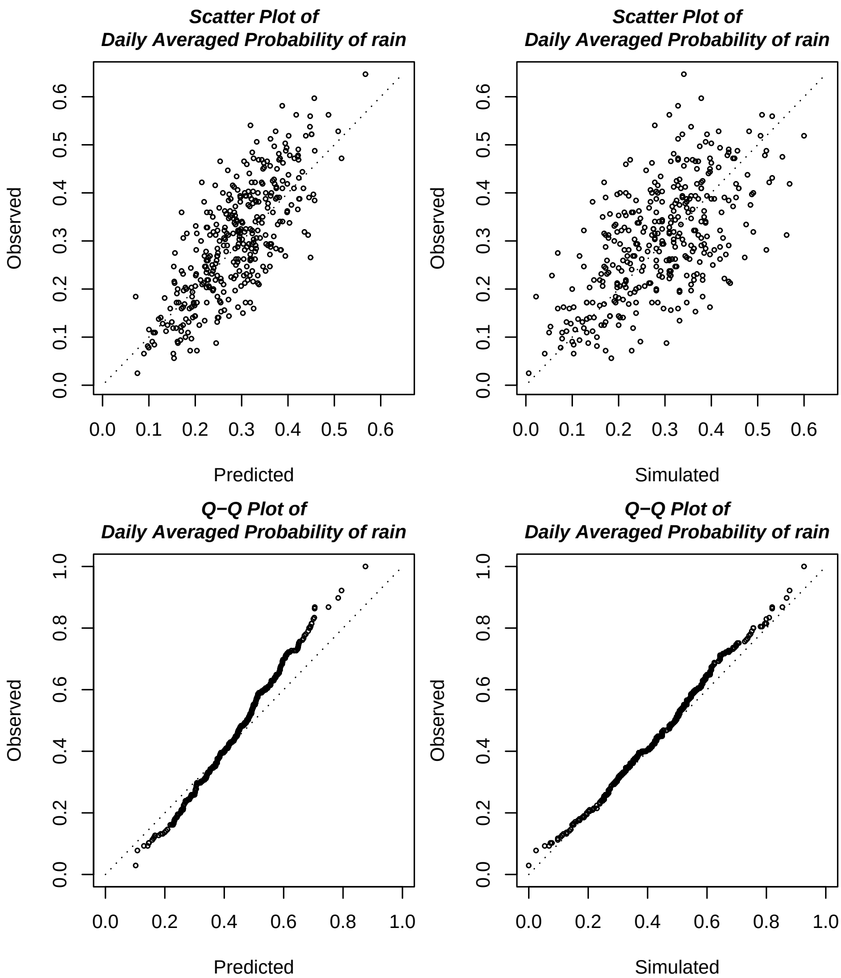

Figure 12. Comparison based on the scatter and Q-Q plots of observed versus model-predicted ($\pi(t)$), and observed versus simulated precipitation probabilities for the reserved period 1991–2000.

spatial mean probabilities are calculated and plotted on the $y-$axis and $x-$axis, respectively. From these plots it is seen these quantities are reproduced well.

4 Conclusion

The main objectives of this paper were to find a suitable stochastic model to describe the process of daily precipitation for a network of stations broadly covering the territory of Bulgaria conditioned upon the values of external atmospheric predictors which vary in both space and time. The study indicates that the GLMs approach provides a powerful tool for interpreting historical precipitation records. We illustrate a few of the wide variety of the questions relating to Bulgarian daily precipitation totals, from daily to annual, that can be answered by the models. These models can be used by hydrologist, climatologist, agriculturalists, geographers and investigators in any field where a knowledge of the precipitation process is required. This is because the simulation is a particularly powerful tool which provides a convenient mean of addressing problems and estimating the probability distributions of conditions that arise in complex processes and systems. These distributions often cannot be obtained analytically or estimated directly from sample which, in the case of daily total precipitation, is too short for reliable estimates to be made. The success of simulation approach depends upon fitting a suitable model to the data, the output of which is shown to preserve the sample distribution of the statistics of interest. Thus the fit of the model needs to be checked carefully and exhaustively. The idea being not to invent new data but to use the sample, via the model, to its full potential.

