Nat. Hazards Earth Syst. Sci., 14, 2321–2335, 2014
www.nat-hazards-earth-syst-sci.net/14/2321/2014/
doi:10.5194/nhess-14-2321-2014

Natural Hazards and Earth System Sciences
Open Access

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 Shose 66, Sofia 1784, Bulgaria
[2]Department of Economic Sciences, Georg-August-Universität, Göttingen, Germany

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

Received: 31 October 2013 – Published in Nat. Hazards Earth Syst. Sci. Discuss.: 10 February 2014
Revised: 26 June 2014 – Accepted: 23 July 2014 – Published: 3 September 2014

Abstract. Stochastic daily precipitation models are commonly used to generate scenarios of climate variability or change on a daily timescale. 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 using 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 modeling 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; see 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 precipitation at that site. These sequences can be used to estimate statistics relating to precipitation events in exactly the way one would do if a long sequence of real precipitation data were available. As a consequence, better risk management strategies and decision-making capabilities can be developed.

In the following, we shall consider precipitation models on daily timescales only. From a statistical point of view, daily precipitation totals are time series with mixed densities, 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 (GLMs) 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, lognormal or a mixture of exponential distributions. More precisely, modeling the occurrence series means modeling the transition probabilities of the two-state 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 the North Atlantic Oscillation, the El Nino–Southern Oscillation, pressure, humidity, temperature, and wind speed, or as slowly varying trend functions over the years. The occurrence (the state transition probabilities) and intensity model components therefore 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) and Fahrmeir and Tutz (2001). The properties and applicabilities of such models on different timescales are discussed by Brandsma and Buishand (1997), Katz and Parlange (1998), Grunwald and Jones (2000), Hyndman and Grunwald (2000), Beckman and Buishand (2002), Chandler and Wheater (2002), Chandler (2005), Yang et al. (2005) and Furrer and Katz (2007), to name a few. Reviews of 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 an estimated shape parameter and a 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 one. 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 that is built by splicing the GP distribution tail to a Gaussian or a truncated Gaussian distribution. 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.

2 Case study – Ihtiman data set

We analyzed 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 were 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 LT), 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 Stochastic daily precipitation models

Let Y_t be the precipitation on day t, and $\boldsymbol{Z}_t = (Z_{1t}, \ldots, Z_{kt})'$ is a vector of associated atmospheric variables or their derivatives for $t = 1, \ldots, T$. Day t is defined as 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 $p_t(\boldsymbol{x}_t)$ represent the probability that day t is wet, conditional on the vector of covariates $\boldsymbol{x}_t = (i_{t-1}, \ldots, i_{t-p}, y_{t-1}, \ldots, y_{t-p}, z_{1t}, \ldots, z_{kt})'$. 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 is missing otherwise, and denote its probability density function, conditional on the atmospheric predictors, by $q(r_t|\boldsymbol{x}_t)$. 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

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

In practice, $q_t(r_t|\boldsymbol{x}_t)$ 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 in extremes intensities, then the GP density can be used.

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

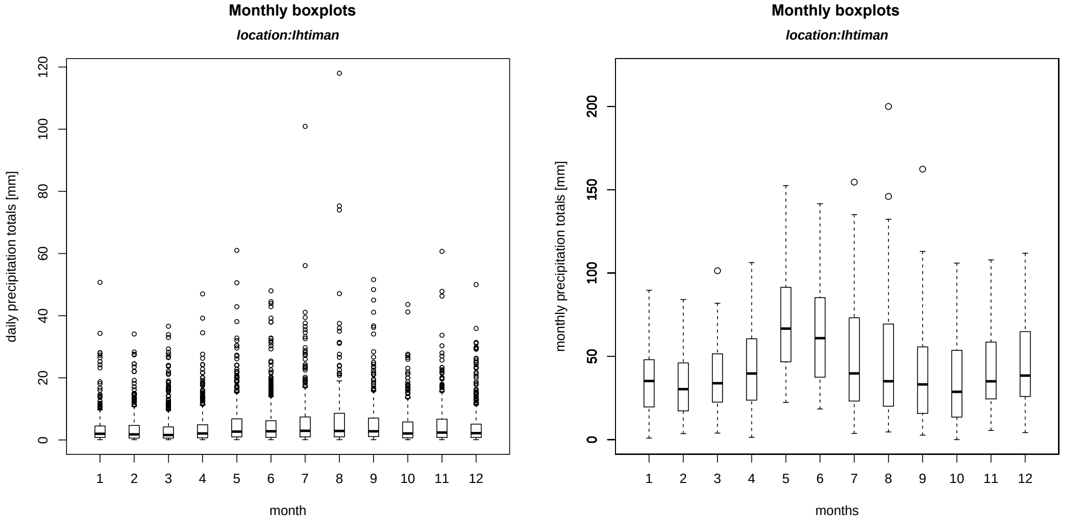

Figure 1. Boxplots of daily precipitation totals (left) shown by months, and monthly (right) precipitation totals. The extreme daily value 234 mm recorded on 5 August 2005 is dropped from the data.

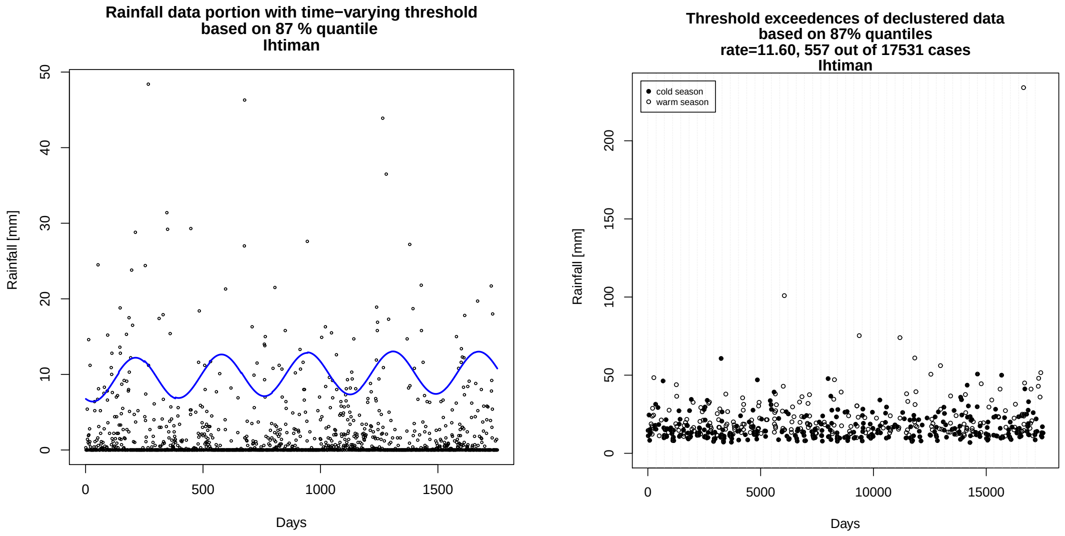

Figure 2. Portion of daily precipitation intensity data with a time-varying threshold (solid line) based on the 87 % quantile (left plot); exceedances based on declustered daily intensity data (right plot).

