*Bulgarian Journal of Meteorology and Hydrology*

**Bulgarian Academy of Sciences**

**National Institute of Meteorology and Hydrology**

**Bul. J. Meteo & Hydro 22/5 (2017) 71-89**

# Stochastic models of daily precipitation totals with hybrid distributions for Zlatograd station

**Nadya Neykova*, Plamen Neytchev**

*National Institute of Meteorology and Hydrology- BAS,*
*Tsarigradsko shose 66, 1784 Sofia, Bulgaria*

**Abstract**: A stochastic daily precipitation model for Zlatograd station in south Bulgaria conditional on atmospheric predictors that characterize the atmospheric circulation over Balkans is developed. The model consists of two components describing the occurrence and intensity precipitation series. The intensity component is based on a hybrid between gamma and generalized Pareto distributions(GP) and Weibull and GP. The results of simulations designed to compare the models based on the hybrid distributions and those based on the standard gamma and Weibull distributions are reported and some potential difficulties are outlined.

**Keywords:** stochastic precipitation modeling, gamma, Weibull, generalized Pareto distribution, hybrid distributions

# Стохастични модели на денонощните суми на валежите с хибридни разпределения за станция Златоград

Надя Нейкова*, Пламен Нейчев

*Национален институт по метеорология и хидрология – БАН,*
*Бул. Цариградско шосе 66, 1784 София, България*

**Резюме**: Предложен е стохастичен модел на денонощните суми на валежите в климатична станция Златоград, който се състои от две компоненти: модел на вероятността за валеж и модел на количеството на валежа. Моделите включват атмосферни предиктори, характеризиращи тропосферата над Балканския

* nadya.neykova@meteo.bg

полуостров. За моделиране на количеството на валежа са използвани гама и Вейбул разпределенията, и хибридни разпределения между тях и обобщеното разпределение на Парето (GP). Сравнителният анализ между историческите и симулираните по моделите данни показва добро сходство. Симулираните редици от данни могат да бъдат използвани за оценка на риска от някои екстремални природни явления (наводнения, засушавания, ерозия на почвите и други).

**Keywords:** стохастично моделиране на валежите, гама, Вейбул, обобщено разпределение на Парето, хибридни разпределения

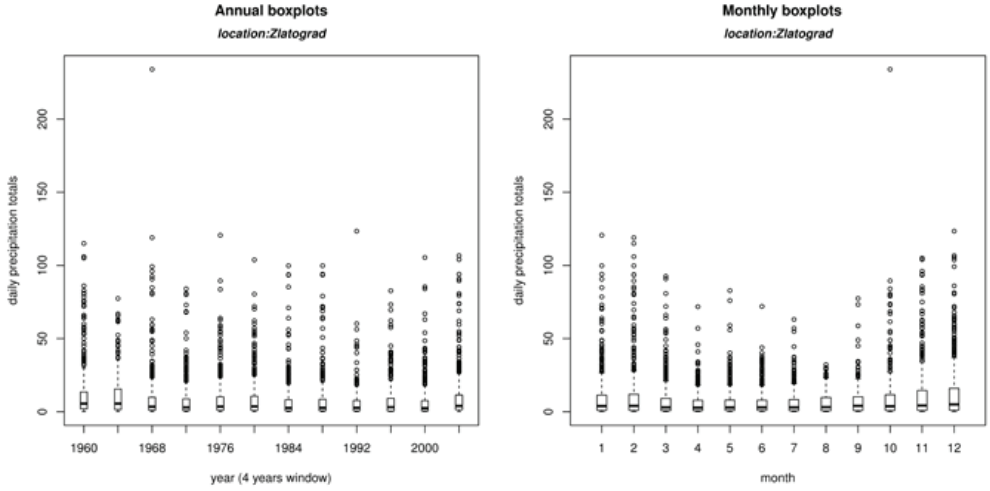
## 1. ВЪВЕДЕНИЕ

Създаването на стохастични модели за валежите е актуална задача в социалната и стопанска дейност. Моделите на валежите се състоят от две компоненти, описващи появата на валеж и количеството на валежа. За моделиране на появата се използва бинарна логистична регресия. За целта данните за денонощните валежни суми се превръщат в бинарни, след което се моделира вероятността за поява на валеж. За моделиране на количеството се използват непрекъснати дясно скосени разпределения, като например гама и Вейбул или логнормално разпределение. За извършване на пресмятанията може да бъде използван стандартен софтуер за обобщени линейни модели (GLMs). Съществен недостатък на тези разпределения е, че чрез тях не могат да бъдат описани добре екстремалните валежни количества. За да избегнем това, ние адаптираме подхода на Furrer & Katz (2008), базиран на хибридни разпределения. По-точно, ние използваме хибридно разпределение между гама и обобщеното разпределение на Парето (GP), и хибридно разпределение между Вейбул и GP, за да направим стохастични модели на валежите за климатична станция Златоград. С помощта на хибридните разпределения става възможно симулирането (генерирането) на денонощни суми на валежите, разпределението на които наподобява разпределението на реалните данни. Целта на симулирането е да се увеличи обемът на извадката от налични данни и да се проведе статистически анализ върху симулираните данни – параметричен бутстрап (bootstrap), както бихме направили, ако разполагахме с дългогодишни наблюдения върху валежите. Симулираните моделни редици от данни могат да бъдат използвани за оценка на риска от някои екстремални природни явления (наводнения, засушавания, ерозия на почвите и други), наблюдавани в Източните Родопи и поречието на река Арда.

В тази статия е представено прилагането и адаптирането на подхода на Furer & Katz (2008), базиран на хибридни разпределения и е предложен подобрен модел, използващ разпределение с по-тежка опашка за описване на денонощните валежи. Моделът описва редицата от денонощните суми на валежите, като за предиктори включва метеорологични променливи, характеризиращи тропосферата над Балканския полуостров.

