



Space-time stochastic daily precipitation occurrence model for Bulgaria

Nadya Neykova*, Plamen Neytchev

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

Abstract. A space-time stochastic daily precipitation occurrence model is developed using precipitation data of 32 stations broadly covering the territory of Bulgaria for the period 1960-2018. The model is based on homogeneous hidden Markov model. The hidden states of the model represent different precipitation patterns. The identified states are found to be physically interpretable in terms of regional climatology related to atmospheric circulations of the Europe-Atlantic.

Keywords: Stochastic model for daily precipitation, Hidden Markov model

Пространствено-времеви стохастичен модел за поява на валеж в България

Надя Нейкова*, Пламен Нейчев

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

Резюме. Създаден е пространствено-времеви стохастичен модел за поява на валеж по данни за денонощните суми на валежите от 32 станции от мрежата на НИМХ в България за периода 1960 - 2018 г. Моделът се основава на хомогенните скрити Марковски модели. Неявните състояния (hidden states) на модела за валежите представляват валежни обстановки. Моделът притежава метеорологична интерпретация, свързана с атмосферната циркулация над Атлантико-Европейския регион.

Ключови думи: стохастичен модел за поява на валеж, скрит Марковски модел

* nadya.neykova@meteo.bg

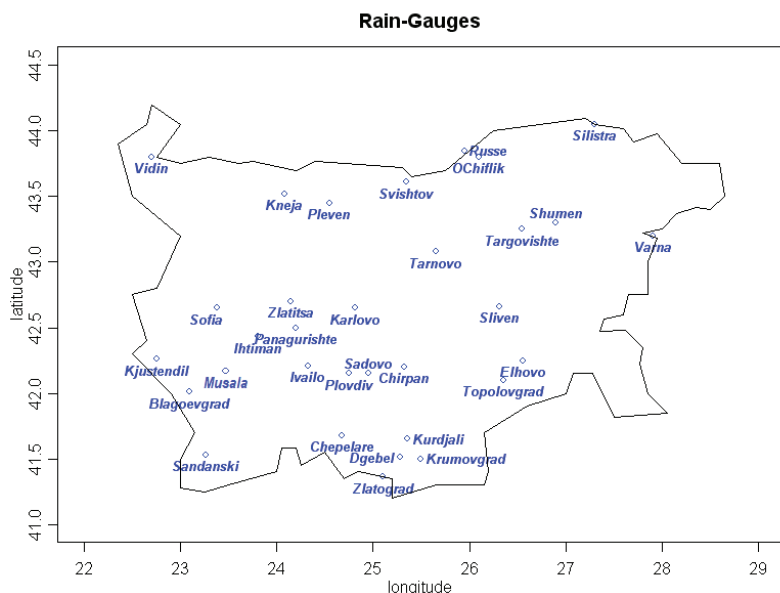
1. ВЪВЕДЕНИЕ

Създаването на пространствено-времеви стохастичен модел на денонощните суми на валежите е актуална задача в социалната и стопанска дейност. Модел от този тип за територията на България е създаден за периода 1960-2000 г. от колектив Neytchev et al. (2006), Neytchev et al. (2011) и Neykov et al. (2012). Необходимо е обновяване и развитие на този модел. Целта на изследването е да бъде създаден пространствено-времеви стохастичен модел за поява на валеж по данни от 32 валежомерни станции от мрежата на НИМХ в България, обхващащи равномерно територията на страната, за периода 1960 – 2018 г. Моделът се основава на скрит Марковски модел (hidden Markov model, НММ), т.е. модел с неявни (ненаблюдаеми, скрити) състояния. В настоящето изследване ще разгледаме хомогенен (homogeneous) НММ. Предимството при моделиране на денонощните суми на валежите чрез НММ е, че с него става възможно синхронно моделирането на валеж в група от станции. НММ се състои от два случайни процеса, които в нашето изследване са пространствен времеви ред на денонощните суми на валежа и дискретен Марковски процес, който приема краен брой състояния. Марковският процес се характеризира с матрица от вероятности за преход от едно състояние към друго и зависи от състоянието на системата в предходния момент. Вероятностите на матрицата на преход от едно състояние в друго се оценяват по матрицата от 32-мерни бинарни вектори от данни. Координатите на тези вектори приемат стойност 1 или 0, в зависимост от това дали е валидно или не в конкретния ден в съответната станция. Оптималният брой ненаблюдаеми състояния на НММs е 8-12 за територията на България, според Neykov et al. (2012). Тези състояния се характеризират чрез валежните обстановки над територията на страната, например: 1) вали над източна България; 2) вали над западна България; 3) вали над северозападна България; 4) вали над югозападна България; 5) вали над югоизточна България; 6) вали над североизточна България; 7) вали над цялата страна; 8) не вали над България. Състоянието без валеж над територията на страната е атрактор, в който се завръща система от състояния, понеже преобладават дните без валеж. Оптималният брой състояния (валежни обстановки), които приема ненаблюдаемият Марковски процес се определя чрез информационния критерий на Бейс (BIC). Предпочитани са моделите с минимална стойност на BIC, които притежават оптимален брой валежни обстановки и физическо съответствие с атмосферната циркулация над Атлантико-Европейския регион. За тази цел са използвани данните от реанализите на NCEP/NCAR

<https://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml>.