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

Standard GLMs software can be used to estimate the unknown parameters due to this factorization of the likelihood; the first product is the likelihood of the binary time series and the second product is the likelihood of the intensity time series. The *vglm* procedure from R package *VGAM* can fit such models (Yee and Stephenson, 2007). This general likelihood maximization procedure, based on an iterative reweighted least squares procedure, is applicable not only to standard GLMs, but also to generalized additive models (GAMs); see Hastie and Tibshirani (1990). Moreover, by this procedure, one can model extreme values easily using generalized extreme value (GEV, block maxima) and peaks over threshold

Table 1. Monthly thresholds in mm based on the 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

(GP) distributions, just like GLMs and GAMs; see Green (1984) and Coles (2001).

The standard approach is to model the probabilities $p_t(x_t)$ within GLMs with a logit-link function:

$$\begin{aligned}\operatorname{logit}(p_t(\boldsymbol{x}_t)) &= \log(p_t(\boldsymbol{x}_t)/(1-p_t(\boldsymbol{x}_t)))\\ &= u(\boldsymbol{x}_t) = \alpha_0 + \sum_{l=1}^{p}(\alpha_l i_{t-l} + g_l(y_{t-l}))\\ &+ \sum_{l=1}^{k} g_{p+l}(z_{lt}) + g_{p+k+1}(t).\end{aligned}$$

The function $u(\boldsymbol{x}_t)$ should be periodic and approximately sinusoidal in shape, in order to reflect the seasonal behavior 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. A simple logit link function, consisting of a seasonal cycle and lagged occurrence and NAO effects, is

$$\begin{aligned}\operatorname{logit}(p_t(\boldsymbol{x}_t)) &= \alpha_0 + \alpha_1 \mathrm{i}_{t-1} + \alpha_2 \mathrm{C}_t + \alpha_3 \mathrm{S}_t + \alpha_4 \mathrm{NAO}_{t-1}\\ &+ [\beta_2 \mathrm{C}_t + \beta_3 \mathrm{S}_t + \beta_4 \mathrm{NAO}_{t-1}]\mathrm{i}_{t-1},\end{aligned}$$

where $\mathrm{C}_t = \cos(2\pi t/365.25)$ and $\mathrm{S}_t = \sin(2\pi t/365.25)$. The covariate vector for this model is $\boldsymbol{x}_t = (1, \mathrm{i}_{t-1}, \mathrm{C}_t, \mathrm{S}_t, \mathrm{NAO}_{t-1}, \mathrm{C}_t\mathrm{i}_{t-1}, \mathrm{S}_t\mathrm{i}_{t-1}, \mathrm{NAO}_{t-1}\mathrm{i}_{t-1})'$. Due to the lagged occurrence and the related interactions included in this logit link function, the conditional two-state non-stationary transition probabilities of a wet day following a dry day, $p_{01}(t) = p_t(\boldsymbol{x}_t)$ for $i_{t-1} = 0$, and a wet day following a wet day, $p_{11}(t) = p_t(\boldsymbol{x}_t)$ for $i_{t-1} = 1$, are allowed different cyclic behaviors in the model. In this way, the parameter estimates of these probabilities can be computed from $p_t(\boldsymbol{x}_t)$ in one run instead of formulating two separate models and the respective data set according to Furrer and Katz (2007). Moreover, based on the total probability formula, one can get the following relationship between the unconditional of previous state probability $\pi(t) = \Pr(I_t = 1|z_t)$ and the two transition probabilities $\pi(t) = \pi(t-1)p_{11}(t) + (1-\pi(t-1))p_{01}(t)$. This representation is very useful in the simulation of artificial rainfall sequences because of the recurrence relationship. Indeed, under the plausible $\pi(t) \approx \pi(t-1)$ for any t, we get $\pi(t) \approx p_{01}(t)/(p_{01}(t) + 1 - p_{11}(t))$; see Zucchini et al. (1992) and Furrer and Katz (2007).

The intensities can be modeled by gamma, Weibull or other right-skewed continuous distributions, 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, and $a > 0$ and $b > 0$ are the shape and rate parameters. The mean, the variance and the scale parameters of the gamma distributions are given by $\mu = a/b$, $\sigma^2 = \mu^2/a$ and $1/b$.

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 $\xi = 0$.

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

$$\begin{aligned}&\log(a) = \theta_1^T \boldsymbol{x}_{1t},\ \log(b) = \theta_2^T \boldsymbol{x}_{2t},\\ &\log(\sigma) = \theta_3^T \boldsymbol{x}_{3t},\ \xi = \theta_4^T \boldsymbol{x}_{4t},\end{aligned}$$

where θ_i is a vector of unknown parameters, and the covariate vector $\boldsymbol{x}_{\mathrm{it}}$ 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

$$\begin{aligned}\log v(\boldsymbol{x}_t) &= u_1(\boldsymbol{x}_t) = \theta_0 + \sum_{l=1}^{p}\left(\theta_p i_{t-l} + h_l(y_{t-l})\right)\\ &+ \sum_{l=1}^{k} h_{p+l}(z_{lt}) + h_{p+k+1}(t),\end{aligned}$$

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	C_t	S_t	NAO	intercept	C_t	S_t	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

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

where the function $u_1(\boldsymbol{x}_t)$ has to be similar to $u(\boldsymbol{x}_t)$, and $h_{p+1},\ldots,h_{p+k+1}$ have to be smooth functions. A simple log-link function, consisting of a seasonal cycle and lagged occurrence and NAO effects, is

$$\log v(x_t) = \theta_0 + \theta_1 \mathrm{i}_{t-1} + \theta_2 \mathrm{C}_t + \theta_3 \mathrm{S}_t + \theta_4 \mathrm{NAO}_{t-1}.$$

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 behavior 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

Figure 1 shows boxplots of daily precipitation totals by month (left) and monthly precipitation totals at Ihtiman (right). In order to get an impression of the precipitation at this location, the extreme daily value 234 mm recorded on 5 August 2005 is excluded from this figure. The second highest daily precipitation total is 120 mm, which is about half of the above value. Moreover, the highest monthly precipitation total (right panel plot) is less than this extreme value. Seasonality is evident from these plots. A time-varying threshold based on the 87 % intensity quantile is displayed in the left panel of Fig. 2. The smooth curve is estimated using a quantile regression model with inter-annual and seasonal periodic (sine–cosine) components of the daily intensities. The *qr* procedure from R package *quantreg* is used for this purpose. Details about quantile regression can be found in Koenker (2005). Threshold selection is based on a range (80–95 %)

