Nat. Hazards Earth Syst. Sci. Discuss., 2, 1223–1258, 2014
www.nat-hazards-earth-syst-sci-discuss.net/2/1223/2014/
doi:10.5194/nhessd-2-1223-2014

Natural Hazards and Earth System Sciences
Open Access
Discussions

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov[1], P. N. Neytchev[1], and W. Zucchini[2]

[1]National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, Tsarigradsko chaussee 66, Sofia 1784, Bulgaria
[2]Department of Economic Sciences, Georg-August-Universität, Göttingen, Germany

Received: 31 October 2013 – Accepted: 27 January 2014 – Published: 10 February 2014

Correspondence to: N. M. Neykov (neyko.neykov@meteo.bg)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Stochastic daily precipitation models are commonly used to generate scenarios of climate variability or change on a daily time scale. The standard models consist of two components describing the occurrence and intensity series, respectively. Binary logistic regression is used to fit the occurrence data, and the intensity series is modeled by a continuous-valued right-skewed distribution, such as gamma, Weibull or lognormal. The precipitation series is then modeled using the joint density and standard software for generalized linear models can be used to perform the computations. A drawback of these precipitation models is that they do not produce a sufficiently heavy upper tail for the distribution of daily precipitation amounts; they tend to underestimate the frequency of large storms. In this study we adapted the approach of Furrer and Katz (2008) based on hybrid distributions in order to correct for this shortcoming. In particular we applied hybrid gamma – generalized Pareto (GP) and hybrid Weibull–GP distributions to develop a stochastic precipitation model for daily rainfall at Ihtiman in western Bulgaria. We report the results of simulations designed to compare the models based on the hybrid distributions and those based on the standard distributions. Some potential difficulties are outlined.

1 Introduction

Stochastic precipitation models are important for forecasting and simulation purposes in climate, hydrological and environmental system studies in modelling runoff, soil water content, crop growth, droughts and floods. These models can aid in understanding the performance of these systems under specific precipitation regimes. Depending on the required precipitation timescale various models such as hourly, daily, weekly, monthly, seasonal or annual have been developed to quantify complex precipitation features, Srikanthan and McMahon (2002) and Yang et al. (2005). Once the model has been calibrated at a given site one uses it to generate long sequences of artificial

NHESSD
2, 1223–1258, 2014

precipitation at that site. 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 precipitation data were available. As a consequence better risk management strategies and decision-making capabilities can be made.

In the following we shall consider precipitation models in daily time scale only. From statistical point of view, daily precipitation totals are time series with a mixed density comprising a discrete component at zero (for dry days) and a continuous positive real-valued component (for rain days). A standard technique of analyzing the series is to decompose it into two components, namely the occurrence and the intensity processes, Stern and Coe (1984), and then to model these separately using standard generalized linear model (GLM) techniques. The occurrence series, consisting of dry and wet states, is modeled by an autoregressive binary logistic regression, and the intensity series by a continuous-valued right-skewed distribution such as gamma, Weibull, log-normal or mixture of exponential distributions. More precisely, modeling the occurrence series means modeling the transition probabilities of the two-states first or higher order Markov chain, Gabriel and Neumann (1962); Katz (1977). The daily precipitation amounts are then modeled using the joint density of the two components. The seasonal behavior of precipitation is accommodated by allowing the model parameters to vary over the year using a finite Fourier representation, Coe and Stern (1982); Stern and Coe (1984); Woolhiser (1992). The parameters can also be modeled as functions of covariates, e.g. atmospheric factors, such as North Atlantic Oscillation, El Nino-Southern Oscillation, pressure, humidity, temperature, wind speed, or as slowly-varying trend functions over the years. Therefore the occurrence (the states transition probabilities) and intensity model components become non-stationary. The required computations can be carried out using standard software procedures for GLMs and generalized additive models (GAMs), e.g., McCullagh and Nelder (1989); Hastie and Tibshirani (1990); Fahrmeir and Tutz (2001). The properties and applicabilities of such models in different time scales are discussed by Brandsma and Buishand (1997); Katz and Parlange (1998); Grunwald and Jones (2000); Hyndman and Grunwald (2000); Beckman and

NHESSD
2, 1223–1258, 2014

Buishand (2002); Chandler and Wheater (2002); Chandler (2005); Yang et al. (2005); Furrer and Katz (2007), to name a few. Reviews about stochastic precipitation modeling can be found in Woolhiser (1992); Wilks and Wilby (1999); Srikanthan and McMahon (2002) and Maraun et al. (2010).

It is well-known that the above continuous distributions tend to underestimate the heavy precipitation. Furrer and Katz (2008) developed a flexible approach, based on gamma and GP distributions, in order to model the whole spectrum of precipitation intensities. A gamma distribution (with covariates) is fitted to the entire intensity data, and then a GP distribution (again with covariates) is fitted to the observations above an appropriately chosen threshold, u. The two estimated density functions are spliced continuously at u by using the gamma density below the threshold and the GP density (with estimated shape parameter and modified scale parameter estimate) above the threshold. The approach of Furrer and Katz (2008) is general, and so other right-skewed distributions, such as Weibull or inverse Gaussian, can be used instead of the gamma. These authors pointed out some of the difficulties with the procedure, e.g. that threshold selection for splicing the distributions is purely subjective. Carreau and Bengio (2009) proposed another hybrid distribution type which is built by splicing the GP distribution tail to a Gaussian or a truncated Gaussian distributions. The usage of the distribution for stochastic downscaling of precipitation and river runoff purposes is discussed in Carreau et al. (2009) and Carreau and Vrac (2011).

This paper describes a practical implementation and adaptation of the Furrer and Katz (2008) approach and offers an improved daily precipitation model with a heavier tail to describe rainfall series in Bulgaria, conditional on atmospheric data. We also study the reliability of the procedure and report our experience in a concrete example for daily precipitation data at Ihtiman.

NHESSD
2, 1223–1258, 2014

2 Case study – Ihtiman data set

We analysed the daily precipitation series at Ihtiman, Bulgaria, for the time period 1 January 1960–31 December 2007. This series is of particular interest because 234 mm of rainfall was recorded for a 24 h period on 5 August 2005. Each observed value represents the total precipitation over a 24 h period ending at 06:00 GMT (08:00 local time) measured using Wild's standard rain gauge mounted 1 m above the ground. The North Atlantic Oscillation (NAO) daily anomaly time series was used in order to study its relationship to daily precipitation at Ihtiman.

3 Daily precipitation modeling

Let Y_t be the precipitation on day t, $t = 1, \ldots, T$, and Z_t a vector of covariates, e.g. associated atmospheric variables or their derivatives. Day t day 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 $I_t = I_{[y_t \geq c]}$ which takes on a value of 1 if day t is wet, and zero if day t is dry. Let $\pi_t(z_t)$ represent the probability that day t is wet, conditional on the covarates z_t. 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 $q(r_t|z_t)$. 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

NHESSD
2, 1223–1258, 2014

distribution is given by

$$
\begin{aligned}
f_t(y_t|X_t = x_t) &= (1-\pi_t(x_t))\, I_{[y_t<c]} + \pi_t(x_t) q_t(r_t|x_t) I_{[y_t\geq c]} \\
&= (1-\pi_t(x_t))\left(1 - I_{[y_t\geq c]}\right) + \pi_t(x_t) q_t(r_t|x_t) I_{[y_t\geq c]} \\
&= (1-\pi_t(x_t))^{\left(1-I_{[y_t\geq c]}\right)} \left(\pi_t(x_t) q_t(r_t|x_t)\right)^{I_{[y_t\geq c]}}
\end{aligned}
$$

where $X_t = (I_{t-1}, \ldots, I_{t-p}, Y_{t-1}, \ldots, Y_{t-p}, Z_{1t}, \ldots, Z_{k-p,t})^T$ is a vector of potential predictors (covariates).

In practice $q_t(r_t|x_t)$ is taken to be gamma (Stern and Coe, 1984), Weibull (Zucchini et al., 1992), log normal or some other continuous right-skewed distribution. If the interest is on extremes intensities then the GP density can be used.

