3rd National Congress on Physical Sciences, 29 Sep. – 2 Oct. 2016, Sofia
Section: Physics of Earth, Atmosphere and Space

Stochastic Downscaling of Daily Precipitation over Bulgaria trough Hidden Markov Models: Precipitation Amount Simulation at Site That Does Not Belong to the Network of Stations

N. Neykov P. Neytchev

National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, Sofia 1784, Bulgaria

Abstract. An eight states multi-site non-homogeneous hidden Markov model that links daily precipitation amounts data at a network of 31 stations broadly covering the territory of Bulgaria to large-scale atmospheric patterns is developed. A technique to simulate daily precipitation amount at sites that belong to the network of stations, not explicitly included in this multi-site model is considered. The results show that the downscaled simulations reproduce well the observed precipitation amount, the wet and dry spell length distributions.

1 Stochastic Daily Precipitation Model Using NHMM

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. These models can be characterized as at-site and multi-site models (un)conditional of the atmospheric variables as well [1, 2]. Once the precipitation model has been calibrated for a given territory one can use it to generate long sequences of artificial daily precipitation data. These sequences can be used to estimate statistics relating to precipitation events in exactly the way one would do so if a long sequence of precipitation data were available.

There are two different main approaches of relating daily precipitation to synoptic atmospheric patterns in development of the multi-site precipitation models, the so called weather state models.

The traditional weather state models are discussed in [3–6], just to name a few. The synoptic atmospheric patterns of this approach have been either subjectively or objectively derived using principal components, canonical correlation analysis, fuzzy rules, neural networks, correlation-based pattern, recognition techniques, analogue procedures, etc. Introducing such an intermediate layer (the weather patterns) precipitation is linked to the circulation patterns using conditional probabilities. Although its physical transparency the above approach has been criticized that suffers from some limitations, e.g., the veracity of weather pattern models depends upon the chosen weather classification system.

The second approach for downscaling uses a Non-homogeneous Hidden Markov Model (NHMM) to simulate precipitation occurrence or amounts. The NHMM relates synoptic-scale atmospheric circulation variables through a finite number of hidden (unobserved, latent) precipitation patterns (states) to multi-site, daily precipitation occurrence or amounts data. The NHMM determines the most distinct patterns in a daily multi-site precipitation occurrence records rather than patterns in atmospheric circulation [7]. These patterns (precipitation states) are then defined as conditionally dependent on a set of atmospheric predictor variables. Unlike other downscaling techniques based on classification schemes these states are not defined a priori. A first-order Markov process defines the daily transitions from precipitation state to another. The process is described as non-homogeneous as the transition probabilities are conditional on a set of atmospheric circulation predictors. The atmospheric predictors may include raw variables such as mean sea level pressure (mslp) or derived variables such as mslp gradient. In this way, the NHMM captures much of the spatial and temporal variability of the precipitation occurrence process. Model selection involves sequential fitting of several NHMMs with an increasing number of weather states and atmospheric predictors. The fits are evaluated in terms of the physical realism and distinctness of the identified precipitation states as well as a Bayesian information criterion (BIC). The objective is to select a NHMM that minimizes the BIC, thus identifying a relatively parsimonious model that fits the data well. The most likely precipitation state sequence is obtained from the selected NHMM using the Viterbi algorithm [8]. This permits the assignment of each day to its respective state. The ability to classify days into states that are distinct in terms of precipitation as well as synoptic situation means that the realism of the states is directly interpretable in terms of regional synoptic climatology. Due to these states the spatial precipitation dependence can be partially or completely captured. The advantage of using the NHMM is that it can simulate stochastically the effects of large-scale atmospheric circulation on local weather at multiple gauge stations, i.e., a multi-site is constituted in this way. We note that the at-site daily precipitation models cannot be easily extended to multiple sites as that requires modeling spatial dependencies between the stations [9]. Moreover, there is no guarantee that these models can generate synchronized daily precipitation occurrence sequences in a small size region.