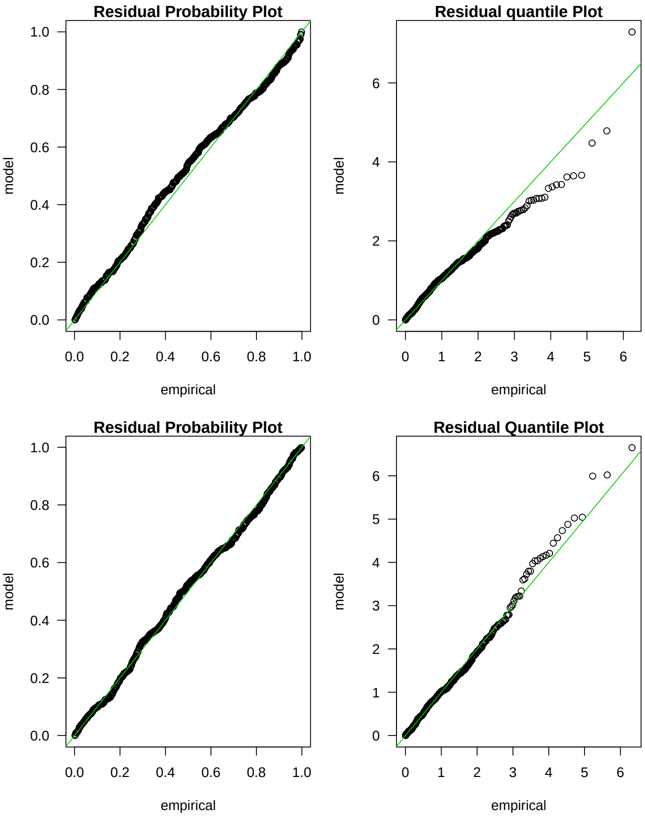

Figure 3. Probability and quantile plots of the declustered precipitation data (in mm) based on a point process with homogeneous parameters (top panels), and with a seasonal cycle in the location parameters (bottom panels).

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 rate $\log(a)$						Weibull scale $\log(\sigma)$					
Intercept	C_t	S_t	NAO	Shape b	BIC	Intercept	C_t	S_t	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**

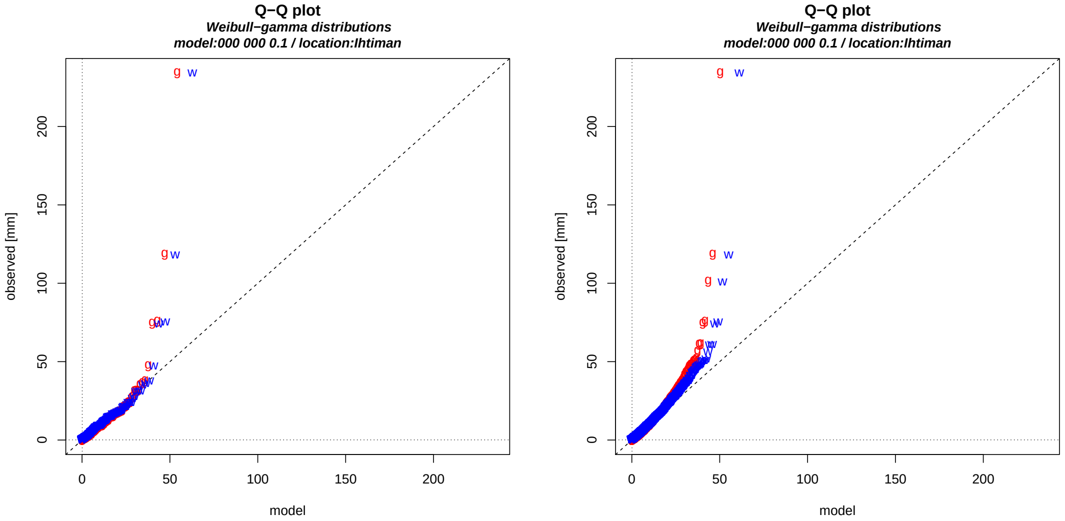

Figure 4. Q–Q plots of observed versus fitted gamma (g) and Weibull (w) quantiles of daily precipitation intensities (standard threshold of 0.1 mm), based on a homogeneous model in August (left plot) and for the entire year (right plot).

of thresholds for which the resulting GP distribution parameter estimates do not change considerably, and at the same time, the fit provides a reasonable model approximation; see the next paragraph for details. 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 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-day 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 a rate of 11.60 in 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-month intensities, respectively. This plot exhibits higher precipitation intensities during the warmer months. These extreme intensities are fitted using a point process model, as in Coles (2001). The advantage of the point process approach is that unifies the classical block maxima (GEV) and peaks over threshold (GP distribution) 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 for non-stationarity effects. The parameter estimates and Bayesian information criterion (BIC) values for several model fits are presented in Table 2. The main message from the results of this table is that the distribution of the extreme precipitation intensity at station Ihtiman has a heavy tail. This is supported by the likelihood ratio test (LRT) concerning the hypothesis $\xi = 0$ (exponential distribution) versus $\xi > 0$ (GP distribution); for all models, the corresponding tail probability is quite small. We note that not all precipitation data series over the territory of Bulgaria exhibit heavy upper tails, especially for sites with relatively short records. This is in agreement with the main conclusions of Papalexiou et al. (2013). The non-stationary model that minimizes the BIC includes a seasonal cycle and a lagged NAO effect in the location parameter, and 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 precipitation intensities synchronized with negative NAO index anomalies. On the other hand, higher scale parameter estimates are associated with

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

logit($p_t(x_t)$) intercept	C$_t$	S$_t$	I_{t-1}	NAO	C$_t I_{t-1}$	S$_t I_{t-1}$	NAOI_{t-1}	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
−1.186	−0.124	0.296	1.218	−0.113	0.260	−0.127	0.207	**20 579.20**