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

$$
\begin{aligned}
L &= \prod_{t=p+1}^{n} f_t(y_t|x_t) = \prod_{t=p+1}^{n} (1-\pi_t(x_t))^{\left(1-I_{[y_t\geq c]}\right)} \left(\pi_t(x_t) q_t(r_t|x_t)\right)^{I_{[y_t\geq c]}} \qquad (1) \\
&= \prod_{t=p+1}^{n} (1-\pi_t(x_t))^{\left(1-I_{[y_t\geq c]}\right)} \left(\pi_t(x_t)\right)^{I_{[y_t\geq c]}} \prod_{t=p+1, y_t>c} q_t(r_t|x_t).
\end{aligned}
$$

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. The *vglm* procedure from the R package *VGAM* package can fit such models, Yee and Stephenson (2007). This general likelihood maximization procedure, based on an iterative reweighted least squares, is applicable not only to standard GLMs but also to generalized additive models (GAMs), Hastie and Tibshirani (1990). Moreover, by this procedure one can model extreme values easily using generalized extreme value (GEV, block maxima) and peaks

NHESSD
2, 1223–1258, 2014

over threshold (GP) distributions, just like GLMs and GAMs, Green (1984) and Coles (2001).

The standard approach is to model the probabilities $\pi_t(x_t)$ within GLMs with logit link function

$$\begin{aligned}\mathrm{logit}(\pi_t(x_t)) &= \log(\pi_t(x_t)/(1-\pi_t(x_t))) \\ &= u(x_t) = \alpha_0 + \sum_{k=1}^{p}\left(\alpha_k i_{t-k} + g_k(y_{t-k})\right) + \sum_{j=p+1}^{r} g_j(z_{j-p,t}) + g_{r+1}(t).\end{aligned}$$

The function $u(x_t)$ should be periodic and approximately sinusoidal in shape in order in to reflect the seasonal behaviour of rainfall occurrence, and a remainder term accounts for deviations from this pattern, i.e., the g_j for $j = p+1,\ldots,r+1$ should be smooth functions. Interaction terms between the predictors can be considered as well. A simple logit link function, consisting of a seasonal cycle and lagged occurrence and NAO effects, is:

$$\begin{aligned}\mathrm{logit}(\pi_t(z_t)) =& \alpha_0 + \alpha_1 i_{t-1} + \alpha_2\cos(2\pi t/365.25) + \alpha_3\sin(2\pi t/365.25) + \alpha_4\mathrm{NAO}(t-1) \\ &+ [\beta_2\cos(2\pi t/365.25) + \beta_3\sin(2\pi t/365.25) + \beta_4\mathrm{NAO}(t-1)]i_{t-1}.\end{aligned}$$

The covariate vector for this model is $z_t = (1, i_{t-1}, \cos(2\pi t/365.25), \sin(2\pi t/365.25), \mathrm{NAO}(t-1), \cos(2\pi t/365.25)i_{t-1}, \sin(2\pi t/365.25)i_{t-1}, \mathrm{NAO}(t-1)i_{t-1})^T$. Due to the interactions terms included in this logit link function, the conditional two-states non-stationary transition probabilities of a wet day following a dry day $p_{01}(t)$ and a wet day following a wet day $p_{11}(t)$ are allowed different cyclic behavior in the model. In this way the parameter estimates of these probabilities can be computed from $\pi_t(z_t)$ in one run instead formulating two separate models and respective data set as follows: $p_{i1}(t) = \pi_t(z_t)$ for $i = i_{t-1} = \{0,1\}$. Moreover, based on the total probability formula one can get the following relationship between the conditional and unconditional of the previous day probabilities: $\pi_t(z_t) = \pi_{t-1}(z_{t-1})p_{11}(t) + (1-\pi_{t-1}(z_{t-1}))p_{01}(t)$.

NHESSD
2, 1223–1258, 2014

This representation is very useful in simulation of artificial rainfall sequences because of the recurrence relationship between $\pi_t(.)$ and the transition probabilities $p_{01}(t)$ and $p_{11}(t)$. Indeed, under the plausible assumption $\pi(t) \approx \pi(t-1)$ for any t we get $\pi(t) \approx p_{01}(t)/(p_{01}(t) + 1 - p_{11}(t))$. More details can be found in Zucchini et al. (1992) and Furrer and Katz (2007).

The intensities can be modeled by gamma, Weibull or other right-skewed continuous distribution, and the extreme intensities by the GP distribution. There exist various parameterisations for these distributions; those used here are listed below. The density function of the gamma distribution is defined by

$$f(x) = \begin{cases} \frac{b^a \; x^{(a-1)} \exp(-bx)}{\Gamma(a)} & x > 0 \\ 0 & x = 0, \end{cases}$$

where $\Gamma(a)$ is the gamma function, $a > 0$ and $b > 0$ are the rate and shape parameters. The mean, variance and the scale parameters of the gamma distributions are given by $\mu = a/b$, $\sigma^2 = \mu^2/a$ and $\sigma = 1/b$.

The density function of the Weibull distribution is given by

$$f(x) = \begin{cases} \frac{a \; x^{(a-1)} \exp(-(x/b)^a)}{b^a} & x > 0 \\ 0 & x = 0, \end{cases}$$

where $a > 0$ and $b > 0$ are the shape and scale parameters.

The density function of the generalized Pareto distribution with threshold u is given by

$$g(x) = \frac{1}{\sigma}\left[1 + \frac{\xi(x-u)}{\sigma}\right]_+^{-\frac{1}{\xi}-1}$$

where $\sigma > 0$ and ξ are the scale and shape parameters and $[A]_+ = \max(A, 0)$. The shape parameter ξ determines the tail behavior of the GP distribution: a heavy tail if ξ is positive, a bounded tail if ξ is negative and a light (exponential type) tail if $\xi = 0$.

NHESSD
2, 1223–1258, 2014

A standard approach in GLMs and extreme value modeling is to link the parameters of these distributions to covariates as follows

$$\log(a) = \theta_1^T \boldsymbol{x_{1t}},\ \log(b) = \theta_2^T \boldsymbol{x_{2t}},\ \log(\sigma) = \theta_3^T \boldsymbol{x_{3t}},\ \xi = \theta_4^T \boldsymbol{x_{4t}},$$

where θ_i is vector of unknown parameters and the covariate vector $\boldsymbol{x}_{it}$ is a subset of $\boldsymbol{x_t}$ for $i = 1, \ldots, 4$. The log-link function is used to ensure positiveness of the scale (σ) and rate (a) parameters in maximization of the intensity likelihood. Details can be found in Yee and Stephenson (2007). An example of such a log-link function is

$$\log v(x_t) = u_1(x_t) = \beta_0 + \sum_{k=1}^{p} \left(\beta_k i_{t-k} + h_k(y_{t-k})\right) + \sum_{j=p+1}^{r} h_j(z_{j-p,t}) + h_{r+1}(t).$$

where the function $u_1(x_t)$ has to be similar to $u(x_t)$ and $h_{p+1}, \ldots, h_{r+1}$ have to be smooth functions. Interaction terms between the predictors can be considered as well. A simple log-link function, consisting of a seasonal cycle and lagged occurrence and NAO effects, is:

$$\log v(x_t) = \beta_0 + \beta_1 i_{t-1} + \beta_2 \cos(2\pi t/365.25) + \beta_3 \sin(2\pi t/365.25) + \beta_4 \mathrm{NAO}(t-1).$$

4 Modeling daily precipitation totals

In this section we consider a number of daily precipitation models for the Ihtiman series. We start with a brief exploratory data analysis to get an overall impression of the behaviour of the series and then proceed to the development of daily precipitation models using gamma and Weibull regressions, and the GP distribution for the extreme intensities.

4.1 Exploratory data analysis

The interannual and seasonal daily precipitation data distributions at Ihtiman for the whole period are displayed Fig. 1. Seasonality is evident in the lower line plot. The time

NHESSD
2, 1223–1258, 2014

varying threshold based on the 87 % intensity quantile is displayed in the left panel of Fig. 2. The smooth curve was estimated using quantile regression model with inter-annual and seasonal periodic (sine-cosine) components to the daily intensities. The *qr* procedure from the R package *quantreg* was used for this purpose. Details about quantile regression can be found in Koenker (2005). We note that the time varying threshold is a continuous analog of the widely-used procedure of splitting the data into seasons and allowing for different thresholds in each season. The right panel gives the exceedances over this threshold are plotted against the days; the monthly threshold values based on this model are given in Table 1.