2. МЕТОДИКА НА ИЗСЛЕДВАНЕТО

Пространствено-времевият модел е създаден по данни за денонощните суми на валежите от 32 климатични и синоптични станции от мрежата на НИМХ за периода 1960-2018 г., обхващащи равномерно територията на България. На Фигура 1 са представени използваните в изследването валежомерни станции.



Фиг. 1. Разположение на валежомерните станции, включени в изследването.

Fig. 1. Map of rain gauge stations over the territory of Bulgaria included in the study.

За проверка на качествата на модела са използвани метеорологични полета, продукт на данните от реанализите на NCEP/NCAR, над Атлантико-Европейския сектор (30°W-60°E, 20°N-70°N) в 777 възли в мрежа от 2.5°x2.5° на стандартни нива 700 и 850 hPa от един и същи срок (12:00 GMT) за периода 1960 – 2018 г., характеризиращи облако - и валежо-образуващите условия над територията на България. Използвани са следните полета - атмосферно налягане, приведено на морско ниво (slp), геопотенциална височина (hgt) на 700 и 850 hPa, температура на въздуха (air) на 850 hPa, специфична (shum) и относителна (rhum) влажност, на 700 и 850 hPa.

Същност на хомогенните HMMs:

Нека $\mathbf{R}_t = (r_t^1, \dots, r_t^n)$ е n -мерен бинарен вектор, характеризиращ появата на валеж в n валежомерни станции: $r_t^i = 1$ ако в денят t вали в станцията i и $r_t^i = 0$ в противен случай за $i = 1, \dots, n$ и $t = 1, \dots, T$, където $T=365 \times NY$, NY е броят на

годините, включени в изследването. Да въведем помощен дискретен случаен процес S_t , който за всяко t приема само една от възможните стойности $1, 2, \dots, k$, съответстващи на броя състояния (типичните валежни обстановки над територията на България). Случайният процес S_t не се наблюдава директно, поради което в литературата се използва терминът скрит от английски hidden, unobservable или latent. Нека $\mathbf{R}_1^T = \{\mathbf{R}_1, \dots, \mathbf{R}_T\}$ и $S_1^T = \{S_1, \dots, S_T\}$. Ще предполагаме, че S_t е хомогенен Марковски процес. Тогава, хомогенният Марковски модел с неяви състояния се дефинира със системата от вероятностни уравнения:

$$\Pr(\mathbf{R}_t | S_1^T, \mathbf{R}_1^{t-1}) = \Pr(\mathbf{R}_t | S_t) \quad (1)$$

$$\Pr(S_t | S_1^T) = \Pr(S_t | S_{t-1}) \quad (2)$$

и началното състояние на системата $\Pr(S_1)$.

Уравнението (1) се нарича модел на „условна независимост“ - появата на валеж $\mathbf{R}_t$ в деня t се определя от типа на валежната обстановка S_t . Уравнението (2) означава, че състоянието на Марковската верига в деня t зависи само от състоянието на редицата в предходния ден $t-1$. Различни НММ могат да се дефинират в зависимост от параметризацията на вероятността $\Pr(\mathbf{R}_t | S_t)$.

При направените предположения, пространственият модел за появата на валеж за n станции при дадено състояние (валежна обстановка) s в деня t се дефинира:

$$\Pr(\mathbf{R}_t = \mathbf{r} | S_t = s) = \prod_{i=1}^n p_{si}^{r_i^i} (1 - p_{si})^{1-r_i^i}, \quad (3)$$

където r_i^i приема стойност 0 или 1 в зависимост от това дали денят t е без или с валеж, $p_{si} = \exp(a_{si}) / (1 + \exp(a_{si}))$ са логистичните вероятности за поява на валеж в станция i и състояние s , a_{si} са неизвестни параметри и $n=32$ е броят на станциите. Тогава, функцията на разпределение на денонощните валежни суми се дефинира както следва

$$L(\theta) = \Pr(\mathbf{R}_1^T | \theta) = \sum_{S_1, S_2, \dots, S_T} \Pr(\mathbf{R}_1^T, S_1^T | \theta) = \sum_{S_1, S_2, \dots, S_T} \Pr(S_1) \prod_{t=2}^T \Pr(S_t | S_{t-1}) \Pr(\mathbf{R}_t | S_t), \quad (4)$$

където $\theta = (a, \gamma)$ е неизвестен вектор от параметрите на модела,

$\Pr(S_t | S_{t-1}) = (\gamma_{ij})_{k \times k}$ е матрица с елементи $\gamma_{ij} = \Pr(S_t = j | S_{t-1} = i)$ за $i, j = 1, \dots, k$.