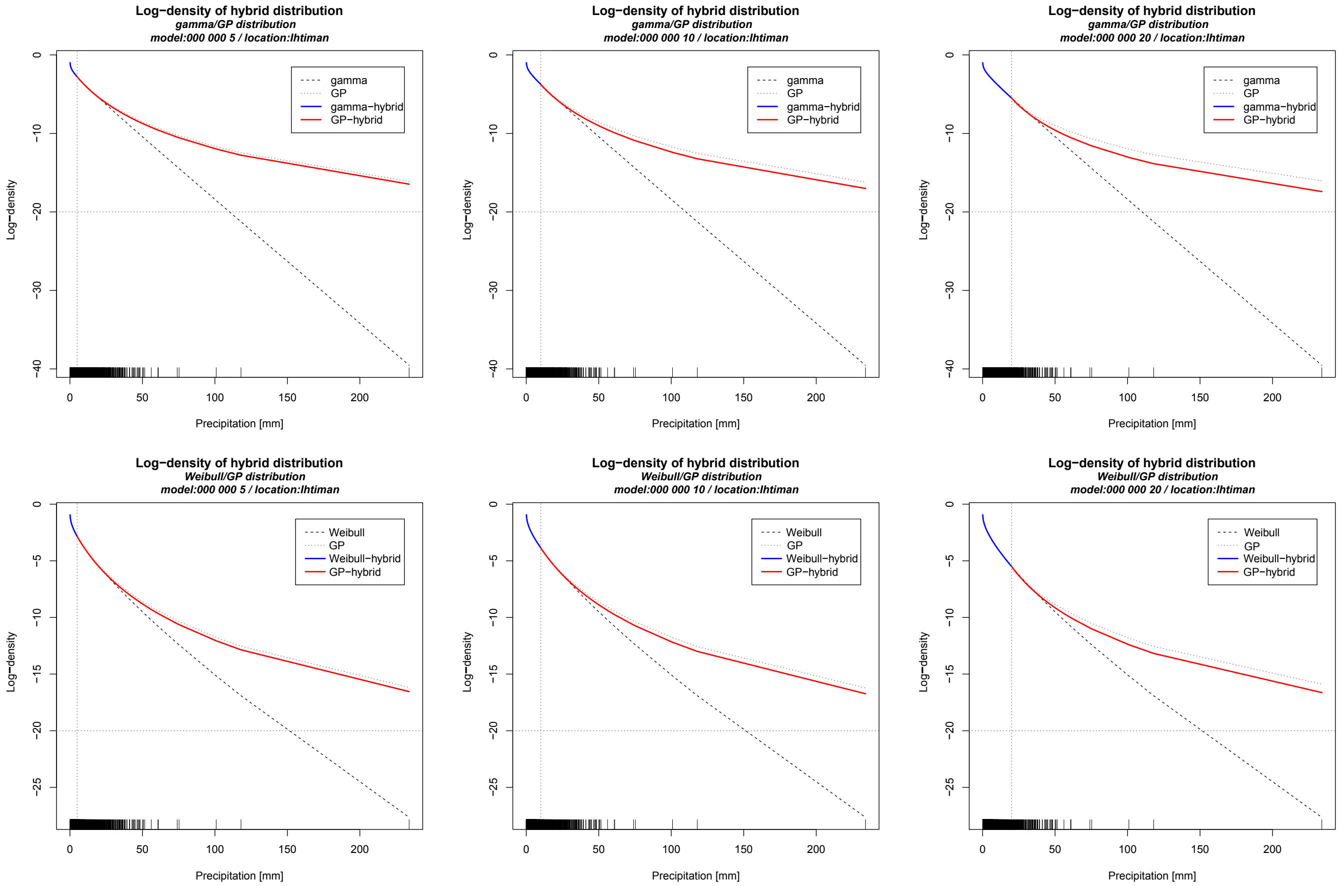

Figure 5. Top panels: log-density functions fitted to daily precipitation intensity with threshold values 5, 10 and 20 mm. Gamma (solid and dashed lines), GP (dotted lines) and hybrid gamma–GP (solid blue and red lines) models are shown. The data are indicated by horizontal ticks, and the threshold u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

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 in Fig. 3. The plots indicate reasonable but by no means perfect fits, and that the non-stationary fits are 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 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. Each model can be identified by its unique abbreviation code “xxx yyy zz” at the top of the individual panel of the corresponding figure. The strings xxx, yyy and zz indicate the occurrence, intensity and threshold

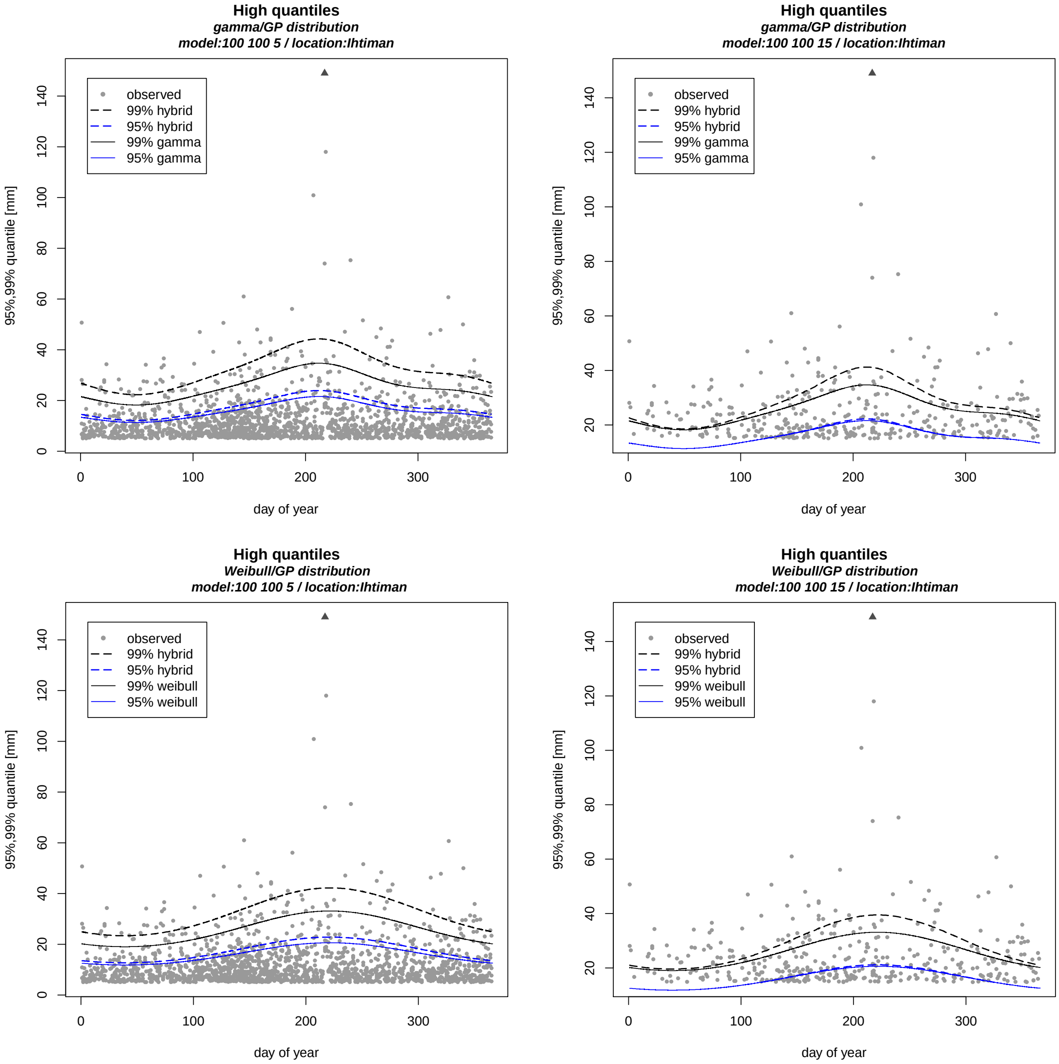

Figure 6. Top row: high quantiles (95 and 99 %) of fitted gamma (solid lines) and hybrid gamma–GP (dashed lines) distributions as functions of the day of the year for Ihtiman precipitation intensity with thresholds 5 mm (left) and 15 mm (right). Lower-line plots: the same as the top row, but based on Weibull and hybrid Weibull–GP distributions.

value model components, respectively. The coding depends on the types of covariates in the model components. For instance, (i) the model without covariates (homogeneous model) is coded as 000 000, (ii) the model with sine and cosine waves and previous day occurrence is coded as 100 100, and (iii) the model (ii) with additional lag of the NAO effect is coded as 110 110. This way, we get various gamma and Weibull precipitation models.

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 the gamma and Weibull models. According to the LRT, the models with seasonal components lead to improvements in comparison with homogeneous models, and the inclusion of NAO effects leads to further improvements. (The LRT and its tail probability values are not presented.) Both models preserve the 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 right 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 completely inadequate, but one can see the BIC value reduction, with the remaining models conditional on seasonal cycle, previous day precipitation occurrence and NAO effect with lag one.