4.2 Fitting of extreme precipitation

Having estimated the time-varying threshold model, clusters of exceedances separated from each other by 3 days run length are identified and each cluster maximum is selected. This is done to avoid dependence in the likelihood specification. In this way 557 peaks out of 17 532 observations were extracted resulting in an rate of 11.60 of excesses per year. The cluster peaks are displayed in the right panel of Fig. 2. The tiny black bullets and circles correspond to cold and warm months intensities, respectively. This plot exhibits higher precipitation intensities during the warmer months but there are no strong grounds for applying a varying threshold. These extreme intensities are fitted using a point process model, as in Coles (2001). The advantage of the point process approach is that it unifies the classical block maxima (GEV) and the peaks over threshold (GPD) approaches and allows modeling of the location, scale and shape parameters of the GEV distributions as functions of time dependent covariates in order to account non-stationarity effects. The parameter estimates and Bayesian information criterion (BIC) values for several models fits are presented in Table 2. The non-stationary model that minimizes the BIC includes a seasonal cycle and a lagged NAO effect in the location parameter, a seasonal cycle in the (logarithm of the) scale parameter, and has a constant shape parameter. The estimated parameters of this model support the notion that higher location values are associated with higher precip-

NHESSD
2, 1223–1258, 2014

itation intensities synchronized with negative NAO index anomalies. On the other hand higher scale parameter estimates are associated with higher variability in precipitation extremes. The scale intercept estimate of this model equals exp(2.413) = 11.167. Residual probability plots for the homogeneous model (with no covariates), and the best among these 6 fits (according to the BIC) are shown on Fig. 3. The plots indicate reasonable but by no means not perfect fits and that the non-stationary fits better than the homogeneous model. The corresponding return levels are given in Table 3. It is seen that the non-stationary model gives reasonable return-level estimates for the historical data. All the computations in this section were done by the *pp.fit* and *pp.diag* procedures from the R package *ismev*.

4.3 Gamma and Weibull intensity models

In this section we compare a number of simple gamma and Weibull models with and without covariates (seasonal cycle and NAO effect) in order to assess their ability to fit the entire intensity series. The corresponding parameter estimates and BIC values are given in Table 4. The homogeneous models are presented for completeness only. It is seen that the inclusion of a periodic component significantly reduces the BIC values of both gamma and Weibull models. According to the likelihood ratio test (LRT) the models with seasonal components lead to improvements in comparison with homogeneous models and that the inclusion of NAO effects leads to further improvements. (The LRT and its tail probability values are not presented.) Both models preserve a physical interpretation that heavier intensities are associated with negative NAO anomalies. The left plot of Fig. 4 shows a quantile-quantile (Q-Q) plot for the model with a seasonal cycle based on the GLMs with gamma and Weibull fits. The left panel of the figure is based on data for a single month (August) whereas the middle panel is for the entire period. The Weibull distribution leads to a slightly better fit but the fits are poor with respect to extreme intensities.

For validation purposes the parameter estimates of some simple daily precipitation occurrence models are presented in Table 5. Obviously the homogeneous model is

NHESSD
2, 1223–1258, 2014

completely inadequate but one can see the BIC value reduction with of the remaining models conditional on seasonal cycle, previous day precipitation occurrence and NAO effect with lag one. As expected the seasonal model with lagged occurrence and NAO effects minimizes the BIC.

5 Hybrid gamma–GP density

Furrer and Katz (2008) define the density function of the hybrid gamma–GP distribution as:

$$h(x) = \begin{cases} f(x) & x \le u \\ (1 - F(u))g(x) & x > u \end{cases}$$

where $f(x)$ is the gamma density, $F(x)$ the gamma distribution functions, and the factor $(1 - F(u))$ ensures $h(x)$ normalization.

In order to attain continuity at the threshold u these authors impose the condition

$$f(u) = [1 - F(u)]g(u) = [1 - F(u)]/\sigma.$$

The resulting GPD scale parameter is equal to $\sigma = (1 - F(u))/f(u)$ which is the inverse of the hazard function of the gamma distribution taken at u. Thus the GPD scale parameter can be written in terms of the parameters of the gamma distribution which accommodates the observations below the threshold. The authors recommend the following estimation procedure: (i) fit a GLM with gamma link function with covariates to the entire intensity dataset; (ii) fit a GP with covariates to the observations above the chosen threshold u; (iii) replace the GP scale parameter by the estimated gamma hazard function. Clearly an analogous procedure is applicable for other hybrid distributions, such as Weibull–GP or inverse Gaussian–GP.

NHESSD
2, 1223–1258, 2014

5.1 Experiments with hybrid GP distributions

We now compare the performance of the gamma–GP and Weibull–GP hybrid distributions in fitting the daily precipitation intensity data. We explore the threshold selection and its effect on generation of artificial daily precipitation data. The methodology of Furrer and Katz (2008) previously described is followed closely.

The QQ-plots Fig. 5 are based on different fits of observed vs. fitted gamma (g) and hybrid gamma–GP (h) quantiles of precipitation intensity seasonal model at Ihtiman with thresholds 5 mm (left), 10 mm (middle) and 15 mm (right) in May (upper line plots) and August (middle line plots) and for the entire year (lower line plots). Results of a similar standard are obtained for the Weibull and Weibull–GP distributions. The hybrid models are significant improvement over the gamma and Weibull models.

Figure 6 shows the fitted gamma, GP, hybrid and gamma–GP (upper line plots) and Weibull, GP and Weibull–GP (lower line plots) log-densities with three threshold values 5 mm, 10 mm and 20 mm for precipitation intensity for the entire year. The homogeneous fits (no covariates in the model) are shown only in order to get a better perception. The hybrid density is indeed continuous and possesses a heavier tail than the gamma and Weibull densities. One can see the effect of the threshold choice in GP distribution tail estimation. A lower threshold choice u gives a larger weight $(1 - F(u))$ of the GP distribution. Thus one can expect that the hybrid distributions quantiles corresponding to these lower threshold GP fits would be larger. The plots of Fig. 7 show the 95 % (green), 98 % (red) and 99 % (black) quantiles of the fitted gamma (solid lines) and hybrid gamma–GP (dashed lines) distributions (upper line plots) as well the fitted Weibull and hybrid Weibull–GP (lower line plots) as functions of the day of the year. The tiny black bullets on the plots correspond to observed precipitation for the entire period. Indeed, the hybrid quantiles (dashed lines) on left column plots are higher than those on the right column plots. The effect due to hybridization is most visible on the highest quantiles. Therefore, threshold determination is of crucial importance of calibration daily precipitation hybrid model conditional on atmospheric covariates.

NHESSD
2, 1223–1258, 2014

Figure 8 shows Q-Q plots of observed vs. simulated gamma (g) and vs. hybrid gamma–GP (h) quantiles of the seasonal intensity model with a lagged NAO effect over a year at Ihtiman with 5 mm (top row), 10 mm (middle row) and 15 mm (bottom row) thresholds. The simulated time series consist of 300 samples of 47 yr of daily precipitation totals from GLMs with the gamma distribution (g) and with the gamma–GP (h) hybrid distributions for each threshold. The majority of the simulated data look like those displayed in the left and middle column plots, but a small percentage look like those displayed in the right column plots. Similar results are obtained for the Weibull–GP hybrid distribution. The monthly box-plots of daily observed (white) and simulated (gray) precipitation totals are presented on the plots of Fig. 9 in order to get an impression about their distributions: the left and middle column panels are based on the hybrid intensity distributions where the right column panels are based on classical GLMs with the gamma and Weibull distributions. The classical GLMs with gamma and Weibull intensity component represent the historical data except for the extreme intensities, which is well known deficiency. The hybrid models are capable of generating series with extremes as large as the observed extremes, or (though unlikely to occur) even larger, depending on the threshold choice. The distributions of the monthly observed and simulated precipitation totals are presented on the plots of Fig. 10. From the right column plots of this figure one can see that the standard intensity GLMs with the gamma and Weibull distributions are not capable of generating monthly precipitation totals with similar magnitude as the historical one whereas the hybrid gamma and Weibull–GP distributions are capable to do so.

The distribution functions of the wet spells and number of wet days within a season are important in applications in various studies. The left plot of Fig. 11 shows the distribution of wet spells for the historical and simulated data. As usual wet spells are defined as the number of consecutive days with precipitation. It is seen that the model captures well the temporal correlation in the data. The middle and right plots of this figure show the monthly number of wet days distribution of historical vs. simulated data. Results of a similar standard are obtained for the Weibull and Weibull–GP distributions.

NHESSD
2, 1223–1258, 2014

The distributions of the observed (solid lines) and simulated (dotted lines) precipitation totals over period of 10 and 60 days are shown on the plots of Fig. 12. The simulated data are generated using intensity GLMs with hybrid Weibull–GPD and threshold 15 mm (left column plots) and standard Weibull (right column plots) distributions. It is seen that the hybrid Weibull–GP distribution simulated data possesses a heavier tail than the standard Weibull distribution. Similar results are obtained for the gamma–GP hybrid distribution.