Acknowledgments

The research was supported by the grant 436 BUL 113/136/0-1 of the Deutsche Forschungsgemeinshchaft within the framework of cooperation between the Deutsche Forschungsgemeinshchaft and the Bulgarian Academy of Sciences. The authors are very thankful and grateful to prof. Walter Zucchini for his encouragements and support.

References

[1] N. Subev, Sv. Stanev (1959) *Climate regions in Bulgaria*. Sofia, BAS Press (in Bulgarian).

[2] R. Kalcheva (1969) *The Precipitation in Bulgaria: A Climate Handbook*, Sofia, BAS Press (in Bulgarian).

[3] E. Koleva, L. Zlatkova (1982) In *The Agroclimate Atlas of Bulgaria*, ed. E. Hershcovich, p. 13 (in Bulgarian).

[4] E. Koleva, R. Peneva (1989) *The Precipitation in Bulgaria: A Climate Handbook*, Sofia, BAS Press (in Bulgarian).

[5] E. Koleva (1991) In *The Climate of Bulgaria*, eds. Sv. Stanev, M. Kiuchukova and St. Lingova, BAS Press, p. 225 (in Bulgarian).

[6] E. Koleva (1991) In *The Climate of Bulgaria* ed. Sv. Stanev, M. Kiuchukova and St. Lingova, BAS Press, p. 238 (in Bulgarian).

[7] B. Efron, R.J. Tibshirani (1994) *An introduction to the bootstrap,* CRC Press, New York.

[8] R. Coe, R.D. Stern (1982) *J. Appl. Meteo.* **21** 1024.

[9] R.D. Stern, R. Coe (1984) *J. Roy. Stat. Soc. A* **147** 1.

[10] G.K. Grunwald and R.J. Jones (2000) *Environmetrics* **11** 327.

[11] R.J. Hyndman, G.K. Grunwald (2000) *Aust. NZ J. Stat.* **42** 145.

[12] R.E. Chandler, H.S. Wheater (2002) *Water Resour. Res.* **38** 1192.

[13] P. McCullagh, J.A. Nelder (1989) *Generalized Linear Models,* Chapman Hall, London.

[14] L. Fahrmeir, G. Tutz (2004) *Multivariate statistical modeling based on Generalized Linear Models*, Springer Verlag, New York.

[15] K.R. Gabriel, J. Neumann (1962) *Q. J. R. Meteorol. Soc.* **88** 85.

[16] W. Zucchini, P. Adamson, L. McNeill (1992) *S. Afr. Jnl. Sci.* **88** 103.

[17] T. Brandsma, A. Buishand (1997) *J. Hydrology* **198** 98.

[18] E.M. Furrer, R.W. Katz (2007) *Clim. Res.* **34** 129.

[19] T.J. Hastie, C. Tibshirani (1990) *Generalized Additive Models,* Chapman Hall, London.

[20] R.W. Katz, M.B. Parlange (1998) *Meteorol. Appl.* **5** 75.

[21] D.S. Wilks, R.L. Wilby (1999) *Prog. Phys. Geog.* **23** 329.

[22] R.E. Chandler (2005) *Environmetrics* **16** 699.

[23] R. Srikanthan, T.A McMahon. (2002) *Hydrol. Earth Syst. Sci.* **5** 653.

[24] N.M. Neykov, P.N. Neytchev, W. Zucchini (2014) *Nat. Hazards Earth Syst. Sci.* **14** 2321.

[25] C. Yang, R.E. Chandler, V.S. Isham, H.S. Wheater (2005) *Water Resour. Res.* **41**; doi:10.1029/2004WR003739.