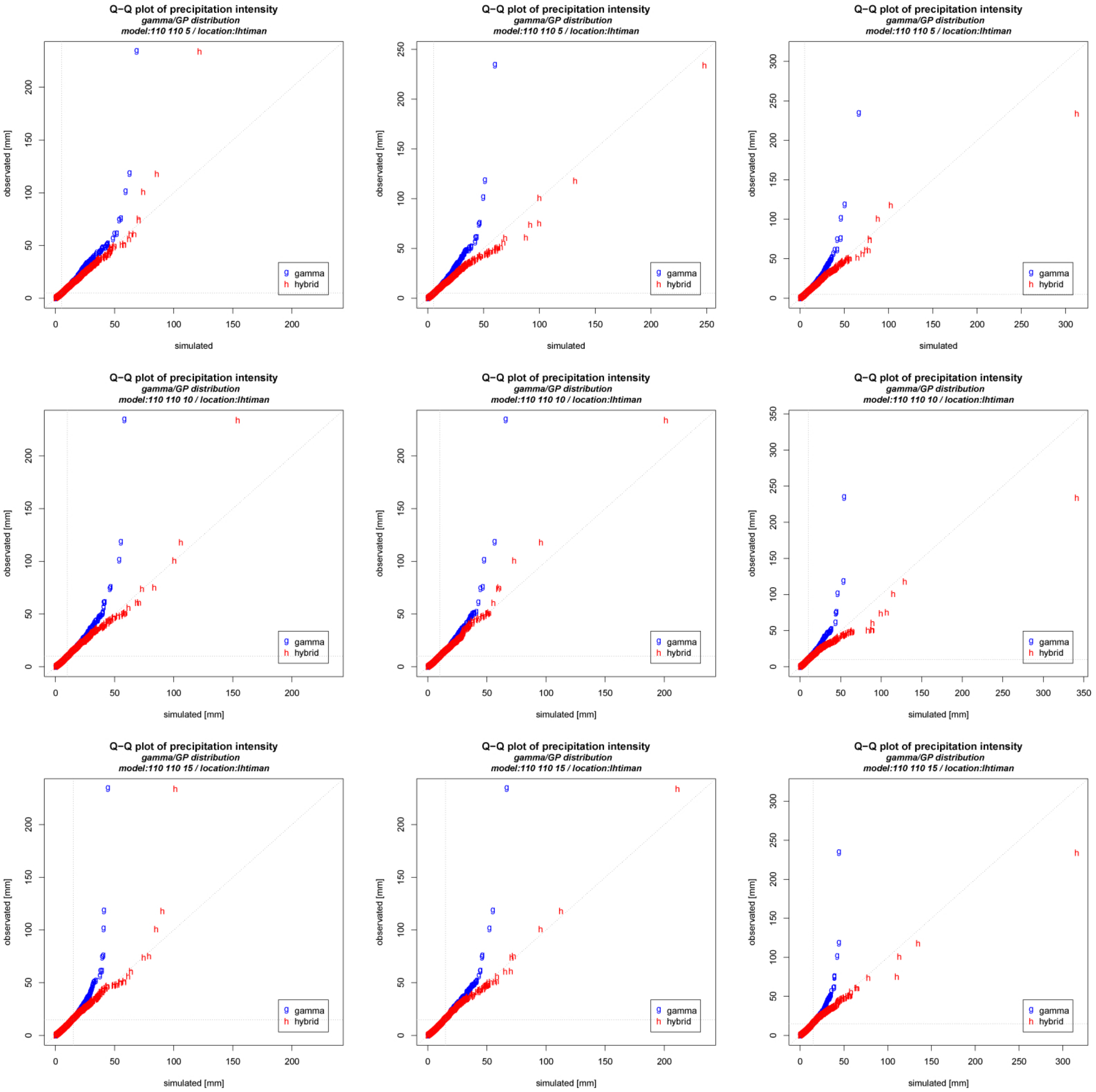

Figure 7. Q–Q plots of simulated versus observed gamma (g) and hybrid gamma–GP (h) quantiles of intensity seasonal models with the NAO effect for the entire year at Ihtiman, with 5 mm (top row), 10 mm (middle row) and 15 mm (bottom row) thresholds.

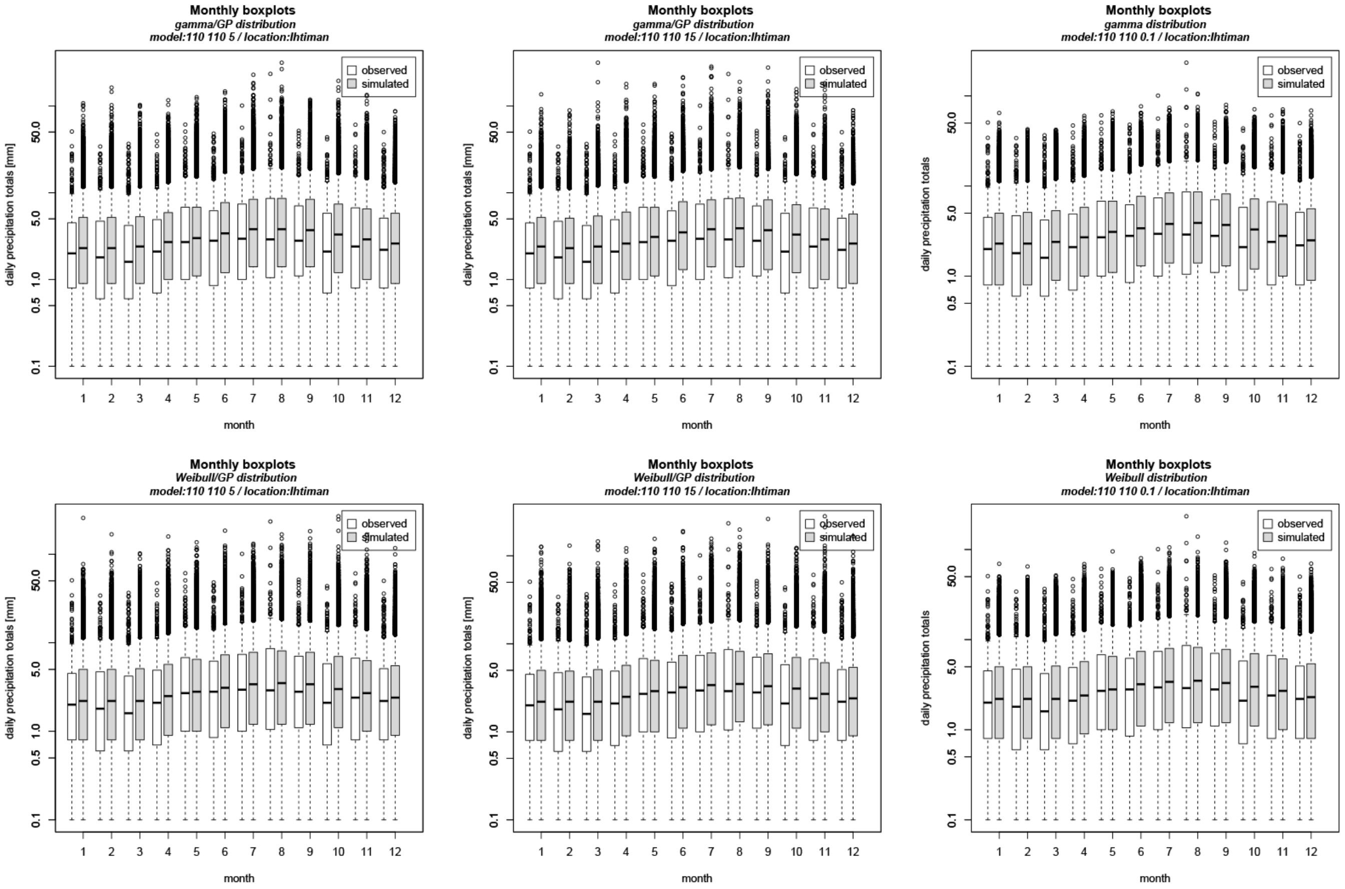

Figure 8. Monthly boxplots of daily observed and simulated precipitation totals in log scale. The simulated data are generated using a seasonal model with a lagged NAO covariate; the intensity component is based on (i) hybrid gamma–GP (top-left and middle plots) and Weibull–GP (bottom-left and middle plots) distributions with threshold values 5 and 15 mm, and (ii) standard gamma and Weibull distributions (right-column plots).