The non-homogenous hidden Markov Models (NHMM) have found widespread application in meteorology and hydrology in Australia, North and South America in studies of climate variability or climate change, statistical downscaling of daily precipitation from observed and numerical climate model simulations. For instance, [10] presented the NHMM for climate change condition in southwestern Australia. The NHMM were extended in [11] by incorporating rainfall amount. The results showed that the extended NHMM accurately simulates the survival curves for dry (wet) spell lengths, wet day probabilities, daily rainfall amount distribution, and intersite correlations for daily rainfall amounts. The NHMM was used in [12] to model the rainfall amount independently at each rainfall station as gamma deviates with gauge-specific parameters in Washington. They verified that the model responded to shifts in atmospheric circulation from a reserved data set. In [13] a NHMM rainfall occurrence was applied over Northeast Brazil and it was found that interannual variability in the frequency of occurrence of dry spells could be simulated well. The ability of the extended NHMM to reproduce observed interannual and interdecadal rainfall variability when driven by observed and modeled atmospheric field was investigated [14]. The NHMM was applied to 11 stations over North Queensland in [15]. The results showed that the model was able to simulate accurate station level simulations of the interannual variability of daily rainfall amount and occurrence. A 4-state HMM is developed in [16] to a network of 13 stations in central western India. Their results have shown enough evidence to the HMM representation of monsoon spatio-temporal variability. The statistical modeling techniques with NHMMs are given in [8] whereas specific details about fitting precipitation data can be found in [12, 20] and [21]. A recent review about multi-site daily precipitation data conditional on the atmospheric circulation data is given in [22].

2 Multi-site Daily Precipitation Model for the Territory of Bulgaria

Since 2005 we have been investigating the usage of the NHMM to link synoptic-scale atmospheric circulation variables to daily precipitation data at a network of rain gauge stations via several hidden (unobserved) states, [17] and [18]. The evolution of these states is modeled as a first-order Markov process with state-to-state transition probabilities conditioned on some indices of the atmospheric variables. Due to these states the spatial precipitation dependence can be partially or completely captured. The NHMM was fitted and independently tested to daily precipitation at 31 rain gauge stations covering broadly the territory of Bulgaria. At each site a 40-year record (1960-2000) of daily precipitation amounts is modeled. For the warmer half-years only the days with precipitation amounts greater then 2mm were included in the study to drop out the days with convective rainfalls. The data consists of all dry and wet days for the cold half-years according to the cutoff $c \geq 0.1mm$ value and all dry and wet days according to the cutoff $c \geq 2$ mm for the warmer half-years. The number of days included in the study is 11323. The atmospheric data consists of daily mean sea-level pressure, geopotential height at 850 hPa, air temperature at 850 hPa and relative humidity at 700 hPa on a $2.5^\circ \times 2.5^\circ$ grid based on NCEP-NCAR reanalysis dataset covering the Europe-Atlantic sector 30°E-60°E, 20°N-70°N for the same period. The first 30 years data were used for model fitting purposes while the remaining 10 years were used for model evaluation.

2.1 Data reduction of dimensionality

The singular value decomposition (SVD) technique, described by [19], was applied to the correlation matrix between the precipitation amounts at each of the 31 sites and the corresponding atmospheric variables at 276 nodes. From Table 1 it is seen that a few summary variables explain most of the correlation between each field and the precipitation process. For instance, the first summary variable of the coupled patterns (atmospheric - precipitation fields) included in the study, accounts for more than 90% of the correlation. However, these summary variables are highly correlated. Therefore a standard principal components analysis was applied to the composed field based on the first summary variable of each of these fields. As a result of this, three new scalar indices (again denoted by F1, F2 and F3) were constructed which account for 95% of the total variance of the composed field as shown in Table 2.

Some occurrence and amounts homogeneous and nonhomogeneous hidden Markov models were calibrated closely following the hierarchical

Table 1: Cumulative proportion of correlation explained by the summary variables

eigenvalues:	% 1st	% 2nd	% 3rd
rhum.700	93.01	96.89	98.34
rhum.850	92.15	96.42	98.07
air.850	94.22	97.82	99.29
hgt.850	94.63	97.40	98.75
slp	93.09	97.79	99.33

Table 2: Cumulative proportion of variance explained by the 1st summary variables of the composite field based on slp, hgt.850, rhum.700, rhum.850 and air.850

names	F1	F2	F3	F4	F5
eigenvalues:	% 1st	% 2nd	% 3rd	% 4rd	% 5th
composite	73.00	86.00	95.00	99.99	100.00

procedure over all days of 1960-1990 period. As could be expected the 8 states NHMM stochastic precipitation model calibrated without Plovdiv station is almost the same as multi-8.1.0.0 developed for the annual 8 states NHMM model for the 1960-1990 period calibrated on all 30 stations. It is characterized by a smaller number of unknown parameters. On the other hand this multi-8.1.0.0 amount model is quite satisfactory not only from formal statistical view point because of the lowest BIC value but because of the distinctiveness and physical interpretability of the weather states.