Броят на неизвестните параметри на НММ модела, описан с уравненията (3) и (4) е $k(k-1) + 3kn$, където k и n са съответно броят на състоянията на процеса

S_t и броят на валежомерните станции, включени в изследването. Оценките на неизвестните параметри на модела се получават като се максимизира функцията на правдоподобие $L(\theta)$, съгласно Zucchini and MacDonald (2009). За избор на модел с оптимален брой параметри се използва информационният критерий на Бейс: $BIC = -2\log L(\hat{\theta}) + npar \log T$, където $\hat{\theta}$ е оценката на стойността, в която функцията на правдоподобие $L(\theta)$ достига максимум, $npar$ е броят на параметрите на модела. BIC служи за избор на модел измежду краен брой модели, като се избира моделът с най-малката му стойност.

3. РЕЗУЛТАТИ

Максимизирайки функцията на правдоподобие (4), с модели от 2 до 10 на брой състояния, бе определен оптимален модел с 8 състояния, за който стойността на BIC е минимална. В Таблица 1 са дадени стойностите на BIC критерия на няколко HMMs.

Таблица 1. Сравнение на няколко HMMs за поява на валеж според BIC критерия

Table 1. Comparison of several HMMs of precipitation occurrence according to the BIC values

Модел	Брой състояния	BIC
HMM.06	6	4.7873e+05
HMM.07	7	4.7307e+06
HMM.08	8	4.6781e+06
HMM.09	9	4.6640e+06
HMM.10	10	4.6621e+06

В резултат на максимизирането на функцията на правдоподобие (4) са получени оценките на преходните вероятности на Марковската верига, които са дадени в Таблица 2. Например, вероятността за преход от състояние 1 в момента $t-1$ ($S_{t-1}, 1$) към състояние 1 в момента t ($S_t, 1$) е равна на 0.70. Аналогично се интерпретират останалите преходни вероятности.

Въз основа на оценените условни вероятности и алгоритъма на Витерби (Viterbi) се определя редицата от най-вероятни състояния на Марковската верига