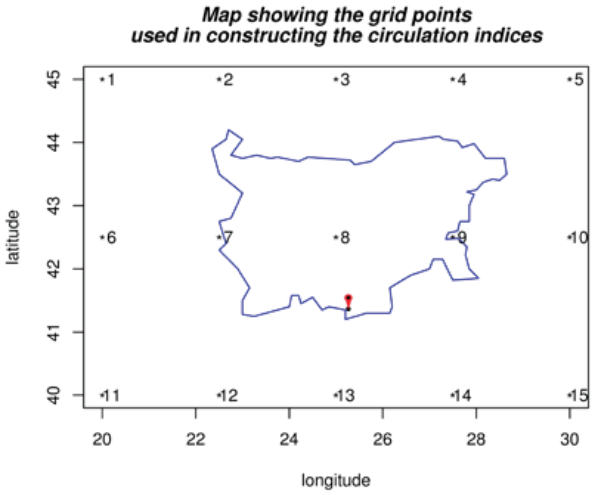
## 2. ОПИСАНИЕ НА ДАННИТЕ

Валеж са течните и/или твърдите продукти на водните пари, които падат от облаците на земната повърхност. В климатичните станции валежите се измерват веднъж в денонощието (към срок 06:00 GMT). В нашето изследване са използвани денонощните суми на валежите в климатична станция Златоград за 48 годишен период (01.01.1960 ÷ 31.12.2007). В тази станция на 03.10.1970 год. е бил измерен рекорден валеж от 234 литра за едно денонощие, а през периода са регистрирани 5402 дни с валеж от общо 17474 дни. На Фиг. 1. е представено разпределението на денонощните валежни суми по години (вляво) и по месеци (вдясно). При създаването на стохастичния модел са използвани атмосферни променливи, продукт на реанализ. Стойностите на променливите във възлите на мрежа (Фиг. 2), на стандартни нива от земната повърхност до 300 hPa, от един и същи срок (12:00 GMT), са комбинирани в конструкции, които отразяват облако- и валежообразуващите условия над Балканския полуостров. Златоград се намира между възли 8 и 13, почти до границата на страната. С влажността на въздуха са свързани променливите shum (специфична влажност в kg/kg) и rhum (относителна влажност в %), prwtr (precipitable water в kg/m2, възможен валеж от атмосферния стълб), prate (precipitatation rate в kg/ m2.s, интензивност на валежа).




**Фиг. 1.** Разпределение на денонощните суми на валежите по години (ляво) и по месеци (дясно). Екстремалната стойност от 234 мм е измерена на 03.10.1970г

**Fig. 1.** Allocation of daily rainfall amounts by years (left) and by months (right). The extreme value of 234 mm was measured on 03.10.1970




**Фиг. 2.** Разположение на възлите на атмосферната циркулация над територията на България

**Fig. 2.** Location of the atmospheric circulation nodes over the territory of Bulgaria

Конструкциите prwtr.u и prwtr.v са градиенти (разлики между стойностите във възли съответно 9 и 7, 13 и 3), а prwtr.x е лапласиан (сумата от стойностите във възли 2, 4, 12 и 14, намалена с 4-кратната стойност в централния възел 8). Подобни са лапласианите shum.x.850 и slp.x.2m съответно за shum на 850 hPa и атмосферното налягане на средно морско ниво. Адвективните конструкции, например adv.u.s.850 и adv.v.s.850 са произведения на стойност на компонента (u или v) на вятъра във възел 8 и съответен градиент на shum на посоченото ниво (в случая 850 hPa). Подобни са адвекциите adv.u.t.sig и adv.v.t.sig – за температурата на въздуха на sig ниво 995 hPa, adv.v.s.10m – за shum на 2 m и вятър на 10 m. Конструкциите gama850 и gama700 са вертикални температурни градиенти над възел 8 в слой съответно между 1000 и 850 и 700 hPa, а sgama300 е вертикалния температурен градиенти в слоя между 1000 и 300 hPa, умножен по средната стойност на shum на нива 850, 700 и 500 hPa. Конструкциите ampl.$t_{max}$ и ampl.$t_{min}$ са разликите между температурата на въздуха air.2m и tmax и tmin за една и съща дата в централния възел 8. Стойностите на променливите са от предходната дата (лаг 1) спрямо денонощните суми. Използвани са също стойностите на северноатлантическата (NAO) и арктическата (AO) осцилации с по-голям лаг.

# 3. МОДЕЛИРАНЕ НА ДЕНОНОЩНИТЕ СУМИ НА ВАЛЕЖИТЕ

## 3.1. Стохастичен модел на валежа

Нека $Y_t$ е сумата на валежа в деня $t$, $Z_t = (Z_{1t}, \ldots, Z_{kt})'$ е вектор от атмосферни променливи свързан с $Y_t$ за $t = 1, \ldots, T$. Денят $t$ е сух (без валеж) ако $Y_t < c$, където $c$ е предварително зададена константа, чиято стандартна стойност е $c = 0.1$ mm, и мокър (с валеж) ако $Y_t \geq c$. Редицата от сухи и мокри дни е представена с помощта на индикаторната функция $I_t = I_{[Y_t \geq c]}$, която приема стойност 0 ако денят $t$ е сух и 1 ако денят $t$ е мокър. Стойностите на $Y_t$ , $Z_t$ и $I_t$ ще означаваме с малки букви. Да означим с $p_t(x_t)$ вероятността денят $t$ да е с валеж, при наблюдавани стойности на вектора $x_t = (i_{t-1}, \ldots, i_{t-m}, y_{t-1}, \ldots, y_{t-m}, z_{1t}, \ldots, z_{kt})'$, в който са включени лагове от ред $m$ на индикаторната променлива и денонощната сума на валежа. Количеството валеж за деня $t$ се дефинира като $R_t = Y_t$ ако $Y_t \geq c$ и $R_t =$ липсващата стойност в противен случай. Означаваме с $q(r_t|x_t)$ условната плътност на разпределение, което е от тип дясно скосено разпределение, понеже преобладават валежите с малко количество.

Редицата от денонощни суми на валежите се моделира като смес от две разпределения. Едното разпределение е дискретно с маса в нулата, отчитащо дните без валеж, докато второто разпределение е непрекъснато дясно скосено, отчитащо количеството валеж на дните с валеж, (Grunwald & Jones, 2000; Stern & Coe, 1984). Понеже двете състояния са взаимно изключващи се (несъвместими), то съответната преходна вероятност е:

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

Най-често използваните разпределения в практиката за $q_t(r_t|x_t)$ са гама, Вейбул, лог-нормално. Ако се интересуваме от модели на екстремалните денонощни валежни суми, тогава се използват обобщеното разпределение на екстремалните стойности (GEV) или обобщеното разпределение на Парето (GP).

При условие, че $p_t(x_t)$ и $q_t(r_t|x_t)$ нямат общи параметри, функцията на правдоподобие за $(y_{t-m-1}, \ldots, y_T)$ се дефинира както следва

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

От тази факторизация на функцията на правдоподобие се вижда, че за оценяването на неизвестните параметри могат да бъдат използвани стандартни програмни процедури за обобщени линейни модели (McCullagh & Nelder, 1989), като *glm*, *vglm* от *VGAM* и други от програмната среда R (R, Core Team 2017). Това е така, понеже първата компонента представлява функцията на правдоподобие на бинарен времеви ред, докато втората компонента е функцията на правдоподобие на количеството на валежа с някое от споменатите по-горе разпределения.