The left plot dots of Fig. 13 represent the estimated conditional probabilities p_{11} and p_{01} and the unconditional precipitation probability $p(t) := \pi_t(z_t)$ (red); the dashed and smoothed lines are based on the R locally weighted scatterplot smoothing procedure loess through the corresponding dots and observed frequencies (not plotted). In the right plot of this figure are given the historical and simulated probabilities (smoothed by loess) of having not less than 40 mm, 80 mm, 120 mm, 160 mm and 200 mm total precipitation for a run of 60 consecutive days, starting on any given day of the year for Ihtiman station. The order of the lines corresponds to their order in the legend. The empirical and the model probabilities match each other closely.

All model fitting and generation of precipitation series was done with the free software environment for statistical computing and graphics: the R Project for Statistical Computing. The *vglm* procedure from *VGAM* package with gamma, Weibull and gpd link functions was used to carry out the estimation (Yee and Stephenson, 2007).

6 Conclusions

Several daily precipitation models with different models for the intensity component were examined. We are able to confirm that, on the whole, the simulated precipitation series based on the hybrid distributions of Furrer and Katz (2008) preserve the properties of the observed series. Although each of the precipitation model components can be estimated using standard software procedures that are widely available, the subjectivity in threshold selection in splicing the distributions is an awkward task.

NHESSD
2, 1223–1258, 2014

Thus development of a daily precipitation models with such distributions, conditional on a large number of atmospheric predictors for downscaling purposes, is still in its early stages. Once this problem would be solved satisfactorily then an extension of the improved at site daily precipitation amount model to a multi-site daily precipitation model would be straightforward on the base of the conditional independence precipitation amount model within the non-homogeneous hidden Markov models framework, Vrac and Naveau (2007) and Neykov et al. (2012).

Acknowledgements. The research was funded partially by the Deutsche Forschungsgemeinschaft, Grant GZ: ZU 237/1-1, the Bulgarian Academy of Sciences and the ESF within the framework of COST Action ES0901 "European Procedures for Flood Frequency Estimation".

References

Beckman, B. R. and Buishand, T. A.: Statistical downscaling relationships for precipitation in the Netherlands and North Germany, Int. J. Climatol., 22, 15–32, 2002. 1225

Brandsma, T. and Buishand, A.: Statistical linkage of daily precipitation in Switzerland to atmospheric circulation and temperature, J. Hydrol., 198, 98–123, 1997. 1225

Carreau, J. and Bengio, Y.: A hybrid Pareto model for asymmetric fat-tailed data: the univariate case, Extremes, 12, 53–76, 2009.

Carreau, J. and Vrac, M.: Stochastic downscaling of precipitation with neural network conditional mixture models, Water Resour. Res., 47, W10502, doi:10.1029/2010WR010128, 2011. 1226

Carreau, J., Naveau, P., and Sauquet, E.: A statistical rainfall-runoff mixture model with heavy-tailed components, Water Resour. Res., 45, W10437, doi:10.1029/2009WR007880, 2009. 1226

Chandler, R. E.: On the use of generalized linear models for interpreting climate variability, Environmetrics, 16, 699–715, 2005. 1226

Chandler, R. E. and Wheater, H. S.: Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland, Water Resour. Res., 38, 1192, doi:10.1029/2001WR000906, 2002. 1226

NHESSD

2, 1223–1258, 2014

Coe, R. and Stern, R. D.: Fitting models to daily rainfall data, J. Appl. Meteorol., 21, 1024–1031, 1982. 1225
Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer, New York, 2001. 1229, 1232
Fahrmeir, L. and Tutz, G.: Multivariate Statistical Modeling Based on Generalized Linear Models, Springer, New York, 1994. 1225
Furrer, E. M. and Katz, R. W.: Generalized linear modeling approach to stochastic weather generators, Clim. Res., 34, 129–144, 2007. 1226, 1230
Furrer, E. M. and Katz, R. W.: Improving the simulation of extreme precipitation events by stochastic weather generators, Water Resour. Res., 2008. 44, W12439, doi:10.1029/2008WR007316, 2008. 1224, 1226, 1234, 1235, 1237
Gabriel, K. R. and Neumann, J.: A Markov chain model for daily rainfall occurrence at Tel Aviv, Roy. Meteo. Soc., 88, 90–95, 1962. 1225
Green, P. J.: Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, J. Roy. Stat. Soc. B, 46, 149–192, 1984. 1229
Grunwald, G. K. and Jones, R. J.: Markov models for time series with mixed distribution, Environmetrics, 11, 327–339, 2000. 1225
Hastie, T. J. and Tibshirani, C.: Generalized Additive Models, Chapman Hall, London, 1990. 1225, 1228
Hyndman, R. J. and Grunwald, G. K.: Generalized additive modelling of mixed distribution Markov models with application to Melbourne's rainfall, Aust. NZ J. Stat., 42, 145–158, 2000. 1225
Katz, R. W.: Precipitation as a chain dependent process, J. Roy. Stat. Soc. B, 16, 671–676, 1977. 1225
Katz, R. W. and Parlange, M. B.: Conditioning on indices of atmospheric circulation, Meteorol. Appl., 5, 75–87, 1998. 1225
Koenker, R.: Quantile Regression, Cambridge University Press, 2005. 1232
Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M. Brienen, S., Rust, H. W., Sauter, T., Themessl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., and Thiele-Eich, I.: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, Rev. Geophys., 48, 1–34, 2010. 1226

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

McCullagh, P. and Nelder, J. A.: Generalized Linear Models, Chapman Hall, London, 1989. 1225

Neykov, N. M., Neytchev, P. N., Zucchini, W., and Hristov, H.: Linking atmospheric circulation to daily precipitation patterns over the territory of Bulgaria, Environ. Ecol. Stat., 19, 249–267, doi:10.1007/s10651-011-0185-9, 2012. 1238

Srikanthan, R. and McMahon, T. A.: Stochastic generation of annual, monthly and daily climate data: A review, Hydrol. Earth Syst. Sci., 5, 653–670, doi:10.5194/hess-5-653-2001, 2001. 1224, 1226

Stern, R. D. and Coe, R.: A model fitting analysis of daily rainfall data, J. Roy. Stat. Soc. A, 147, 1–34, 1984. 1225, 1228

R Development Core Team: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, 2013.

Vrac, M. and Naveau, P.: Stochastic downscaling of precipitation: from dry events to heavy rainfalls, Water Resour. Res., 43, W07402, doi:10.1029/2006WR005308, 2007. 1238

Wilks, D. S. and Wilby, R. L.: The weather generation game: a review of stochastic weather models, Prog. Phys. Geog., 23, 329–357, 1999. 1226

Woolhiser, D. A.: Modeling daily precipitation – progress and problems, in: Statistics in the Environmental and Earth Sciences, edited by: Walden, A. T. and Guttorp, P., Halsted Press, 71–89, 1992. 1225, 1226

Yang, C., Chandler, R. E., Isham, V. S., and Wheater, H. S.: Spatial-temporal rainfall simulation using generalized linear models, Water Resour. Res., 41, W11415, doi:10.1029/2004WR003739, 2005. 1224, 1226

Yee, T. W. and Stephenson, A. G.: Vector generalized linear and additive extreme value models, Extremes, 10, 1–19, doi:10.1007/s10687-007-0032-4, 2007. 1228, 1231, 1237

Zucchini, W., Adamson, P., and McNeill, L.: A model of southern African rainfall, S. Afr. J. Sci., 88, 103–109, 1992. 1228, 1230

NHESSD

2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

Table 1. Monthly thresholds based on time varying threshold.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
7.99	8.45	9.37	10.55	11.65	12.38	12.53	12.05	11.10	9.91	8.82	8.12

NHESSD
2, 1223–1258, 2014

Table 2. Estimated parameters and BIC values (minimum in bold) for candidate point process models for daily precipitation extremes over the entire year with a time-varying threshold at Ihtiman.

	Location μ			Scale $\log\sigma$					
Intercept	Cosine	Sine	NAO	Intercept	Cosine	Sine	NAO	Shape ξ	BIC
34.447				2.469				0.131	2018.098
33.680	−1.222	−2.361		2.455				0.159	2012.348
34.458				2.513	0.033	0.076		0.183	2009.993
33.828	−5.481	−5.951		2.412	−0.176	−0.149		0.126	2003.701
33.834	−5.477	−5.962	−0.062	2.413	−0.176	−0.149		0.127	**2002.516**
33.966	−5.236	−5.237	−0.041	2.388	−0.171	−0.121	−0.111	0.116	2010.117

NHESSD
2, 1223–1258, 2014

Table 3. Return levels based on point process model fit: (i) homogeneous model; (ii) seasonal cycle in the location and scale parameters and NAO index as location parameter.