Таблица 2. Матрица на преходните вероятности на хомогенната HMM с 8 състояния

Table 2. Transition probability matrix of the homogeneous HMM with 8 states

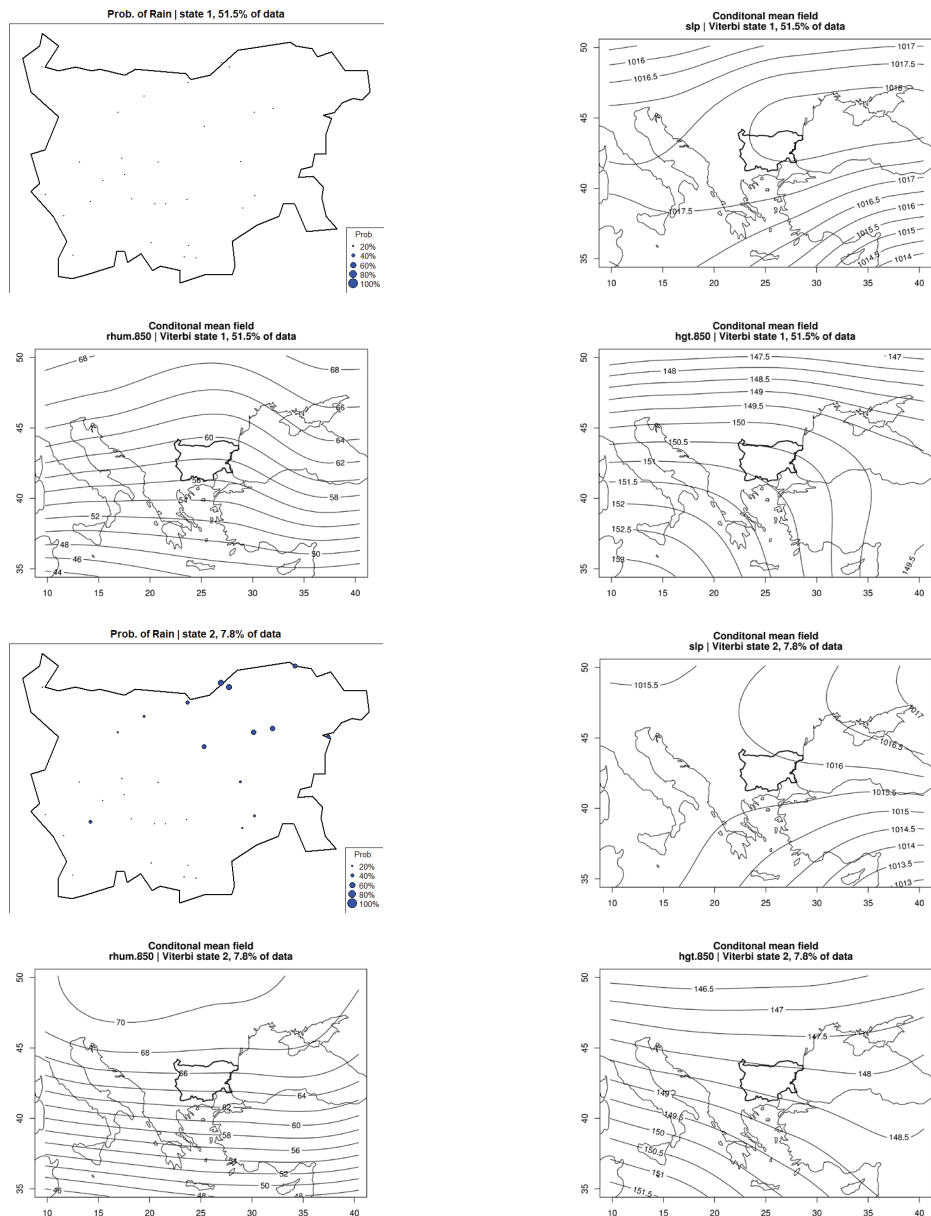
състояние		към							
		S _t .1	S _t .2	S _t .3	S _t .4	S _t .5	S _t .6	S _t .7	S _t .8
о т	S _{t-1} .1	0.70	0.09	0.05	0.02	0.03	0.05	0.02	0.03
	S _{t-1} .2	0.33	0.22	0.08	0.07	0.06	0.12	0.03	0.10
	S _{t-1} .3	0.25	0.15	0.13	0.10	0.11	0.08	0.08	0.10
	S _{t-1} .4	0.11	0.06	0.06	0.27	0.20	0.07	0.14	0.08
	S _{t-1} .5	0.25	0.11	0.06	0.11	0.17	0.10	0.14	0.05
	S _{t-1} .6	0.27	0.12	0.04	0.10	0.10	0.19	0.06	0.12
	S _{t-1} .7	0.47	0.11	0.06	0.03	0.07	0.09	0.14	0.04
	S _{t-1} .8	0.12	0.10	0.06	0.24	0.11	0.13	0.08	0.16
стабилни състояния		0.515	0.078	0.107	0.066	0.053	0.055	0.064	0.062

като решение на оптимизационната задача (Zucchini and MacDonald, 2009):

$$\arg \max_{1 \leq i_1 \dots i_T \leq k} \Pr(\{S_1 = i_1, \dots, S_T = i_T \mid R_1^T = r_1^T\})$$