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 distribution function, $f(x)$ and $g(x)$ are the gamma and GP densities, 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 GP distribution 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. The GP distribution scale parameter can thus be written in terms of the parameters of the gamma distribution that accommodates the observations below the threshold. The authors recommend the following estimation procedure: (i) fit a GLMs with a gamma link function with covariates to the entire intensity data set; (ii) fit a GP distribution with covariates to the observations above the prespecified threshold u, just as in the conventional extreme value methodology; and (iii) replace the GP distribution scale parameter with the estimated gamma hazard function. Clearly, an analogous procedure is applicable for other hybrid distributions, such as Weibull–GP or inverse Gaussian–GP.

5.1 Experiments with hybrid GP distributions

As in the previous section, the models discussed hereafter can be identified by their unique code. However, gamma and Weibull intensity model components are combined with the GP distribution at the prespecified threshold values, e.g., 5, 10 and 15 mm. This way, we consider several hybrid gamma–GP and Weibull–GP distribution models. We now compare the fit of these models to Ihtiman 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.

Figure 5 shows the fitted gamma, GP, and hybrid gamma–GP (upper-line plots), and Weibull, GP and Weibull–GP (lower-line plots) log densities, with three threshold values 5, 10 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. One can thus expect that the hybrid distribution quantiles corresponding to these lower-threshold GP fits would be larger. The plots of Fig. 6 show the 95 % (blue) and 99 % (black) quantiles of the fitted gamma (solid lines) and hybrid gamma–GP (dashed lines) distributions (upper-line plots), as well as the fitted Weibull and hybrid Weibull–GP (lower-line plots) distributions as functions of the day of the year. The tiny black bullets in the plots correspond to observed precipitation totals exceeding the prespecified threshold. Indeed, the hybrid quantiles (dashed lines) in the left-column plots are higher than those in the right-column plots. The effect due to hybridization is most visible in the highest quantiles. Therefore, threshold determination is of crucial importance to calibration daily precipitation hybrid models, conditional on atmospheric covariates.

Figure 7 shows Q–Q plots of observed versus simulated gamma (g) and 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. Results of a similar standard are obtained for the Weibull and Weibull–GP distributions. The hybrid models are a significant improvement over the gamma and Weibull models.

The monthly boxplots of daily observed (white) and simulated (gray) precipitation totals given on logarithmic scales are presented in the plots of Fig. 8 in order to get an impression of their distributions: the left and middle column panels are based on the hybrid intensity distributions, with 5 and 15 mm thresholds where the right column panels are based on classical GLMs with the gamma and Weibull distributions. The classical GLMs with gamma and Weibull intensity components represent the historical data, except for the extreme intensities, which is a 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 in the plots of Fig. 9. 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 magnitudes as the historical ones, whereas the hybrid gamma and Weibull–GP distributions are capable of doing so. The distribution functions of the wet spells and the number of wet days within a season are important in applications in various studies. The left plot of Fig. 10 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 versus simulated data. Results of a similar standard are obtained for the Weibull and Weibull–GP distributions.

The distributions of the observed (solid lines) and simulated (dotted lines) precipitation totals over periods of 10 and 60 days are shown in the plots of Fig. 11. The simulated data are generated using intensity GLMs with hybrid Weibull–GP distributions and a threshold of 15 mm (left-column plots), and standard Weibull (right-column plots) distributions. It is seen that the hybrid Weibull–GP distribution-simulated data possess a heavier tail than the standard Weibull distribution. Similar results are obtained for the gamma–GP hybrid distribution.

The left-plot dots of Fig. 12 show the different seasonal cycles as well as the different magnitudes of the estimated precipitation probabilities $p_{11}(t)$ and $p_{01}(t)$, and the unconditional precipitation probability $p(t) := \pi(t)$ (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, 80, 120, 160 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 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 the *VGAM* package with gamma, Weibull and GP distribution link functions was used to carry out the estimation (Yee and Stephenson, 2007).