Years	10	20	50	100	500	1000	5000	10 000
(i) Return Levels	65.73	78.01	95.92	110.99	152.11	172.90	229.78	258.57
(ii) Return Levels	65.44	77.46	94.82	109.29	148.17	167.52	219.65	245.61

NHESSD
2, 1223–1258, 2014

Table 4. Estimated parameters and BIC values for candidate gamma (left) and Weibull (right) models for daily precipitation intensity over the entire year at Ihtiman.

gamma log(rate)						Weibull scale $\log\sigma$					
Intercept	Cosine	Sine	NAO	Shape b	BIC	Intercept	Cosine	Sine	NAO	Shape a	BIC
−1.857					28 616.93	1.409				0.798	28 410.37
−1.834	0.169	0.219		0.751	28 483.62	1.418	−0.162	−0.214		0.808	28 305.52
−1.832	0.174	0.215	−0.057	0.752	**28 312.39**	1.416	−0.167	−0.211	−0.053	0.808	**28 135.98**

NHESSD
2, 1223–1258, 2014

Table 5. Estimated parameters and BIC values for daily precipitation occurrence models over the entire year at Ihtiman.

logit($\pi(x_t)$) Intercept	Cosine	Sine	Occurrence	NAO	BIC
−0.736					22 091.27
−1.186	−0.025	0.239	1.216		20 678.49
−1.186	−0.022	0.240	1.213	−0.036	**20 607.01**

NHESSD
2, 1223–1258, 2014

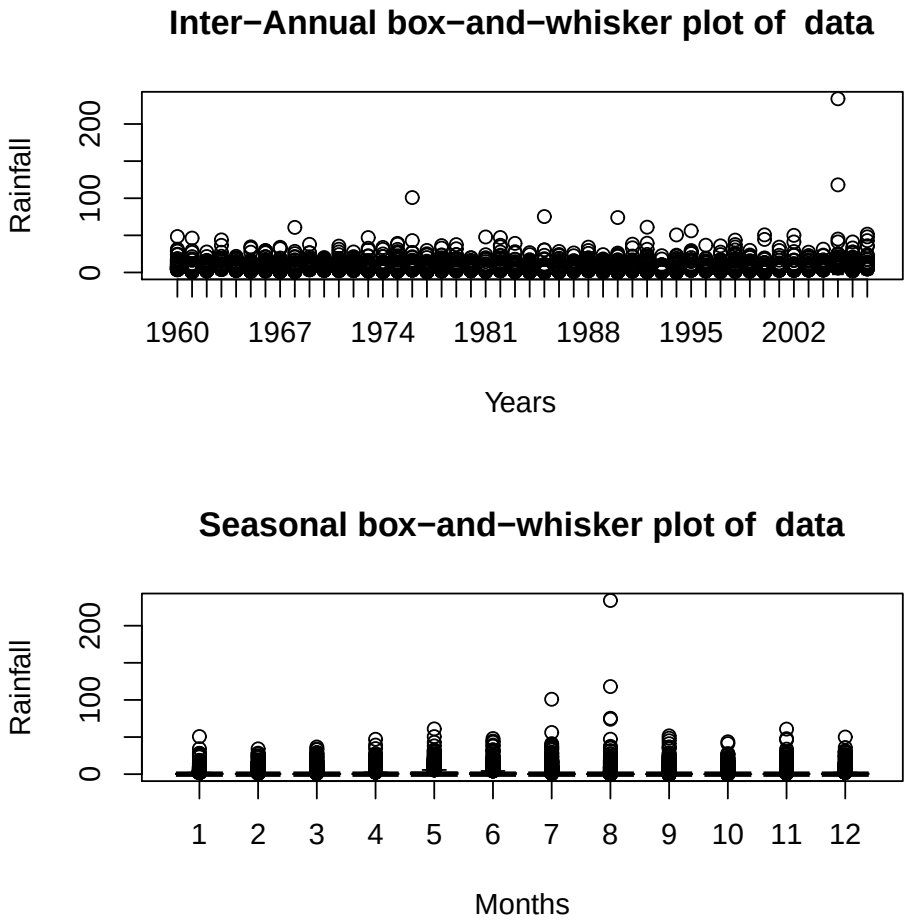

Fig. 1. Station Ihtiman – Box-and-whisker plots of daily precipitation data shown by month.

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

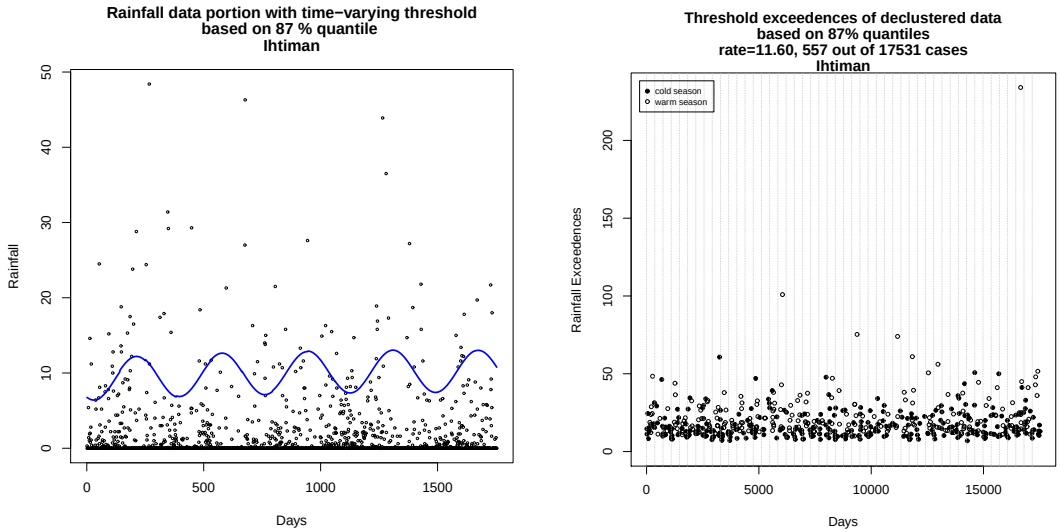

Fig. 2. Station Ihtiman: portion of daily precipitation intensity data with time-varying threshold (solid line) based on 87 % quantile (left plot); exceedences based on declustered daily intensity data (right plot).

NHESSD
2, 1223–1258, 2014

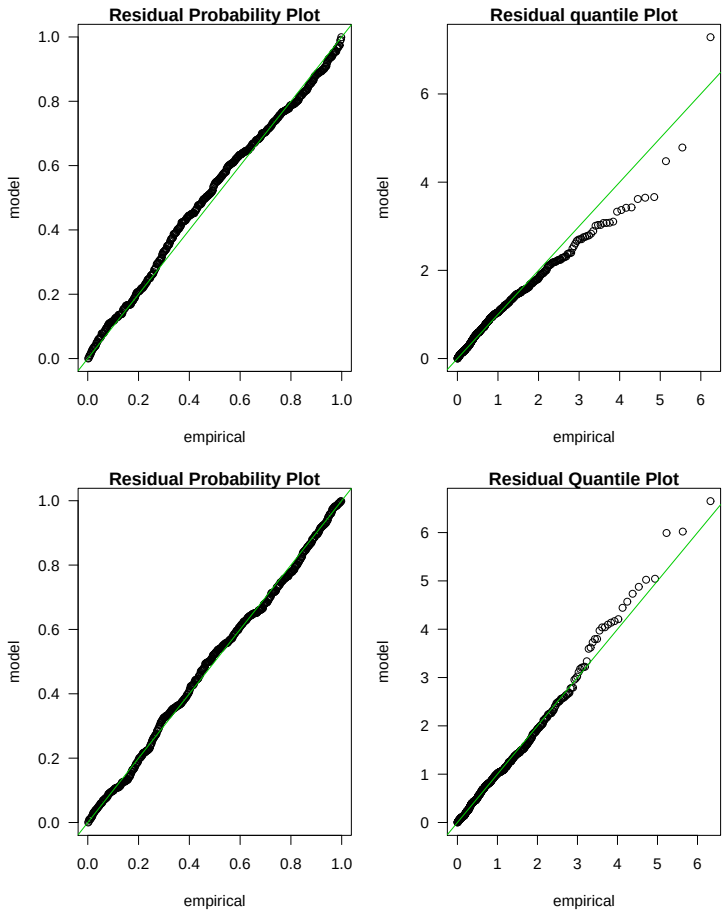

Fig. 3. Station Ihtiman: probability and quantile plots of the declustered precipitation data (in mm) based on a point process with homogeneous parameters (top panels), and with a seasonal cycle in the location parameters (bottom panels).