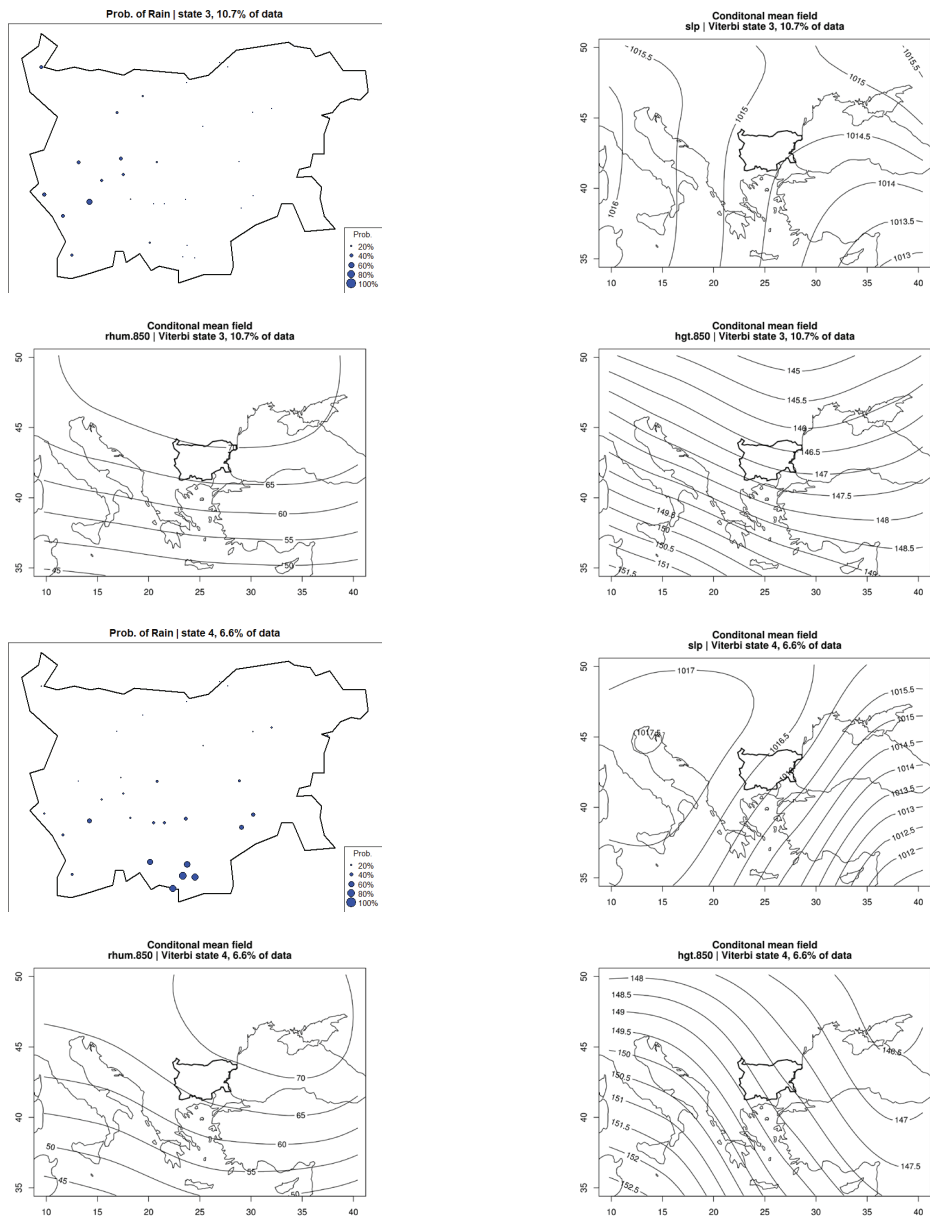
където i_j за $j=1, \dots, T$ са възможните състояния на Марковския процес. Вследствие на тази редица от най-вероятни състояния, дните от 1 до T се класифицират в 8 групи. Всяко състояние се характеризира с 32-мерен вектор от оценени вероятности за валеж и съответства на валежна обстановка. На Фигура 2 и нейните продължения са дадени по два панела, състоящи се от по 4 графики. На Фигура 2 горе вляво на всеки панел е представено едно от 8-те състояния валеж - Prob. of Rain | state, съответният му номер и честотата на поява в проценти за периода 01.01.1960-31.12.2018 г. Вероятностите за поява на валеж за всяка станция за всяко състояние са дадени като кръгчета с диаметри, пропорционални на стойностите им. Композиционните полета на slp, rhum.850 hPa и hgt.850 във всеки панел са получени чрез осредняване във възлите за всяка от 8-те групи, формирани от 8-те състояния. Вижда се много добро съответствие между състоянията на модела (валежна обстановка) с композиционните полета. Състоянията 1 и 8 се характеризират с ниски и високи стойности на вероятностите за валеж във всяка от станциите. Сухото състояние 1, свързано с композиционните полета на slp, rhum.850 hPa и hgt.850, се характеризира с антициклонални обстановки над Балканския полуостров и ниска влажност. Състоянията 5-8 съответстват на валежите от Средиземноморски циклони и се характеризират с по-ниски стойности на налягането и по-високи стойности на относителната влажност.

С модела са генерирани денонощни валежни суми от 200 годишни редици за периода 01.01.1960-31.12.2018 г. Тези редици от данни са сравнени с историческите по отношение на синхронност на валеж в групи от станции и последователен брой дни с валеж.