Our presentation will report some of our results and findings about the calibrated multi.8.1.0.0 model in [18]. The plots of 1st, 2nd and 3rd columns of Figure 1 present the NHMM estimated precipitation occurrence and intensity states and the composite mean sea-level pressure (slp) patterns associated with these states. Each state is related with a precipitation occurrence patterns and are found to be physically interpretable in terms of regional climatology. The diameters of circles indicate daily precipitation occurrence probabilities and intensities at each site with the largest circle 1.0 and above 15mm, respectively. About the slp patterns associated with the precipitation states we note that each day is first classified into its most likely state according to the Viterbi algorithm and, second, all days in a particular state are then averaged at each grid node for the atmospheric variables to obtain the corresponding composite fields. The precipitation states 1 and 8 occur on 46% and

Table 3: Comparison of homogeneous and nonhomogeneous annual HMMs using the conditional independence occurence and amount model for $P(\mathbf{R}_t|S_t)$. The models are sorted according BIC values. The number of the days included in the study is 11323

Model	Number of model predictors	Number of model pameters	Negative log-likelihood	BIC
Amount.multi.8.1	F1	830	347543.3	702834.3
Amount.HMM.8	0	776	358475.2	724194.0

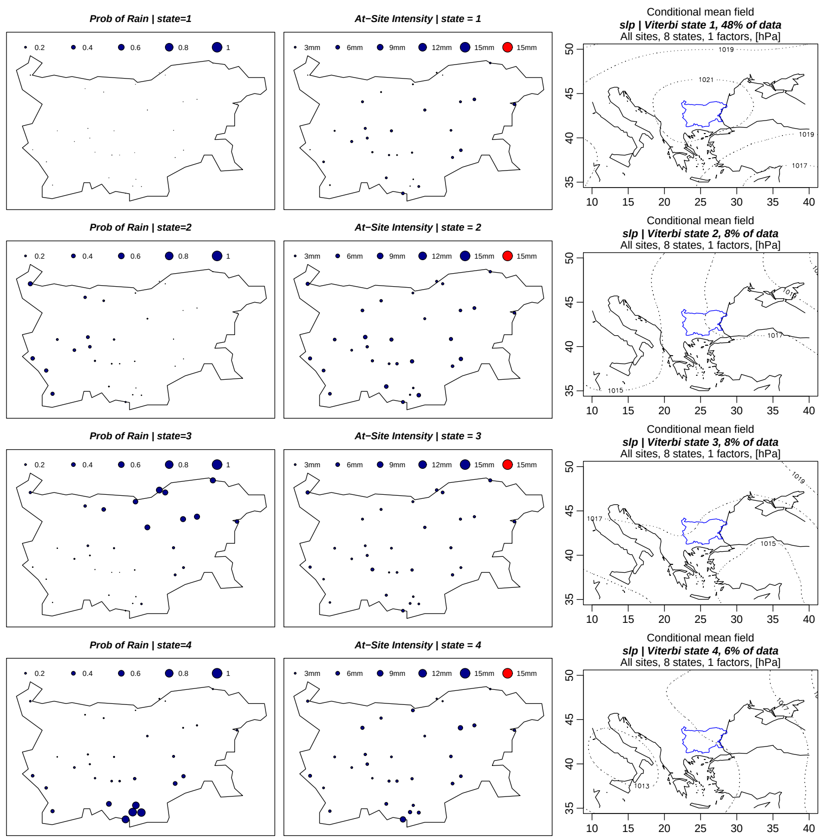

Figure 1: Precipitation occurrence probabilities (diameters of circles proportional to probability of a wet-day with the largest circle 1.0), at-site intensity (diameters of circles proportional to intensity with the largest circle greater than 15mm) and composite sea-level pressure fields.

12% of days, respectively. The synoptic pattern associated with state 1 is a typical dominant high pressure system centered at Balkan Peninsula. The states from 2 till 8 are characterized by various spatial precipitation probabilities. For instance, the circulation pattern related with state 2 can be associated with Mediterranean cyclones centered over Northern Italy moving to Hungary due to southwestern flow. State 3 exhibits high probability of rain at the southeast stations but low probability in the rest of the country. The state 4 demonstrates the well known fact that when the axis of the upper-level trough is over Eastern Bulgaria and the Black sea the precipitation events are more likely in the North-Eastern Bulgaria. The weather states from 5 till 8 are associated with a depres-