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

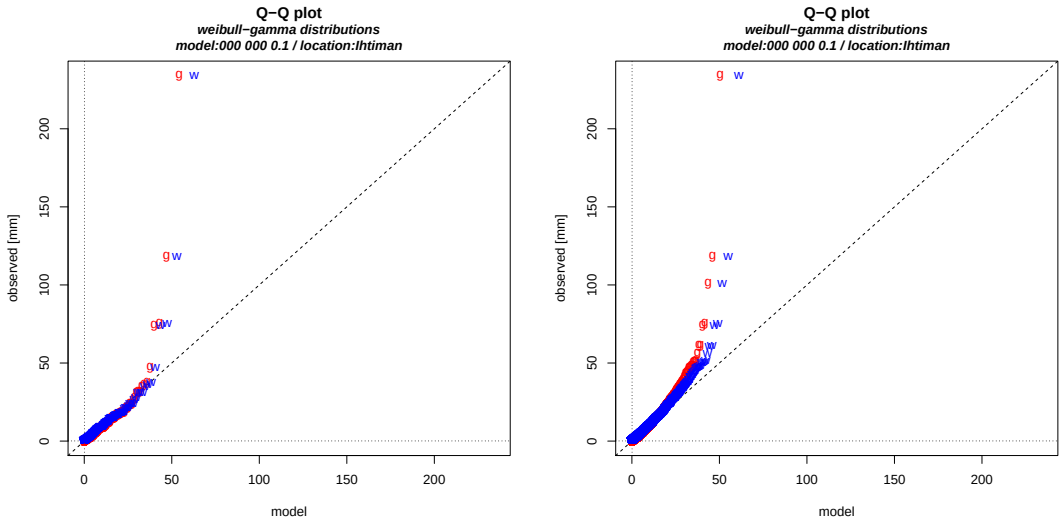

Fig. 4. Q-Q plots of observed vs. fitted gamma (g) and Weibull (w) quantiles of daily precipitation intensities (standard threshold of 0.1 mm) based on homogeneous model at Ihtiman in August (left plot) and for the entire year (right plot).

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

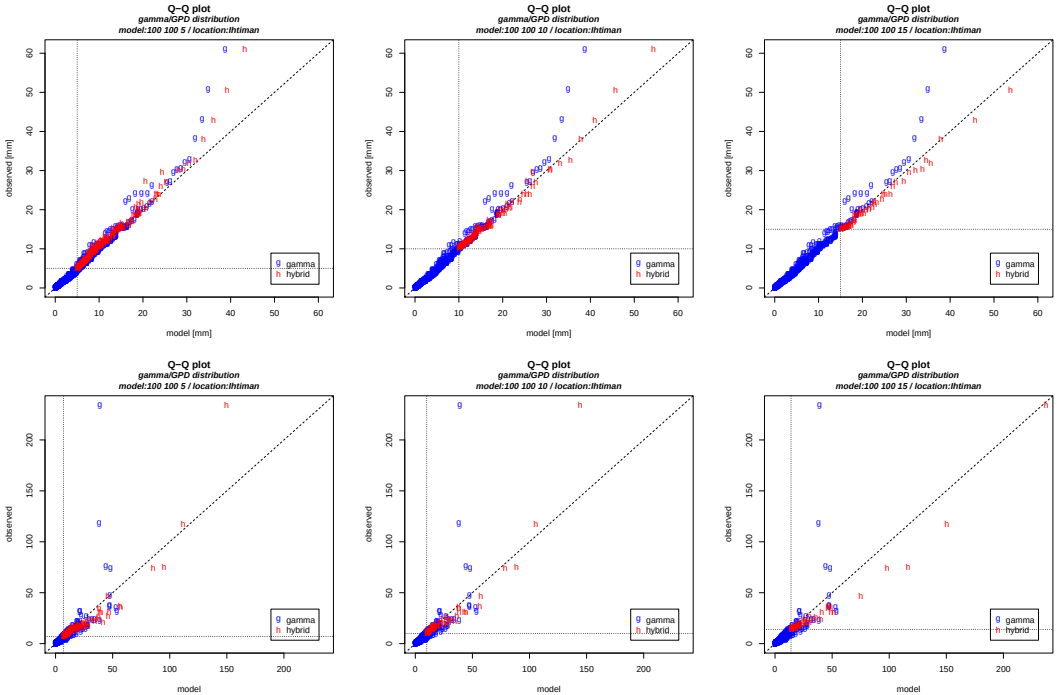

Fig. 5. Q-Q plots of observed vs. fitted gamma (g) and hybrid gamma–GP (h) quantiles of precipitation intensity seasonal model at Ihtiman with thresholds 5 mm (left), 10 mm (middle) and 15 mm (right) in May (top line) plots) and August (bottom line).

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

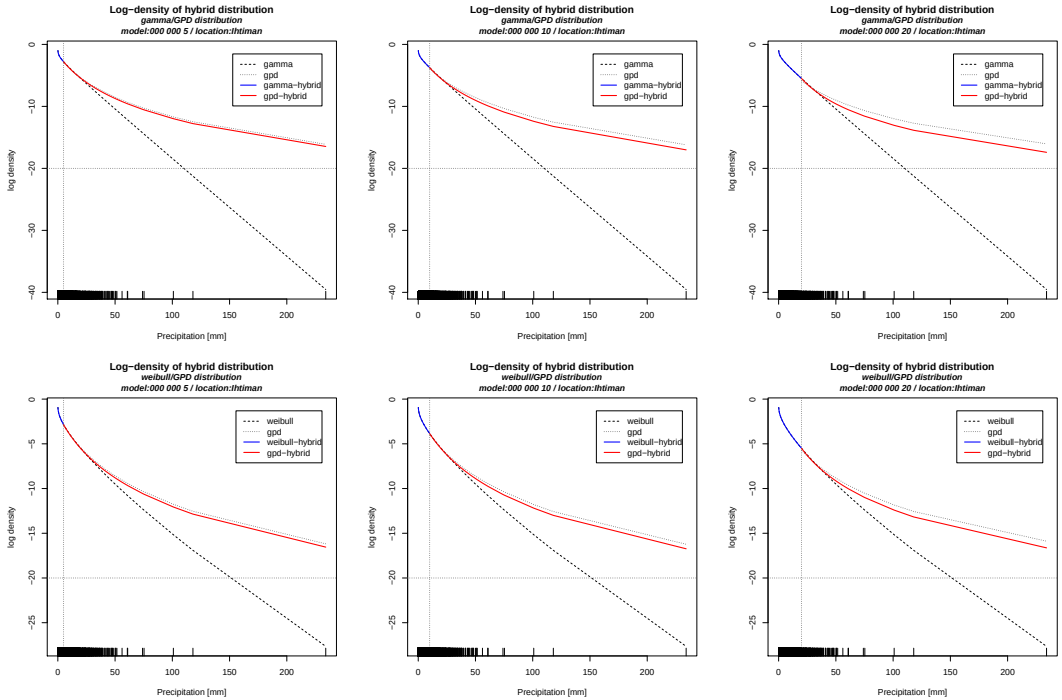

Fig. 6. Top panels: log-density functions fitted to daily precipitation intensity at Ihtiman with threshold values 5 mm, 10 mm and 20 mm. Gamma (solid and dashed lines), GP (dotted lines) and hybrid gamma–GP (solid blue and red lines) models are shown. The data are indicated by horizontal ticks and the threshold u by a vertical line. Bottom line: The same as the top row but based on the Weibull and hybrid Weibull–GP distributions.