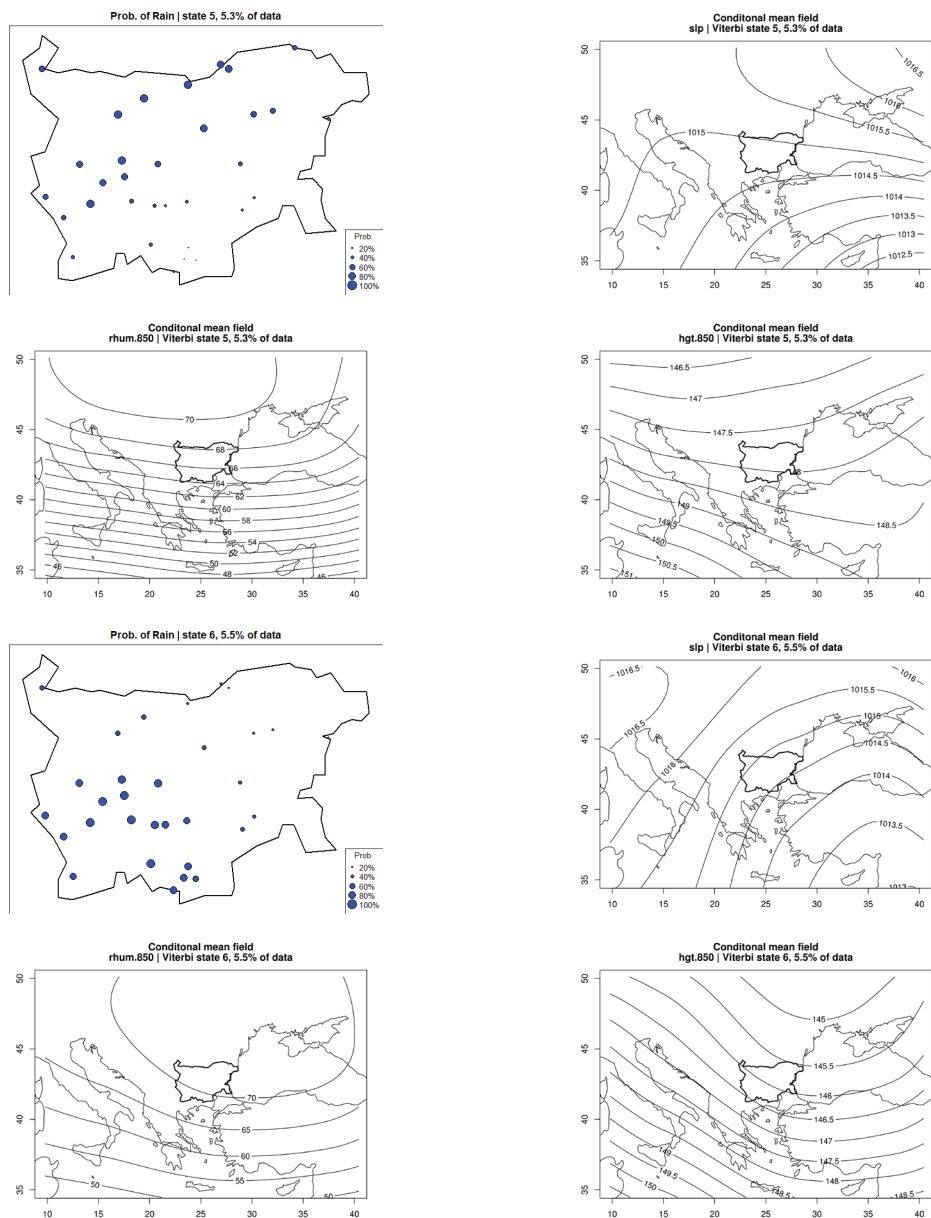
Фиг. 2. Вероятности за поява на валеж, диаметрите на кръгчетата са пропорционални на вероятността за валеж и варират в интервала $[0,1]$. Осреднени композиционни полета на slp, rhum.850 hPa и hgt.850 за всяко състояние.

Fig. 2. Precipitation occurrence probabilities, the diameters of circles are proportional to the probability of precipitation and vary in the interval $[0,1]$. Composite slp, rhum.850 hPa and hgt.850 fields averaged over all days classified under each state.