## 3.2. Модел за появата на валеж

За моделиране на вероятността за поява на валеж $p_t(x_t)$ ще използваме модел на логистичната регресия:

$$\text{logit}\big(p_t(x_t)\big) = \log(p_t(x_t))/(1-p_t(x_t))) = u(x_t) =$$
$$= \alpha_0 + \sum_{l=1}^{p}(\alpha_l i_{t-l} + g_l(y_{t-l})) + \sum_{l=1}^{k} g_{p+l}(z_{lt}) + g_{p+k+l}(t).$$

В линейния предиктор $u(x_t)$ се включват лагови предиктори на появата $i_{t-l}$ и количеството $y_{t-l}$ на валежа, краен ред на Фурие, за да бъде отчетена сезонността на валежите, както и атмосферни предиктори, където $g_l$ за $l = 1, \ldots, p+k+1$ са непрекъснати функции. Пример за подобен модел:

$$\text{logit}\big(p_t(x_t)\big) = \alpha_0 + \alpha_1 i_{t-1} + \alpha_2 C_t + \alpha_3 S_t + \alpha_4 NAO_{t-1} + [\beta_2 C_t + \beta_3 S_t + \beta_4 NAO_{t-1}] i_{t-1},$$

където $C_t = \cos(2\pi t/365.25)$ и $S_t = \sin(2\pi t/365.25)$, а NAO е северно-атлантическата осцилация. Предикторите в този модел са $x_t = (1, i_{t-1}, C_t, S_t, NAO_{t-1}, C_t, \quad i_{t-1}, S_t, i_{t-1}, NAO_{t-1}, i_{t-1})$, $\alpha_i$ са неизвестни параметри.

## 3.3. Разпределения на количеството валеж

За моделиране на разпределението на количеството на валежа са използвани гама, Вейбул и GP разпределенията. Поради различните параметризации на тези разпределения, в тази част са дадени техните плътности, които са използвани в библиотеката VGAM (Yee, 2015), за да бъдат избегнати недоразумения с други алтернативни представяния в литературата.

Плътност на гама разпределението:

$$f(r,a,b) = \begin{cases} \dfrac{b^a\, r^{(a-1)} \exp(-br)}{\Gamma(a)}, & r > 0 \\ 0 & r = 0, \end{cases}$$

където $\Gamma(a)$ е гама функцията, $a > 0$ е параметър на формата (shape), $b > 0$ е параметърът, свързан с мащаба (scale). При тази параметризация очакването и дисперсията са съответно равни на $\mu = a/b$ и $\sigma^2 = \mu^2/a = a/b^2$.

Плътност на обобщеното разпределение на Парето (GP) с праг *u*:

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

където $r > u$, $u, \sigma > 0$ и $\xi$ са параметрите на положението, мащаба и формата, $[A]_+ = \max(A, 0)$. Параметърът $\xi$ характеризира типа на GP разпределението, както следва: с тежка опашка ако $\xi > 0$, с крайна опашка ако $\xi < 0$ и експоненциално (изместено с *u*) разпределение ако $\xi = 0$.

При моделиране на данни с обобщени линейни модели (GLMs) и разпределения на екстремалните стойности се използват следните свързващи функции, чрез които се задава връзката между предикторните (регресионните) променливи и параметрите на гама, Вейбул и GP разпределенията

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

където $\theta_i$ са неизвестните векторни параметри, $x_{it}$ са образувани от вектора на предикторите $x_t$ за $i = 1, \ldots, 4$. Използването на логаритъм като свързваща функция осигурява положителност на параметрите $a$, $b$ и $\sigma$ при максимизирането на функцията на правдоподобие. Пример за свързваща функция, състояща се от периодична компонента и лаг на северно атлантическата осцилация NAO за параметъра *a*, е:

$$\log(a) = \theta_0 + \theta_1 i_{t-1} + \theta_2 C_t + \theta_3 S_t + \theta_4 NAO_{t-1}.$$

Определянето на оценките на неизвестните параметри $\alpha_i$ в модела на поява и $\theta_i$ в модела на количеството се получава чрез максимизирането на функцията на правдоподобие.

## 3.4. Хибридни разпределения

Furrer & Katz (2008) дефинират хибридното разпределение (гама-GP) на основата на гама и GP разпределенията, както следва:

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

където *F(x)* е функцията на гама разпределението, *f(x)* и *g(x)* са плътностите на гама и GP разпределенията, а 1–*F(u)* е нормиращ множител. За да осигурят непрекъснатост в праговата стойност *u* тези автори налагат условието

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

В резултат на това за параметъра $\sigma$ на GP разпределението получаваме $\sigma = (1 - F(u))/f(u)$. Това означава, че параметърът на мащаба на GP разпределението се дефинира изцяло в термините на гама разпределението, с което успешно могат да бъдат моделирани наблюденията с по-малки стойности от прага *u*. Авторите предлагат следната процедура за оценяване на параметрите на хибридното разпределение: (i) моделиране на количеството валеж със стандартна процедура за обобщен линеен модел с гама разпределение по пълната извадка от данни; (ii) моделиране на количествата валеж, превишаващи прага *u*, със стандартна процедура за анализ на екстремалните стойности, основана на GP разпределението, като за целта бъдат използвани подходящи свързващи функции; (iii) заместване на параметъра σ на GP разпределението в стойността на прага *u*. Предложената процедура за дефиниране на хибридни разпределения е универсална, тъй като не зависи от вида на разпределението *F(x)* . В работата на Neykov et al (2014) се разглеждат модели на денонощните суми на валежите за станция Ихтиман с гама, Вейбул и хибридни разпределения между гама и GP, и Вейбул и GP, в които модели NAO е използвана като предиктор.

## 4. РЕЗУЛТАТИ

### 4.1. Модел за поява на валеж

За моделиране на вероятността за поява на валеж е използвана логистичната регресия с линеен предиктор, в който са включени северно-атлантическата и арктически осцилации NAO и AO и PNA с лагове до 5 дни, краен ред на Фурие, за да бъде отчетена сезонността на появата на валеж, изброените в секция 2 атмосферни индекси с лаг 1 както и взаимодействия на състоянието поява на валеж в предишния ден $i_{t-1}$ с някои от тях. Голям брой от включените атмосферни индекси в модела на появата на валеж се оказаха статистически незначими предиктори, според критерия на Стюдент при стандартните нива на съгласие. Това се потвърждава и от резултатите при проверката на хипотези за значимост на оценките на неизвестните параметри пред съответните атмосферни индекси със статистическия критерий (тест), основан на отношението на правдоподобие. Полученият модел е с изразен сезонен характер, а оценките на неизвестните параметри пред някои от атмосферните индекси, за които е известно че формират появата на валежа, са статистически значими. Някои от атмосферните индекси