NHESSD
2, 1223–1258, 2014

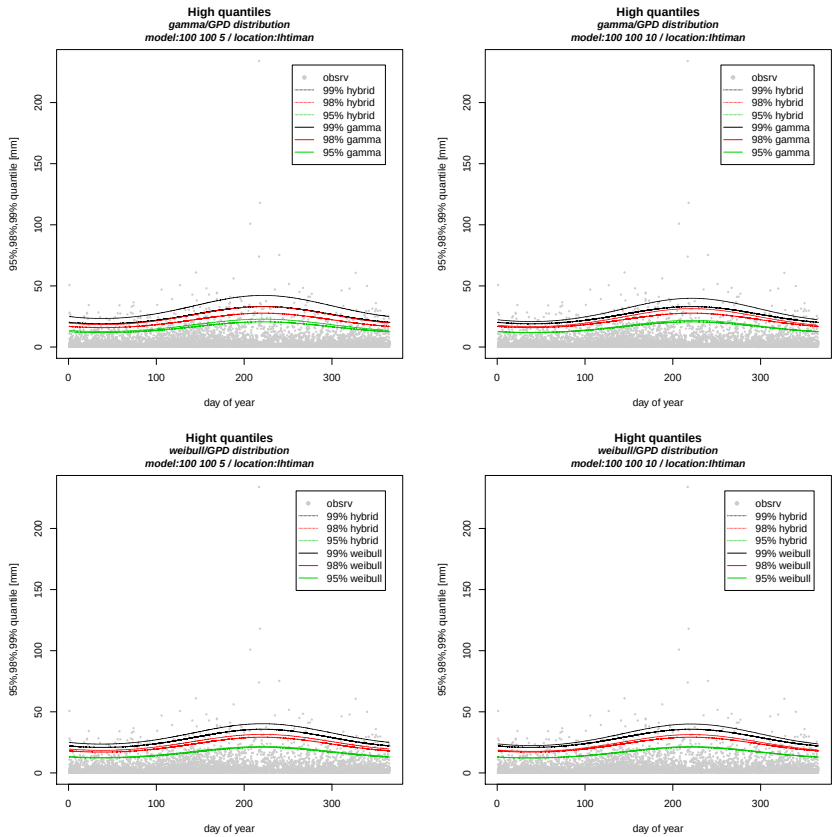

Fig. 7. Top row: high quantiles (95 %, 98 % and 99 %) of fitted gamma (solid lines) and hybrid gamma–GP (dashed lines) distributions as functions of the day of the year for Ihtiman precipitation intensity with thresholds 5 mm (left) and 10 mm (right). Lower line plots: the same as the top row but based on Weibull and hybrid Weibull–GP distributions.

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

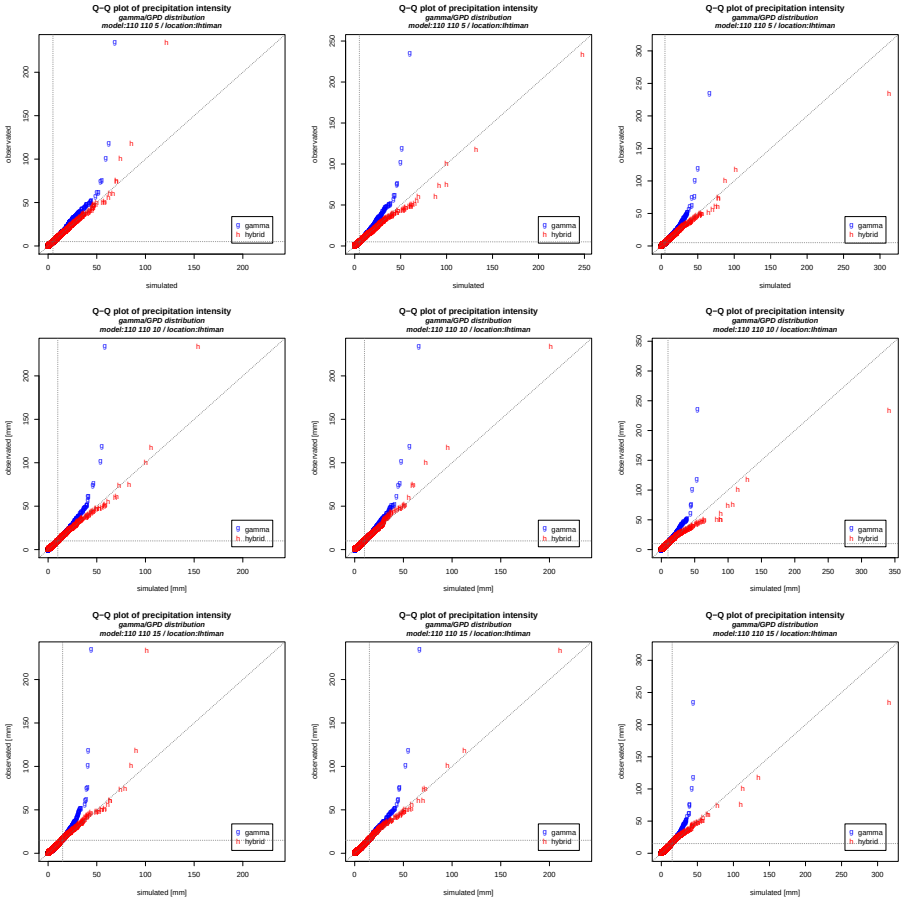

Fig. 8. Q-Q plots of observed vs. simulated gamma (g) and hybrid gamma–GP (h) quantiles of intensity seasonal models with NAO effect for the entire year at Ihtiman with 5 mm (top row), 10 mm (middle row) and 15 mm (bottom row) threshold.

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

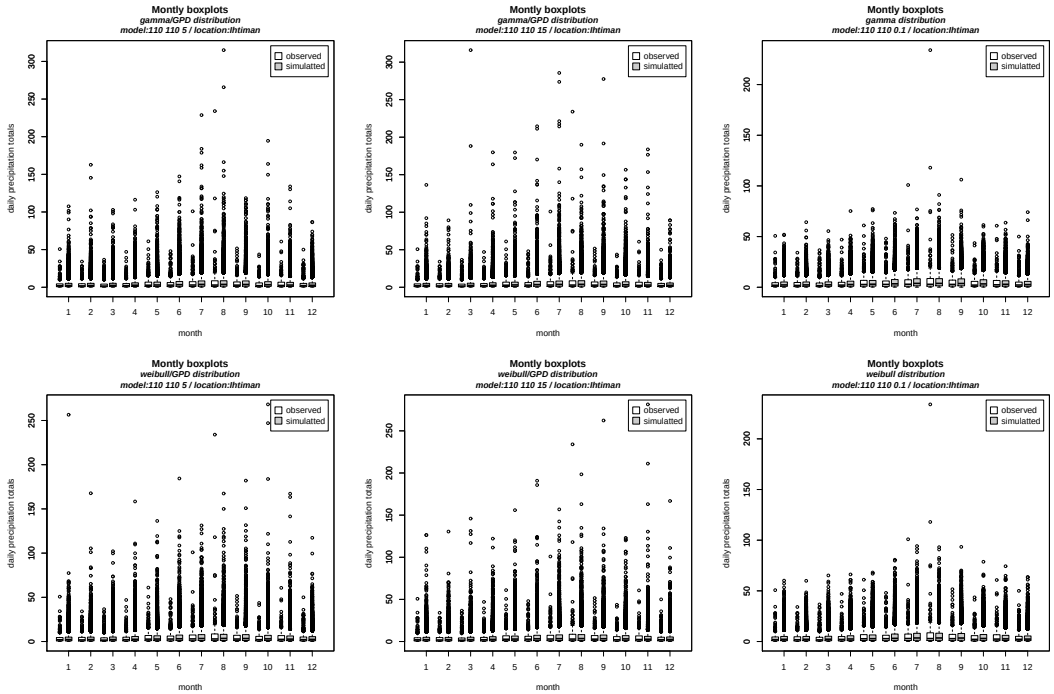

Fig. 9. Monthly box-plots of daily observed and simulated precipitation totals for Ihtiman station. The simulated data are generated using seasonal model with a lagged NAO and occurrence covariates; the intensity component is based on: (i) hybrid gamma–GP (top left and middle plots) and Weibull–GP (bottom left and middle plots) distribution with threshold values 5 mm and 15 mm; (ii) standard gamma and Weibull distributions (right column plots).

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.