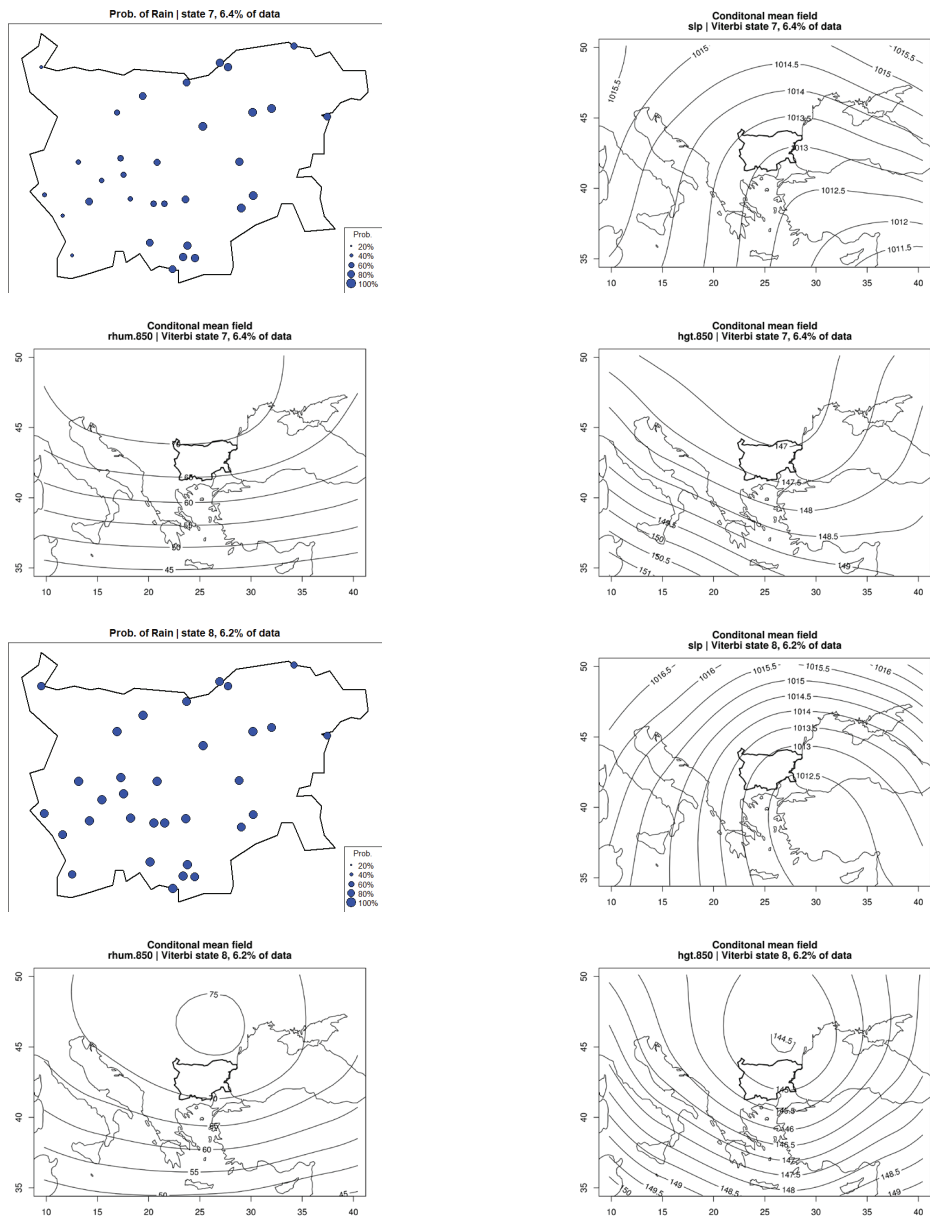
Фиг. 2. (продължение) Вероятности за поява на валеж, диаметрите на кръгчетата са пропорционални на вероятността за валеж и варират в интервала $[0,1]$. Осреднени композиционни полета на slp, rhum.850 hPa и hgt.850 за всяко състояние.

Fig. 2. (continued) Precipitation occurrence probabilities, the diameters of circles are proportional to the probability of precipitation and vary in the interval $[0,1]$. Composite slp, rhum.850 hPa and hgt.850 fields averaged over all days classified under each state.



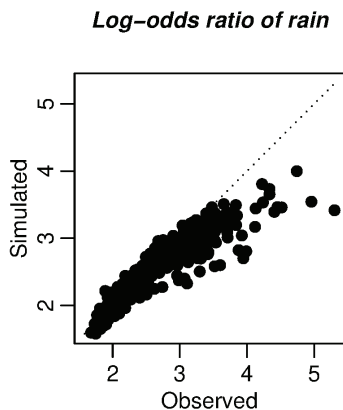
Фиг. 2. (продължение) Вероятности за поява на валеж, диаметрите на кръгчетата са пропорционални на вероятността за валеж и варират в интервала $[0,1]$. Осреднени композиционни полета на slp, rhum.850 hPa и hgt.850 за всяко състояние.

Fig. 2. (continued) Precipitation occurrence probabilities, the diameters of circles are proportional to the probability of precipitation and vary in the interval $[0,1]$. Composite slp, rhum.850 hPa and hgt.850 fields averaged over all days classified under each state.



Фиг. 2. (продължение) Вероятности за поява на валеж, диаметрите на кръгчетата са пропорционални на вероятността за валеж и варират в интервала $[0,1]$. Осреднени композиционни полета на slp, rh_{um.850} hPa и hgt.850 за всяко състояние.