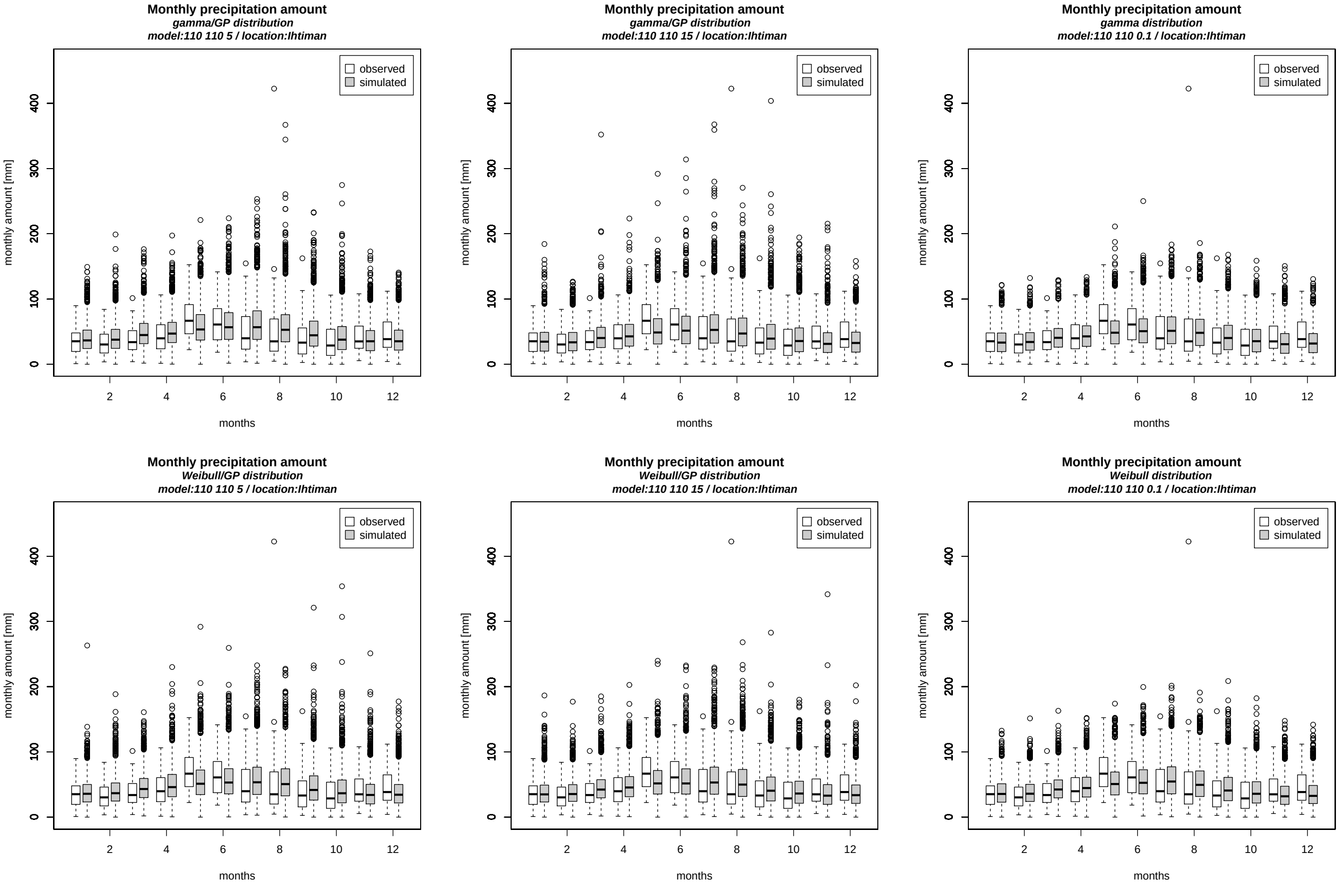

Figure 9. Boxplots of monthly observed and simulated precipitation totals. The simulated data are generated using a seasonal model with a lagged NAO covariate; the intensity component is based on (i) hybrid gamma–GP (top-left and middle plots) and Weibull–GP (bottom-left and middle plots) distributions with threshold values of 5 and 15 mm, and (ii) standard gamma and Weibull distributions (right-column plots).

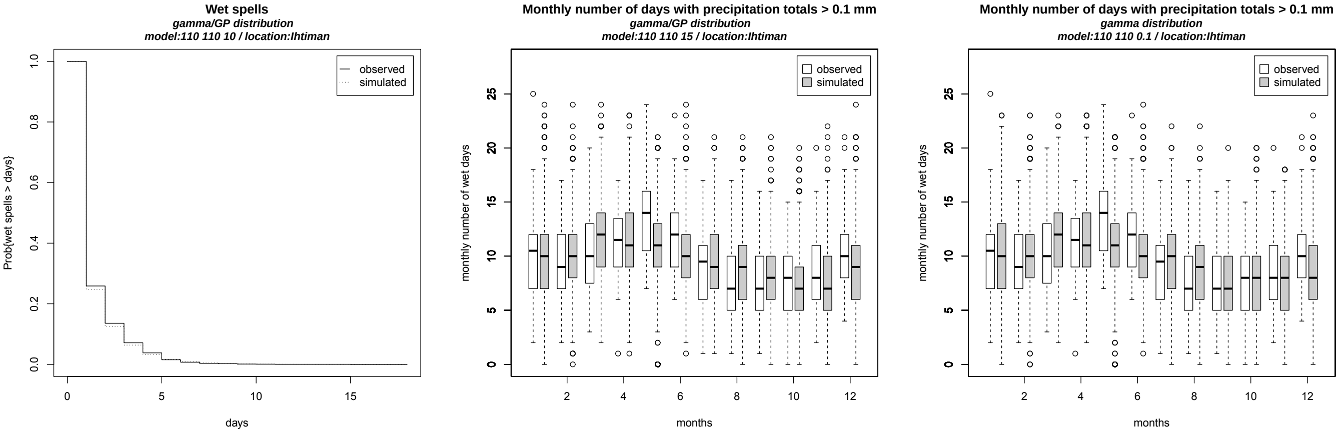

Figure 10. Left plot: observed (solid lines) and simulated (dotted lines) wet spell distribution. Box-plots of monthly observed (white) versus simulated (gray) number of wet days: the simulated data are based on the hybrid gamma–GP distribution with threshold 15 mm and standard gamma (right plot) distribution.

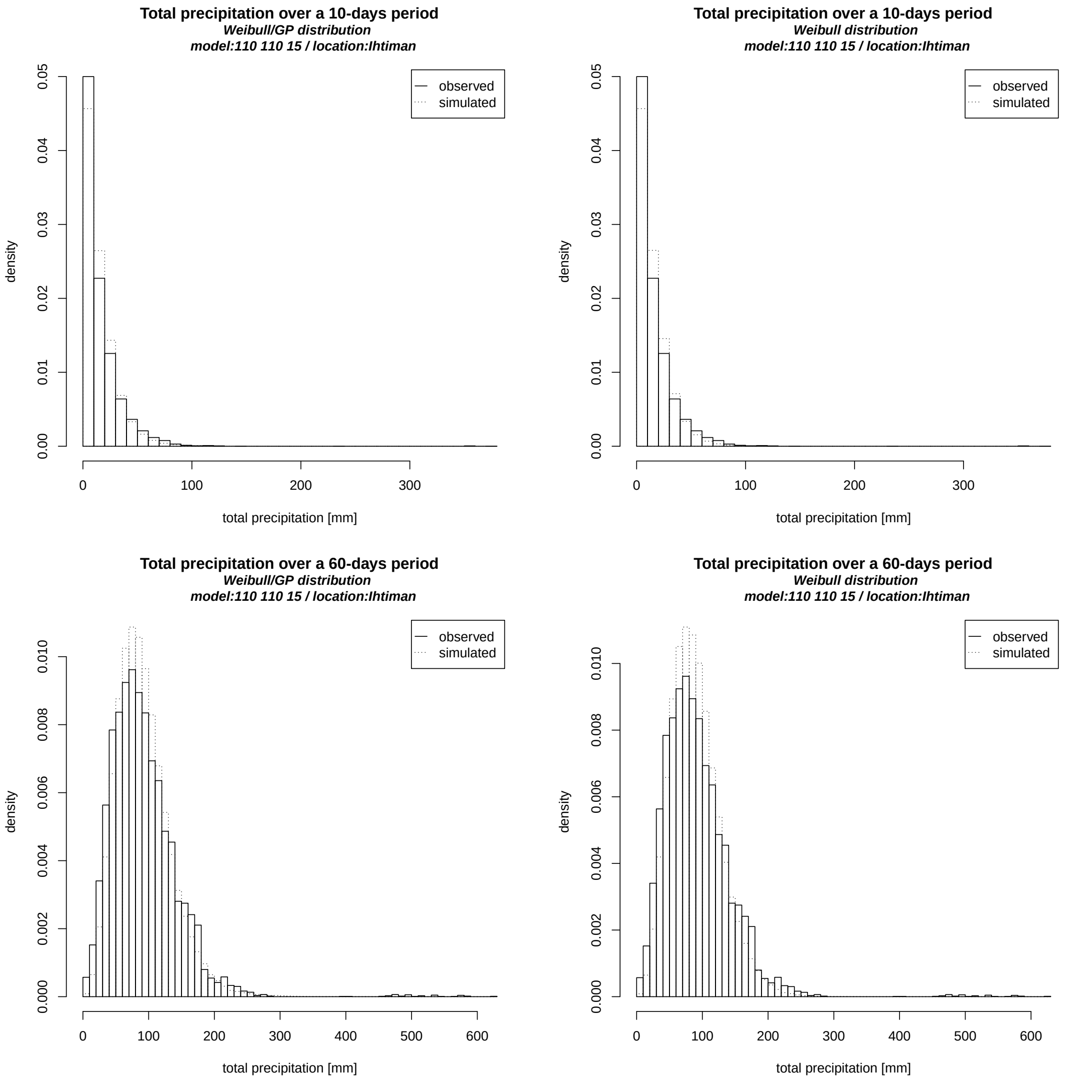

Figure 11. The distributions of the observed (solid lines) and simulated (dotted lines) precipitation totals over periods of 10 and 60 days. The simulated data are generated using intensity GLMs with hybrid Weibull–GP distributions and threshold 15 mm (left column) and standard Weibull (right column) distributions.