като NAO, AO и PNA с лаг по-голям от 2, както и взаимодействията, формирани чрез сезонните компоненти с $i_{t-1}$ са статистически незначими предиктори, понеже съответните им параметри са статистически незначими. Проведен бе допълнителен анализ чрез стъпковата процедура stepAIC от библиотеката MASS за избор на статистически значими предиктори с помощта на информационния критерий на Бейс (BIC). езултатите се дадени в Таблица 1 на отклоненията (Deviance table) от изхода на glm процедурата. В резултат на това се достигна до редуциран модел на появата на валеж, включващ най-значимите предиктори със съответните оценени параметри:

$$\begin{aligned}&\mathrm{logit}\big(\mathrm{p_t(x_t)}\big) = \\ &-1.29130 + 0.12396\,\sin(\mathrm{arg1}) + 0.24469\,\cos(\mathrm{arg1}) - 0.14800\,\sin(\mathrm{arg2}) \\ &+0.03803\cos(\mathrm{arg2}) + 0.50874\mathrm{i_{t-1}} - 0.50580\mathrm{prwtrv_{t-1}} - 0.20326\mathrm{prwtrx_{t-1}} \\ &+0.43005\,\mathrm{gama850_{t-1}} + 0.53271\,\mathrm{gama700_{t-1}} + 0.42198\,\mathrm{shumx850_{t-1}} \\ &-0.09687\mathrm{adv.\,u.\,s.\,850_{t-1}} - 0.18907\mathrm{adv.\,u.\,s.\,700_{t-1}} + 0.19096\mathrm{adv.\,u.\,s.\,10m_{t-1}} \\ &-0.35527\mathrm{adv.\,u.\,t.\,sig_{t-1}} - 0.08767\mathrm{adv.\,v.\,s.\,850_{t-1}} - 0.20304\mathrm{adv.\,v.\,s.\,700_{t-1}} \\ &+0.10928\mathrm{adv.\,v.\,s.\,500_{t-1}} - 0.28963\mathrm{adv.\,v.\,t.\,300_{t-1}} - 0.09892\mathrm{adv.\,v.\,s.\,10m_{t-1}} \\ &+0.12684\,\mathrm{adv.\,v.\,t.\,sig_{t-1}} - 0.44172\,\mathrm{ampl.\,tmax_{t-1}} + 0.17676\,\mathrm{ampl.\,tmin_{t-1}} \\ &+0.24329\,\mathrm{slpx.\,2m_{t-1}} + 0.54637\mathrm{rhumv_{t-1}} + 0.21776\,\mathrm{airv2m_{t-1}}\,, \qquad (1)\end{aligned}$$

където arg1=2πt/365.25, arg2=3πt/365.25 за $t = 1, \ldots, \mathrm{T}$ (01.01.1960 – 31.12.2007).

Ще отбележим, че стойността на функцията на отклоненията без предикторни променливи, т.е., само модел със свободен парметър $a_0$ (NULL), е 21579 в таблицата при 17473 степени на свобода. Стойността ѝ се редуцира до 14660 при 17448 степени на свобода след използването на 20те атмосферни индекса, двете сезонни компоненти и свободният член като предиктори в модела. Тези атмосферни индекси са статистически значими предиктори, понеже съответните им параметри са статистически значими. Така например, влажността в атмосферата $shumx850_{t-1}$ и състоянието „валеж“ от предходния ден $i_{t-1}$ редуцират стойността на функцията на отклоненията, съответно с 1575.73 и 1555.09. Линейната комбинция от sin и cos, характеризираща сезонното поведение на появата на валеж, е статистически значима, тъй като поне една от нейните компоненти е значима. Резултатът от процедурата stepAIC е еквивалентен на определянето на оптимален модел сред множество от алтернативни модели по обучаваща и валидираща извадки от данните. Разгледаният модел за вероятността за поява на валеж характеризира много добре историческите бинарни данни.

**Таблица 1**. ANOVA таблица на модел на вероятността за поява на валеж – редуциран модел след използване на процедурата stepAIC

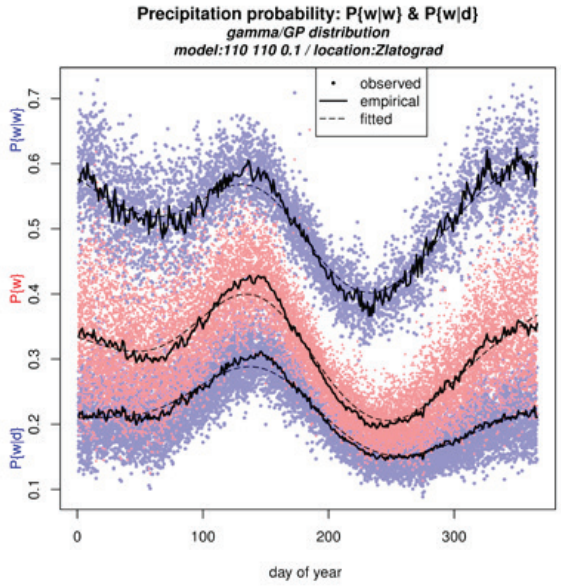
**Table 1**. ANOVA table of model of probability for the occurrence of rainfall - reduced model after using the stepAIC procedure