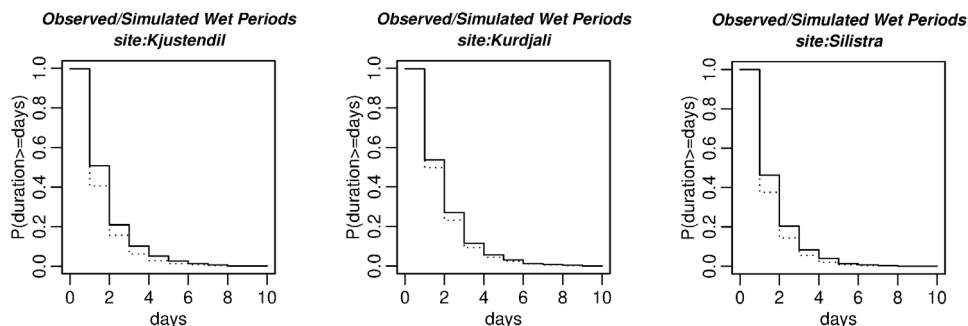
Fig. 2. (continued) Precipitation occurrence probabilities, the diameters of circles are proportional to the probability of precipitation and vary in the interval $[0,1]$. Composite slp, rh_{um.850} hPa and hgt.850 fields averaged over all days classified under each state.



Фиг. 3. Пространствена мярка за сходство между исторически и симулирани по модела данни за поява на валеж.

Fig. 3. Observed versus model based log odds ratios.

На Фигура 3 е дадена пространствената мярка за сходство на измерените и генерирани по модела данни за поява на валеж (1 ако вали и 0 ако не вали) за всевъзможните двойки станции. Мярката за сходство се дефинира чрез отношението $\log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right)$, където n_{11} и n_{00} са броят на дните, в които вали или не вали едновременно в станциите i и j , съответно, докато n_{10} и n_{01} са броят на дните с валеж в станция i , но без валеж в станция j и обратно. Забелязва се много добро сходство и синхронност на валежа между историческите и моделни данни, тъй като отношенията за повечето станции следват диагонала на квадрата. Забелязва се известно различие с валежите от станцията, разположена на връх Мусала, което се дължи на нейния различен климат.



Фиг. 4. Разпределение на продължителност от валежни дни по наблюдавани (плътни линии) и симулирани (пунктирни линии) данни за избрани станции.

Fig. 4. Observed (solid lines) and model-based (dotted lines) wet spells distribution at selected stations.

От графиките на Фигура 4 се вижда, че за станции Кюстендил, Кърджали и Силистра оценките на вероятностите за продължителност на определен брой от валежни дни, пресметнати по наблюдаваните и симулирани данни, са сходни. Аналогични резултати са получени за останалите валежимерни станции. Това показва, че моделът запазва времевата зависимост в данните, което е важно за приложенията му в практиката.

4. ЗАКЛЮЧЕНИЕ

Показано е че, хомогенният НММ е полезно средство за моделиране на пространствени валежи. Предимството на този тип модели е, че чрез тях е възможно да бъдат генерирани синхронни редици от поява на валеж в група от станции. В процес на създаване е стохастичен модел за интензивността на валеж по данни за дните, в които е наблюдаван валеж, с експоненциалното и гама разпределенията и с вероятностна смес от две експоненциални разпределения. В процес на създаване е и нехомогенен НММ, в който за предиктори се използват данните от реанализите. Предстояща и актуална задача е разширено изследване на характеристиките на моделите с методите на имитационното моделиране по данни на обучаващата извадка и тестови периоди.

БЛАГОДАРНОСТИ

Това изследване е финансирано от Национална програма „Млади учени и постдокторанти“ на Министерството на образованието и науката (МОН).

ЛИТЕРАТУРА

- Neytchev, P., Zucchini, W., Hristov, H. and Neykov, N.M. (2006) Development of a multisite daily precipitation model for Bulgaria using hidden Markov models. In: Proc. of the XXIIIrd conference of Danubian countries on the hydrological forecasting and hydrological bases of water management. Belgrade, Serbia, 28-31 August, S. Bruk and T. Petkovic (eds.).
- Neytchev, P., Zucchini, W., Neykov, N. and Hristov, H. (2011) Drought study based on multi-site daily precipitation model applied for Bulgaria. *Bulg. J of Meteorology and Hydrology*. 16, 54-61.
- Neykov, N.M., Neytchev, P.N., Zucchini W. and Hristov, H. (2012). Linking atmospheric circulation to daily precipitation patterns over the territory of Bulgaria. *Environ. Ecol. Stat.*, 19, 249-267.
- Zucchini, W. and MacDonald, I.L. (2009). Hidden Markov models for time series. An introduction using R. Chapman & Hall/CRC, London.