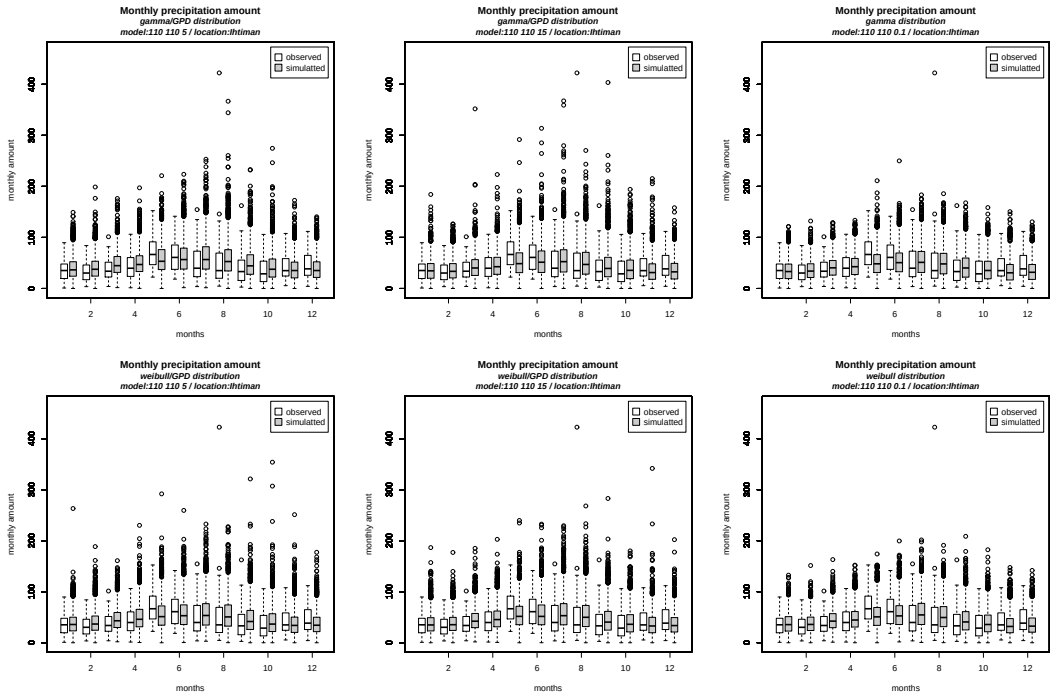

Fig. 10. Box-plots of monthly observed and simulated precipitation totals for Ihtiman station. The simulated data are generated using seasonal model with a lagged NAO and occurrence covariates; the intensity component is based on: (i) hybrid gamma–GP (top left and middle plots) and Weibull–GP (bottom left and middle plots) distributions with threshold values 5 mm and 15 mm; (ii) standard gamma and Weibull distributions (right column plots).

NHESSD
2, 1223–1258, 2014

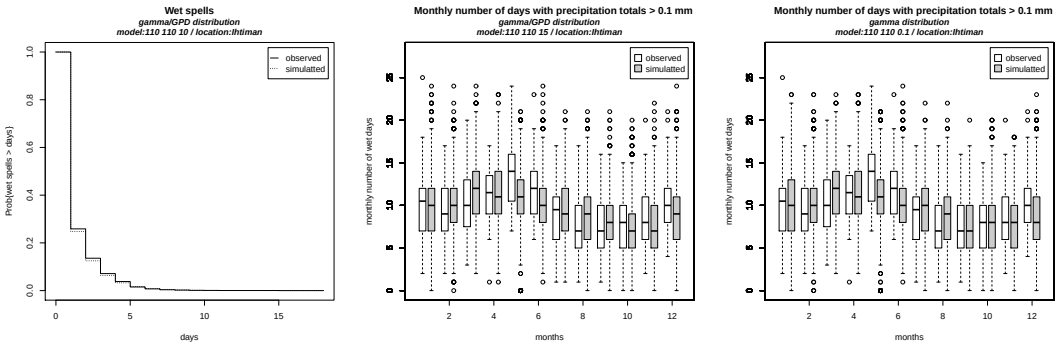

Fig. 11. Left plot: observed (solid lines) and model-based (dotted lines) wet spells distribution. Box-plots of monthly observed (white) vs. simulated (gray) number of wet days for Ihtiman station: the simulated data are generated using hybrid gamma–GPD with threshold 15 mm and standard gamma (right plot) distribution.

NHESSD
2, 1223–1258, 2014

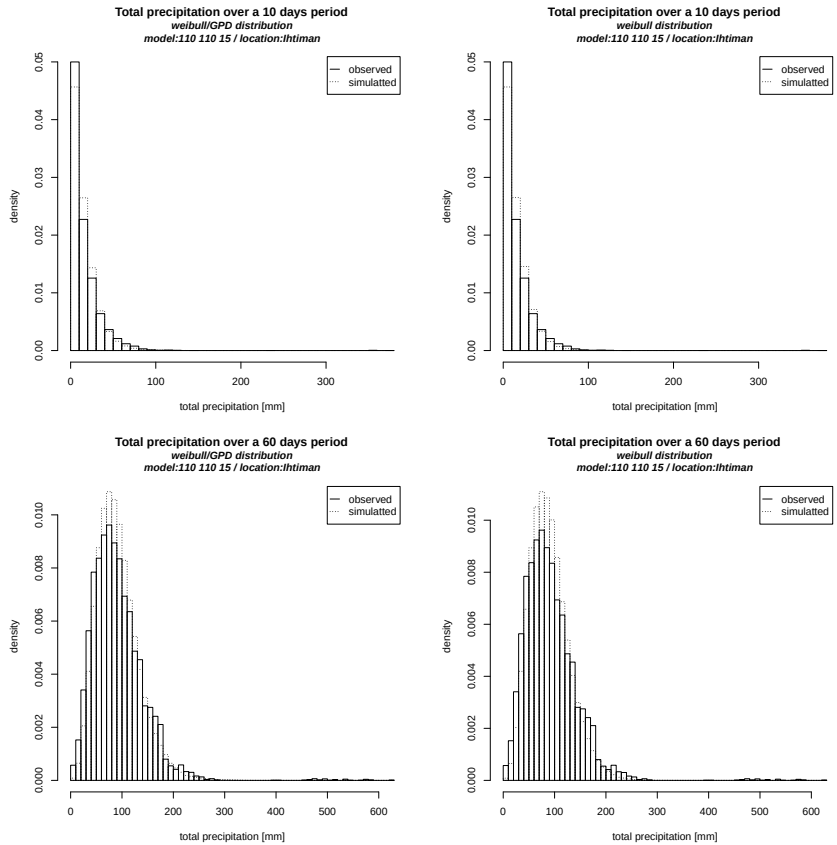

Fig. 12. The distributions of the observed (solid lines) and simulated (dotted lines) precipitation totals over period of 10 and 60 days. The simulated data are generated using intensity GLMs with hybrid Weibull–GPD and threshold 15 mm (left column) and standard Weibull (rigth column) distributions.

NHESSD
2, 1223–1258, 2014

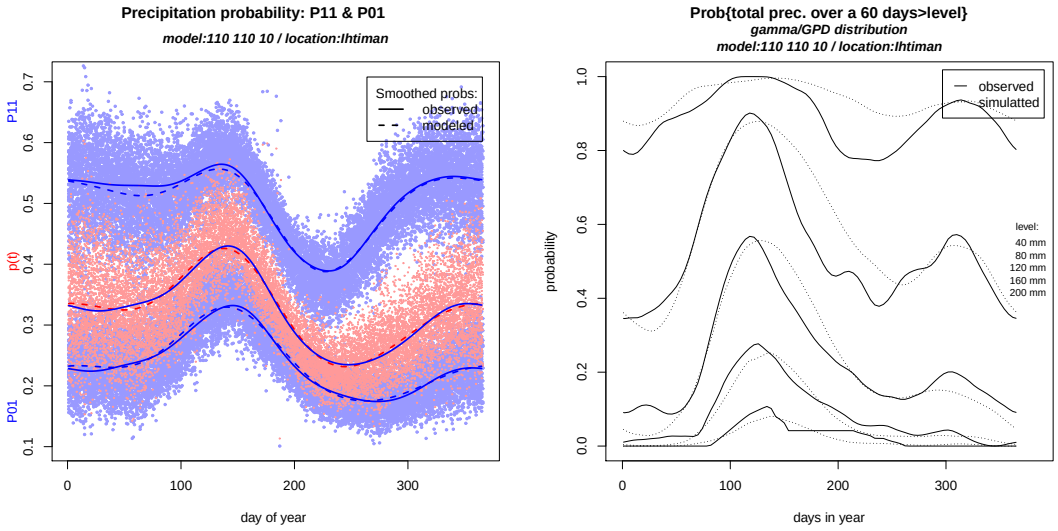

Fig. 13. Left plot: (i) the dots represent the estimated conditional probabilities p_{11} and p_{01} and the unconditional precipitation probability $p(t) := \pi_t(z_t)$ (red); the dashed and smoothed lines are based on the R locally weighted scatterplot smoothing procedure loess through the corresponding dots and observed frequencies (not plotted). Right plot: the historical and simulated probabilities (smoothed by loess) of having not less than 40 mm, 80 mm, 120 mm, 160 mm and 200 mm total precipitation for a run of 60 consecutive days, starting on any given day of the year for Ihtiman station.

NHESSD
2, 1223–1258, 2014

Stochastic daily precipitation model with a heavy-tailed component

N. M. Neykov et al.