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 the threshold selection in splicing the distributions is an awkward task. The development of a daily precipitation model with such distributions, conditional on a large number of atmospheric predictors for downscaling purposes, is thus still in its early stages. Once this problem has been 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 basis of the conditional independence precipitation amount model within the non-homogeneous hidden Markov model framework; see Vrac and Naveau (2007) and Neykov et al. (2012).

The hybrid distribution daily precipitation model we discussed so far can be extended to model hourly precipitation data. In order to account for various dependencies, the precipitation model occurrence and intensity link functions have to incorporate some additional finite Fourier series, with harmonics and appropriate autoregressive covariates varying over the daily hours. Also, threshold selection must be performed on an hourly timescale. Essentially, this would be an adaptation of the methodology proposed by

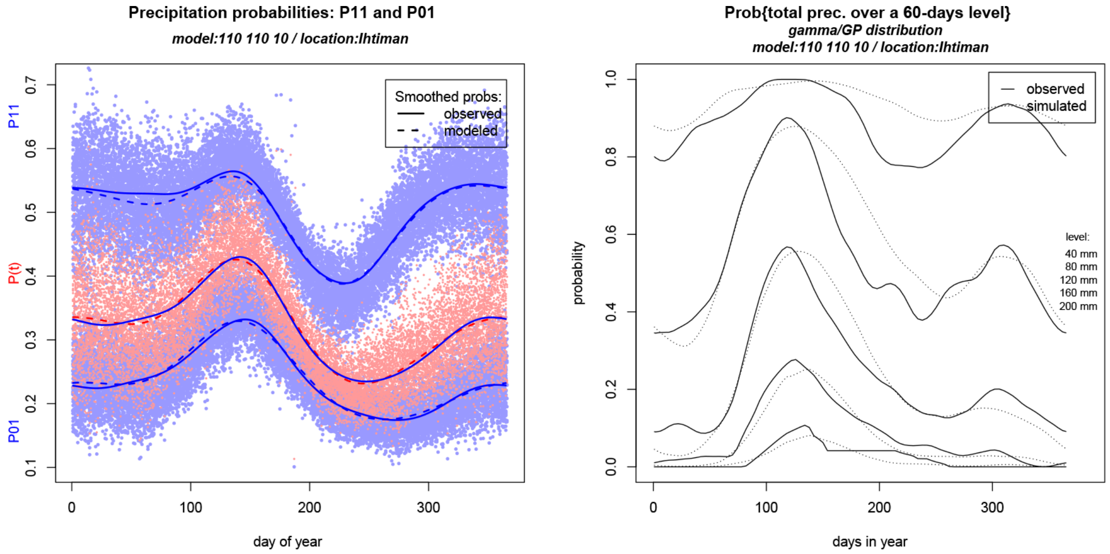

Figure 12. Left plot: (i) the dots represent the estimated probabilities p_{11} (blue), $p(t) := \pi(t)$ (red) and p_{01} (blue); the dashed and smoothed lines are based on the *loess* procedure through the corresponding dots and observed frequencies (not plotted). Right plot: the observed and simulated probabilities (smoothed by *loess*) of having not less than 40, 80, 120, 160 and 200 mm total precipitation for a run of 60 consecutive days, starting on any given day of the year.

Katz and Parlange (1995) and Chappell et al. (2009). However, estimation problems might arise when hourly precipitation data series are short. Finally, in order to achieve finer temporal resolution, the at-site hybrid GLMs on a daily precipitation scale can be combined with a single-site disaggregation model based on Poisson cluster processes. For instance, simulations of long sequences of sub-daily precipitation data can be obtained from hybrid GLMs-simulated daily precipitation totals using the HYETOS software; see Koutsoyiannis and Onof (2001).

Acknowledgements. The research was funded partially by the Deutsche Forschungsgemeinschaft, grant GZ: ZU 237/1-1, the Bulgarian Academy of Sciences, and the European Science Foundation within the framework of COST Action ES0901 "European Procedures for Flood Frequency Estimation" and COST ES1102 "Validating and integrating downscaling methods for climate change research".

Edited by: T. Kjeldsen
Reviewed by: two anonymous referees

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.

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

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.

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.

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

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.

Chappell, N. A., Discenza, A. R., Tych, W., Whittaker, J. and Bidin, K.: Simulating hourly rainfall occurrence within an equatorial rainforest, Borneo Island, Hydrol. Sci. J., 54, 571–581, 2009.

Coe, R. and Stern, R. D.: Fitting models to daily rainfall data, J. Appl. Meteorol., 21, 1024–1031, 1982.

Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer, New York, 2001.

Fahrmeir, L. and Tutz, G.: Multivariate Statistical Modeling Based on Generalized Linear Models, Springer, New York, 1994.

Furrer, E. M. and Katz, R. W.: Generalized linear modeling approach to stochastic weather generators, Clim. Res., 34, 129–144, 2007.

Furrer, E. M. and Katz, R. W.: Improving the simulation of extreme precipitation events by stochastic weather generators, Water Resour. Res., 44, W12439, doi:10.1029/2008WR007316, 2008.

Gabriel, K. R. and Neumann, J.: A Markov chain model for daily rainfall occurrence at Tel Aviv, Roy. Meteorol. Soc., 88, 90–95, 1962.

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.

Grunwald, G. K. and Jones, R. J.: Markov models for time series with mixed distribution, Environmetrics, 11, 327–339, 2000.

Hastie, T. J. and Tibshirani, C.: Generalized Additive Models, Chapman Hall, London, 1990.

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.

Katz, R. W.: Precipitation as a chain dependent process, J. Roy. Stat. Soc. B, 16, 671–676, 1977.

Katz, R. W. and Parlange, M. B.: Generalizations of chain dependent processes: Application to hourly precipitation, Water Resour. Res., 31, 1331–1341, 1995.

Katz, R. W. and Parlange, M. B.: Conditioning on indices of atmospheric circulation, Meteorol. Appl., 5, 75–87, 1998.

Koenker, R.: Quantile Regression, Cambridge University Press, 2005.

Koutsoyiannis, D. and Onof, C.: Rainfall disaggregation using adjusting procedures on a Poisson cluster model, J. Hydrol., 246, 109–122, 2001.

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.

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

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.

Papalexiou, S. M., Koutsoyiannis, D., and Makropoulos, C.: How extreme is extreme? An assessment of daily rainfall distribution tails, Hydrol. Earth Syst. Sci., 17, 851–862, doi:10.5194/hess-17-851-2013, 2013.

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.

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

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.

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

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.

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.

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.

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