| Terms | Df | Deviance | Resid.Df | Resid.Dev | Pr(>Chi) | Signif. |
|---|---|---|---|---|---|---|
| NULL | | | 17473 | 21579 | | |
| sin(arg1) | 1 | 181.82 | 17472 | 21397 | < 2.2e-16 | *** |
| cos(arg1) | 1 | 1.34 | 17471 | 21395 | 0.2468671 | |
| sin2(arg2) | 1 | 133.86 | 17470 | 21262 | < 2.2e-16 | *** |
| cos2(arg2) | 1 | 30.83 | 17469 | 21231 | 2.81E-08 | *** |
| $i_{t-1}$ | 1 | 1555.09 | 17468 | 19676 | < 2.2e-16 | *** |
| $prwtrv_{t-1}$ | 1 | 410.52 | 17467 | 19265 | < 2.2e-16 | *** |
| $prwtrx_{t-1}$ | 1 | 682.72 | 17466 | 18582 | < 2.2e-16 | *** |
| $gama850_{t-1}$ | 1 | 371.91 | 17465 | 18210 | < 2.2e-16 | *** |
| $gama700_{t-1}$ | 1 | 457.18 | 17464 | 17753 | < 2.2e-16 | *** |
| $shumx850_{t-1}$ | 1 | 1575.73 | 17463 | 16178 | < 2.2e-16 | *** |
| $adv.u.s.850_{t-1}$ | 1 | 47.4 | 17462 | 16130 | 5.79E-12 | *** |
| $adv.u.s.700_{t-1}$ | 1 | 50.57 | 17461 | 16080 | 1.15E-12 | *** |
| $adv.u.s.10m_{t-1}$ | 1 | 93.08 | 17460 | 15986 | < 2.2e-16 | *** |
| $adv.u.t.sig_{t-1}$ | 1 | 236.57 | 17459 | 15750 | < 2.2e-16 | *** |
| $adv.v.s.850_{t-1}$ | 1 | 143.79 | 17458 | 15606 | < 2.2e-16 | *** |
| $adv.v.s.700_{t-1}$ | 1 | 12.39 | 17457 | 15594 | 0.0004309 | *** |
| $adv.v.s.500_{t-1}$ | 1 | 22.71 | 17456 | 15571 | 1.89E-06 | *** |
| $adv.v.t.300_{t-1}$ | 1 | 211.58 | 17455 | 15359 | < 2.2e-16 | *** |
| $adv.v.s.10m_{t-1}$ | 1 | 15.13 | 17454 | 15344 | 0.0001004 | *** |
| $adv.v.t.sig_{t-1}$ | 1 | 38.06 | 17453 | 15306 | 6.85E-10 | *** |
| $ampl.tmax_{t-1}$ | 1 | 193.92 | 17452 | 15112 | < 2.2e-16 | *** |
| $ampl.tmin_{t-1}$ | 1 | 20.31 | 17451 | 15092 | 6.57E-06 | *** |
| $slpx.2m_{t-1}$ | 1 | 127.86 | 17450 | 14964 | < 2.2e-16 | *** |
| $rhumv_{t-1}$ | 1 | 243.86 | 17449 | 14720 | < 2.2e-16 | *** |
| $airv2m_{t-1}$ | 1 | 60.17 | 17448 | 14660 | 8.69E-15 | *** |
| Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | | | | |

Използвайки (1) пресмятаме условните вероятности на нехомогенната Марковска верига на прехода с две състояния $p_{01}(t) = p_t(x_t)$ за $i_{t-1} = 0$ (вероятността да вали в деня $t$ при условие, че предишния ден не е валяло) и $p_{11}(t) = p_t(x_t)$ за $i_{t-1} = 1$. От формулата за пълната вероятност следва връзката между $\pi(t) = \Pr(I_t = 1|z_t)$ и вероятностите на матрицата на прехода $p_{01}(t)$ и $p_{11}(t)$:

$$\pi(t) = \Pr(I_t = 1|z_t) = \pi(t-1)p_{11}(t) + (1 - \pi(t-1))p_{01}(t) \qquad (2)$$

При предположение, че вероятностите за валеж във всеки два съседни дни са близки, т.е. $\pi(t) \approx \pi(t-1)$ то за всяко $t$ получаваме $\pi(t) \approx p_{01}(t)/(p_{01}(t) + 1 - p_{11}(t))$, Furrer and Katz (2007). Марковската матрица на прехода е хомогенна, когато $p_{01}(t)$ и $p_{11}(t)$ зависят само от $I_{t-1} = i_{t-1}$, т.е. не се използва информацията от данните $Z_t = z_t$, за $t = 1, \ldots, T$.




**Фиг. 3.** Вероятност за поява на валеж. Точките представляват предсказаните стойности на условните вероятности $p(x_t)$ при $i_{t-1} = 1$ (горе в синьо), $p(x_t)$ при $i_{t-1} = 0$ (долу в синьо) и безусловната вероятност π (*t*) (средата в розово), пунктирните линии са съответните им предсказани стойности чрез линеен предиктор.

**Fig. 3.** Probability of precipitation occurrence. The points represent the predicted values of conditional probabilities $p(x_t)$ for $i_{t-1} = 1$ (up in blue), $p(x_t)$ for $i_{t-1} = 0$ (below in blue) and the unconditional probability π (*t*) (middle in pink), the dashed lines are the respective predicted values using a linear predictor.

На Фиг. 3 са представени: (a) с назъбени горна и долна линии честотите за поява на валеж пресметнати по данните при условие, че предният ден е бил с или без валеж; (b) оценените вероятности $p_{01}(t)$, $p_{11}(t)$ и $\pi(t)$, като функции само на двете хармоники (без атмосферни предиктори), са дадени с пунктирна линия; (c) сините точки (горе при $i_{t-1} = 1$ и долу при $i_{t-1} = 0$ съответстват на оценените вероятности $p_t(x_t)$, т.е. на $p_{01}$ и $p_{11}$ с предикторите от оптималния модел, докато розовите точки в средата съответстват на оценените вероятности π (*t*) от рекурентната формула (2).

## 4.2. Модел за количеството валеж

За моделиране на количеството на денонощните валежи сме използвали гама и Вейбул разпределенията. В тази секция ще бъдат показани резултатите от моделирането с гама разпределението, т.к. резултатите за разпределението на Вейбул са подобни. По аналогия с модела на вероятността за валеж, при създаването на модел за количеството на денонощния валеж са използвани стойностите на всички атмосферни индекси от предходния ден както и някои взаимодействия с лаг на количеството от предишния ден, докато за NAO, AO и PNA са използвани лагове до 5 дни. За модела на количеството на валежа са използвани 5406 дни, в които е регистриран валеж в станцията Златоград. Голям брой от включените атмосферни индекси се оказаха статистически незначими предиктори, според критерия на Стюдент при стандартните нива на съгласие. Това се потвърждава и от резултатите при проверката на хипотези за значимост на оценките на параметрите пред съответните атмосферни индекси с теста, основан на отношението на правдоподобие. Полученият модел е с изразен сезонен характер, а оценките на неизвестните параметри пред някои от атмосферните индекси, за които е известно че формират количеството валеж, са статистически значими. Проведен бе допълнителен анализ чрез стъпковата процедура stepAIC. В резултат на това се достигна до редуциран модел, включващ най-значимите предиктори със съответните оценени параметри:

$$\begin{aligned} p_t(x_t) = {} & 1.445864 - 0.081567\sin(\mathrm{arg1}) + 0.238677\cos(\mathrm{arg1}) + 0.002521\sin(\mathrm{arg2}) \\ & + 0.064810\cos(\mathrm{arg2}) + 0.013826\sin(\mathrm{arg3}) - 0.057494\cos(\mathrm{arg3}) \\ & + 0.003447\,\mathrm{y}_{t-1} - 0.053933\,\mathrm{AO}_{t-1} + 0.078619\,\mathrm{NAO}_{t-2} - 0.185295\,\mathrm{prwtrv}_{t-1} \\ & + 0.078402\,\mathrm{prwtru}_{t-1} + 0.162051\,\mathrm{gama700}_{t-1} + 0.219000\,\mathrm{sgama300}_{t-1} \\ & + 0.055719\,\mathrm{shumx850}_{t-1} - 0.072066\,\mathrm{adv.u.s.700}_{t-1} - 0.191617\,\mathrm{adv.u.t.sig}_{t-1} \\ & - 0.074729\,\mathrm{adv.v.s.850}_{t-1} - 0.060022\,\mathrm{adv.v.s.700}_{t-1} \\ & - 0.178242\mathrm{adv.v.t.300}_{t-1} - 0.064057\mathrm{adv.v.s.10m}_{t-1} - 0.106929\mathrm{adv.v.t.sig}_{t-1} \\ & + 0.223056\,\mathrm{slpx.2m}_{t-1} + 0.074354\,\mathrm{rhumv}_{t-1} + 0.052751\,\mathrm{pratev}_{t-1} \end{aligned} \quad (3)$$

Стойността на функцията на отклоненията, която е мярката за качеството на модела, без използването на атмосферни индекси е 11559.9 при 5380 степени

на свобода. Тя се редуцира до 8904.0 при 5356 степени на свобода в резултат на включването на значимите атмосферни величини и сезонни компоненти.