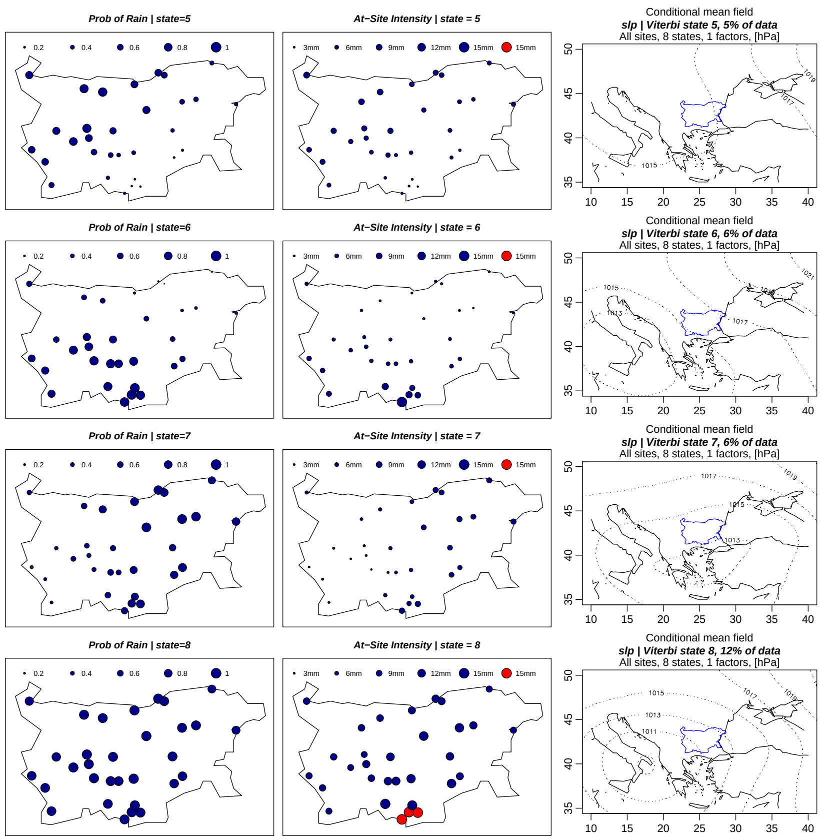

Figure 1 (cont.): Precipitation occurrence probabilities (diameters of circles proportional to probability of a wet-day with the largest circle 1.0), composite sea-level pressure (hPa), 850 hPa geopotential height (m), At-site Intensity (diameters of circles proportional to intensity with the largest circle greater than 15mm), air temperature and 700hPa relative humidity fields averaged over all days classified under each weather state for the multi.8.1 model.

sion centered over the central Mediterranean and the Southern Italy in the mean sea-level pressure field and an upper-level trough with different amplitude and tilt in the geopotential height at 850 hPa. The model adequately accounts what the general sense suggests that the most likely precipitation events occur below the most humid middle troposphere as well as within the polar frontal zone which follows from the patterns associated with the relative humidity at 700 and 850 hPa and the air temperature field at 850 hPa (these plots are not included due to space limitation).

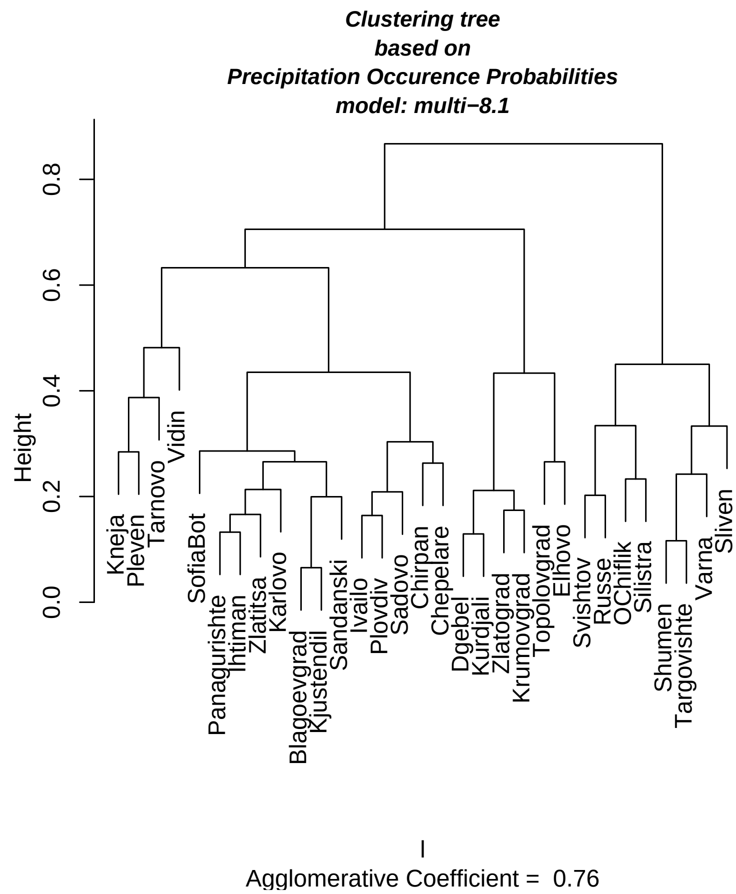

Figure 2: Clustering of the 31 rain stations according to the estimated precipitation occurrence probabilities for the eight state conditional independence model.