## 4.3. Поверка на хипотезата за тежка опашка на разпределението на екстремалните валежи

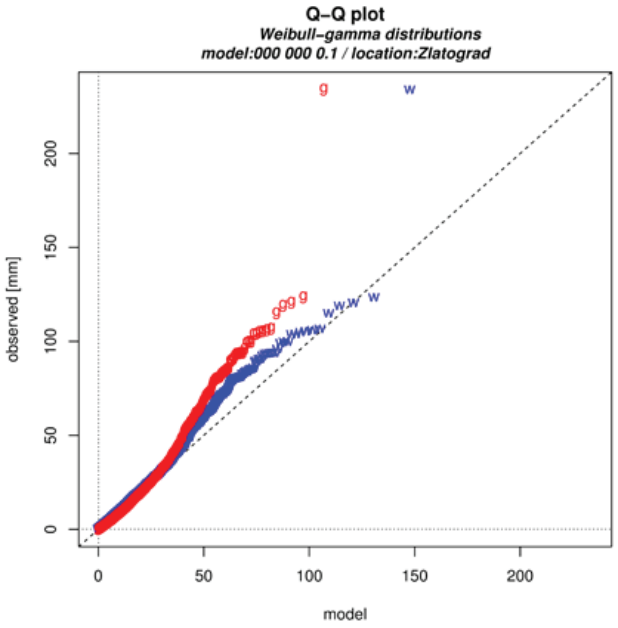
Създаването на хибридно разпределение е целесъобразно при условие, че данните са генерирани от разпределение с тежка опашка. Проверява се хипотезата $H_0: \xi = 0$ за експоненциален закон на разпределението срещу алтернативната $H_1: \xi > 0$, т.е. разпределението на количеството валеж да е с тежка опашка. За целта, денонощните суми на валежите се деклъстерират, като се извличат максималните стойности на всяка група от данни във времето, превишаващи предварително зададен праг (емпиричен квантил на количеството на денонощните валежи), след което се използва стандартната методология за оценяване на неизвестните параметри на GP разпределението на есктремалните стойности. Стойността на прага се избира между 80% и 95% емпиричен квантил на количеството на валежа, за който качеството на апроксимацията на теоритичните GP квантили спрямо емпиричните е удовлетворителна. Максимално правдоподобните оценки на параметъра ξ на GP разпределението за прагове 10 и 15 са 0.2214157 и 0.2016251, съответно. Съгласно критерия на отношението на правдоподобие, отхвърляме хипотезата за екпоненциален закон на разпределението на количеството валеж за станция Златоград, тъй като стойностите на вероятността на опашката (p-value) са по-малки от стандартните нива на съгласие: 3.104358e-13 и 3.363121e-08, съответно. Т.е. разпределението на количеството на денонощните суми на валежа е с тежка опашка. Необходимите пресмятания са извършени в средата на *vglm* процедурата.

## 4.4. Хибридни разпределения. Сравнителен анализ

Известен недостатък на гама и Вейбул разпределенията е, че чрез тях не могат да бъдат описани добре екстремалните стойности на валежите. Това се вижда добре на Фиг. 4. За да избегнем този недостатък ние адаптираме подхода на Furrer & Katz (2008), базиран на хибридни разпределения .

Изборът на праг е от изключително значение за конструирането на хибридните разпределения. Разгледани са модели с прагове 5 мм, 10 мм и 25 мм. Така например: (1) ако бъде избран праг с голяма стойност, оценяването на параметрите на GP разпределението ще бъде по извадка с твърде малък обем, което ще се отрази на надеждността на резултатите, поради големите стойности на оценката на стандартните грешки; (2) ако бъде избран праг с малка стойност, тогава ще бъдат нарушени предположенията за валидност на GP разпределението, докато формалното му използване ще доведе до голяма тежест на опашката, вследствие

на което бихме получавали твърде големи предсказани квантилни стойности за денонощни валежни суми.

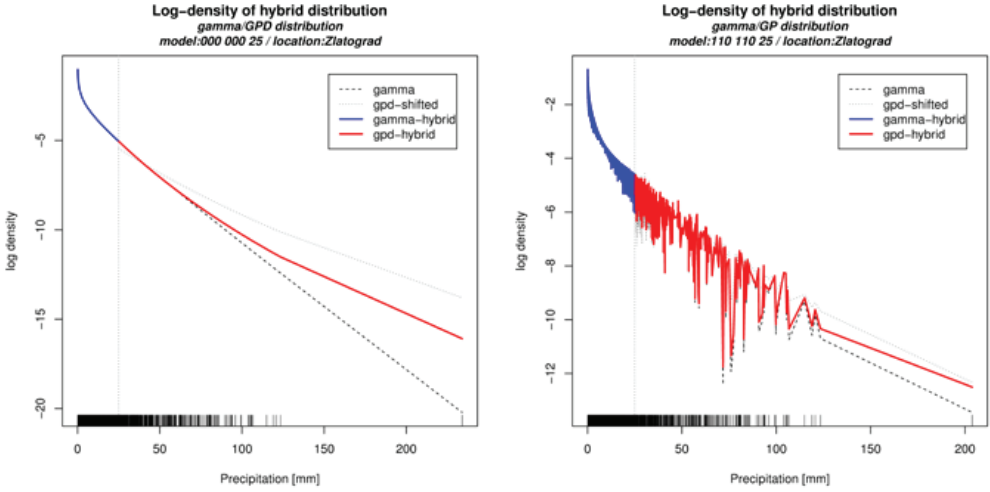



**Фиг. 4.** Q-Q плот на измерените денонощни количества и оценените квантили от модели на гама (g) и Вейбул (w) разпределенията за периода 1960 – 2007 г.

**Fig. 4.** Q-Q plot of measured 24-hour quantities and estimated quantils by models of gamma (g) and Weibull distributions (w) for the period 1960-2007

На Фиг. 5а) са показани логаритмите от максимално правдоподобните оценки на плътности с хомогенни (без използване на предиктори) параметри. Оценките на гама разпределението са представени с непрекъсната синя линия до праг с 25мм и пунктирна линия след съответния праг. Тези на GP са представени с линия от многоточие след прага, а на хибридно гама-GP − с непрекъсната линия (синя и червена). Вертикалната права на плота представлява прага, а количеството на денонощните валежи е дадена с малките вертикални черти по хоризонталната ос. Интерпретацията на резултатите от плота на Фиг. 5.б) е като на плот а), но за МП оценка на плътността на гама разпределението са използвани атмосферни предикторни променливи. Във връзка с надеждността и приложимостта на разглежданите модели на денонощните суми на валежа бе проведен сравнителен анализ между симулираните и наблюдаваните валежи. На Фиг.6. а) са показани 95% и 99% емпирични квантили (точките) и моделните квантили с гама (непрекъсната линия) и хибридно гама-GP (пунктирна линия) разпределения на денонощните количества валеж с праг 5мм за модел без атмосферни предиктори. Аналогични

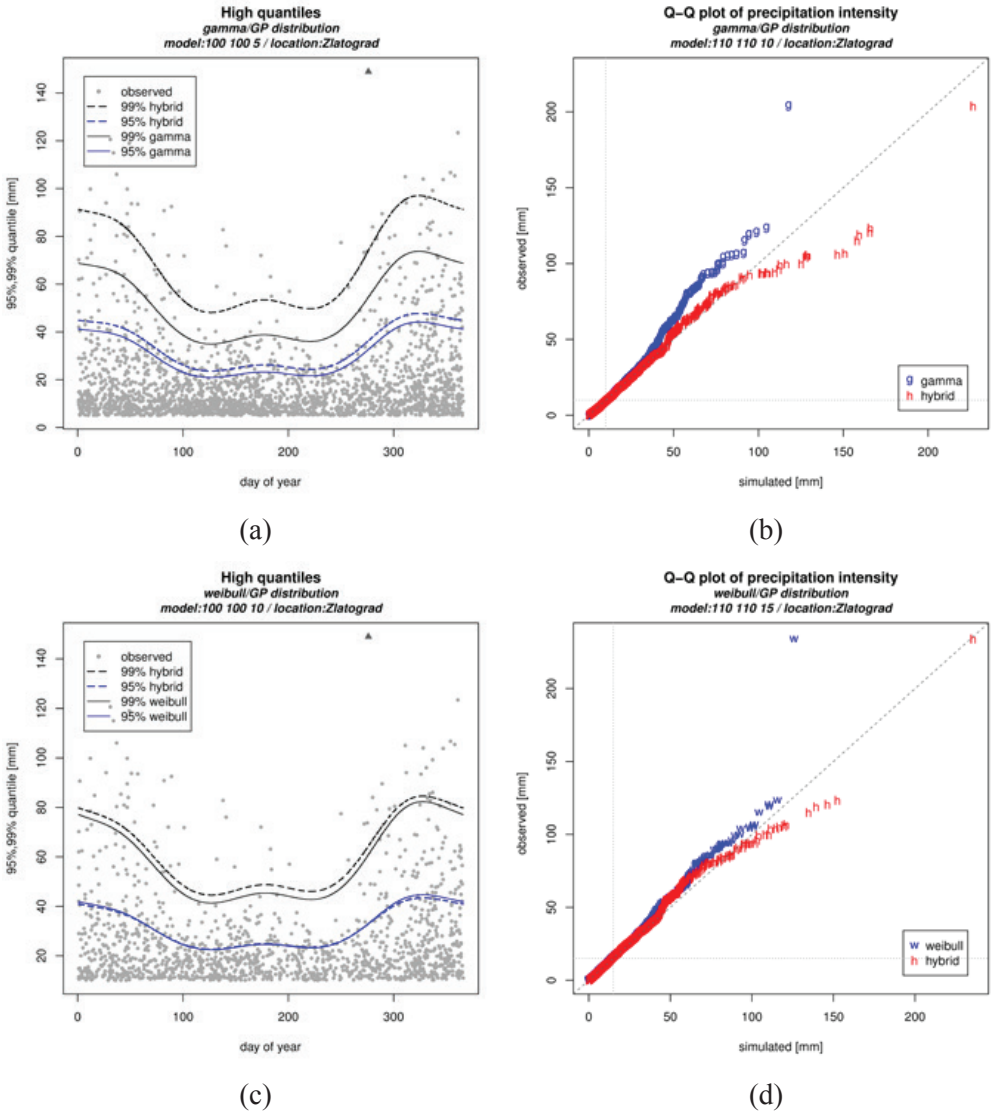
резултати, но за Вейбул разпределението са показани на Фиг.6. c). Екстремната стойност от 234 мм, измерена на 03.10.1970 год., е маркирана с плътен триъгълник. Ефектът от хибридизацията на разпределенията се откроява ясно при по-високите квантили и по-малка стойност от 5мм за праг. На Фиг. 6. b) е показан Q-Q плот на наблюдавани и симулирани чрез гама (g) и хибридно гама-GP (h) квантили на денонощните суми на валежите с праг 10 мм. Ефектът от хибридизацията на разпределенията се откроява при избор на праг с по-малка стойност.




(a) (b)

**Фиг. 5.** Логаритми от максимално правдоподобните оценки на плътностите на гама, GP и хибридно гама – GP разпределенията. (a) без използване на предиктори в модела; (b) с използване на атмосферни предикторни променливи

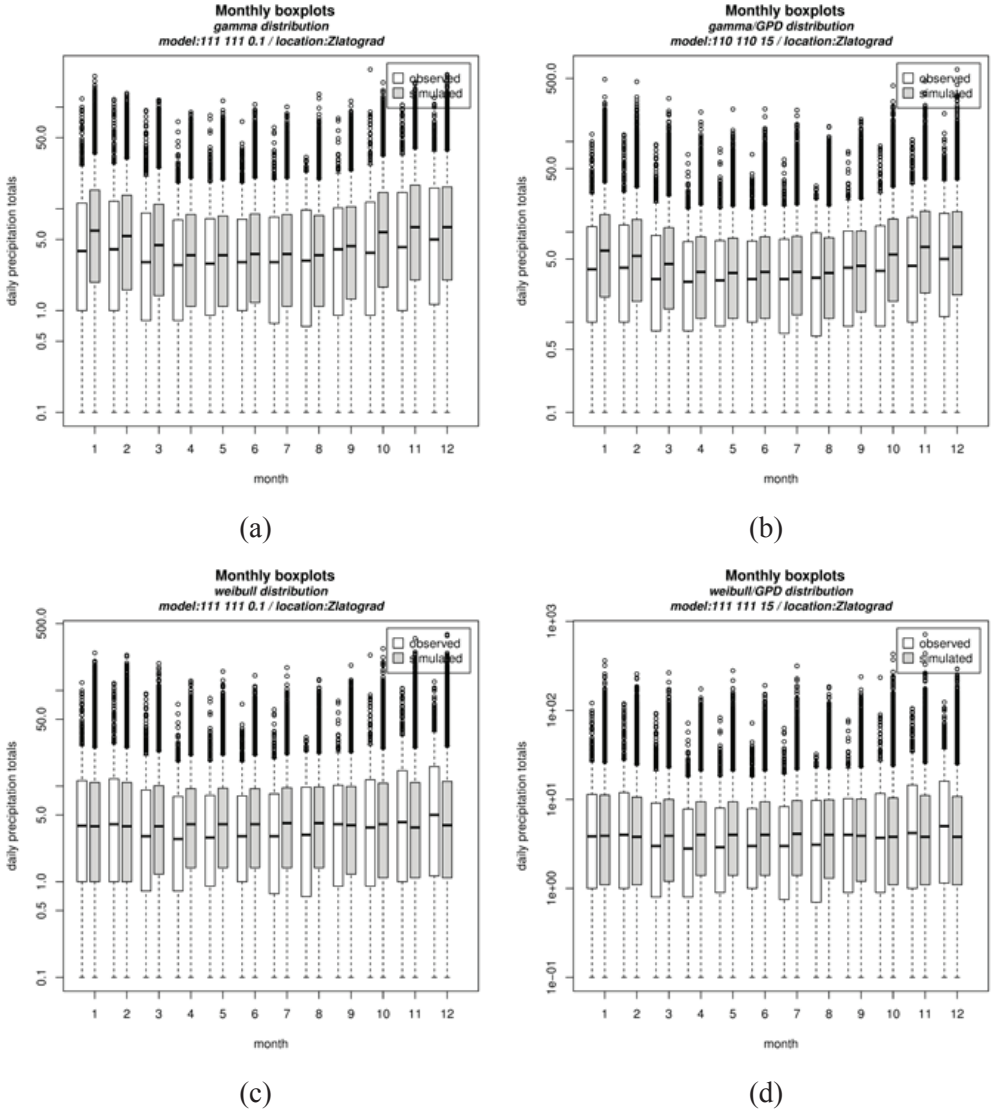
**Fig. 5.** Logarithms of the most plausible densities estimates of gamma, GP and hybrid gamma - GP distributions. (a) without using predictors in the model; (b) using atmospheric predictor variables




**Фиг. 6.** (a) и (c): 95% и 99% емпирични квантили (точките) и моделните квантили с гама и Вейбул (непрекъсната линия) и хибридно гама-GP и Вейбул-GP (пунктирна линия); (b) и (d) Q-Q плот на наблюдавани и симулирани чрез гама (g) и Вейбул(w) и хибридно гама-GP(h) и Вейбул-GP(h)

**Fig. 6.** (a) and (c): 95% and 99% empirical quantils (dots) and model quantiles with gamma and Weibul (solid line) and hybrid gamma-GP and Weibul-GP (dashed line); (b) and (d) Q-Q plot of observed and simulated by gamma (g) and Weibull (w) and hybrid gamma-GP (h) and Weibul-GP (h)

На Фиг. 7 са представени плотове (box-plots) на денонощните суми на наблюдаваните и симулирани валежи по месеци в логаритмична скала. Симулациите са базирани на 50 повторения над разглеждания 47 годишен период, т.е., генерирани са 2350 годишни денонощни суми на валежите с модел, включващ предиктори $i_{t-1}$, една хармоника, и значимите атмосферни предиктори.




**Фиг. 7.** Денонощни суми на наблюдаваните и симулирани валежи по месеци в логаритмична скала.

**Fig. 7.** Daily sums of observed and simulated rainfall by months in logarithmic scale.

На фигурата, симулациите са направени от модел: само с гама разпределение (a), с хибридно гама - GP разпределение, 15 мм праг (b); само с Вейбул (c) и хибридно Вейбул—GP, 15 мм праг (d). Вижда се, че моделите на количеството, използващи стандартните гама и Вейбул разпределения генерират сходни с наблюдаваните денонощни количества с изключение на екстремалните стойности на денонощните валежни суми. Ефектът от хибридизацията е виден: чрез хибридните модели можем да генерираме денонощни валежни суми в целия спектър на наблюдаваните валежи, дори по-големи от наблюдаваните, макар и с малка вероятност.

## 5. ЗАКЛЮЧЕНИЕ

В това изследване бяха разгледани няколко модела за появата и количеството на денонощните валежи за станция Златоград, включващи атмосферни предиктори, характеризиращи поведението на атмосферната циркулация над Балканския полуостров. За моделиране на разпределението на количеството валеж бяха използвани гама и Вейбул разпределенията, и хибридните разпределения между тях и опашката на обобщеното разпределение на Парето. В резултат на изследването беше установено, че разпределението на денонощните суми на валежите за станция Златоград е с тежка опашка. Беше проведен сравнителен анализ на историческите данни със симулираните по моделите данни, основани на хибридни разпределения от тип гама - GP, Вейбул - GP и беше установено, че симулираните данни наподобяват наблюдаваната редица от данни. Необходимите изчисления бяха проведени със стандартни статистически процедури като glm и vglm от R. Определянето на прага за свързване на основното разпределение с GP разпределението е напълно субективен, което се оказа сериозен проблем дори при сравнително малък брой атмосферни предикторни променливи.

Придобитият опит ще бъде от полза при създаването на модели на часовите суми на валежите чрез хибридни разпределения, което е актуална задача, тъй като те се използват за оценка на риска от наводнения, при проектирането на канализационни мрежи за управление на водите и водните ресурси, за борба с ерозията на почвите и други.

## REFERENCES

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

Furrer, E. and Katz R., (2008), Improving the simulation of extreme precipitation events by stochastic weather generators, Water Resour. Res., 44, W12439

Grunwald, G. and Jones, R., (2000), Markov models for time series with mixed distribution, Environmetrics, 11, 327-339

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

McCullagh P. and Nelder J., (1989), Generalized linear models. Chapman and Hall.

R Development Core Team, (2017), A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Yee, T., (2015), Vector generalized linear and additive models with an implementation in R, Springer-Verlag New York

Neykov, N., Neytchev, P. and Zucchini, W., (2014),.Stochastic daily precipitation model with a heavy-tailed component, Nat. Hazards Earth Syst. Sci., 14, 2321-2335.