Clusters of rain gauges based on the precipitation occurrence probabilities parameter estimates of model.8.1.0.0, the occurence component of the model, are presented in the plots of Figure 2. It is seen that the clusters belong to distinct geographical climate regions of Bulgaria. Therefore a method to simulate precipitation amount at sites that belong to the network of stations, not explicitly included in the model can be developed on the basis of this regionalization. We propose estimating the parameters of such a station of interest as a weighted linear combination of parameter estimates of the neighbouring rain gauge stations according to occurrence clusters. The weights can be the inverse distances between the station of interest and the neighbouring precipitation stations. In this way, this would make the calibrated NHMM available at sites where precipitation records are not available.

3 At-Site Precipitation Amount Simulation Which Is Not from the Stations Network

In the following we will demonstrate the potential applicability of this approach dropping out Plovdiv station from the stations network.

The plots in Figure 3 show quantile-quantile plots in logarithmic scale of the observed versus model-based precipitation intensities for Plovdiv

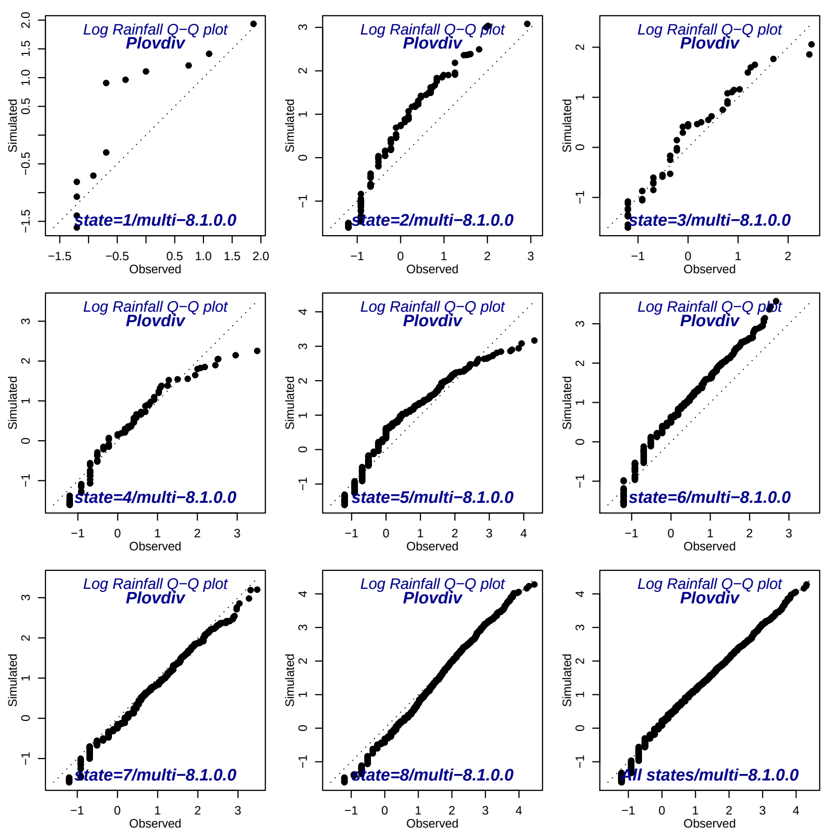

Figure 3: QQ-plot in logaritmic scale of simulated vs observed precipitation intensities for Plovdiv station. The simulated and historical precipitation intensities are calculated over all days classified under each weather state separately and totaly. The 8-state NHMM daily model is calibrated for the 1960-1990 period on the whole year data.

station based on multi-8.1.0.0 model. Recall that simulated and historical precipitation intensities are calculated over all days classified under each weather state separately and totally. Figure 4 are similar, however, the simulated precipitation intensity data are based on the precipitation model parameter estimate defined as a weighted linear combination of the parameter estimates of Plovdiv neighbouring gauge sta-

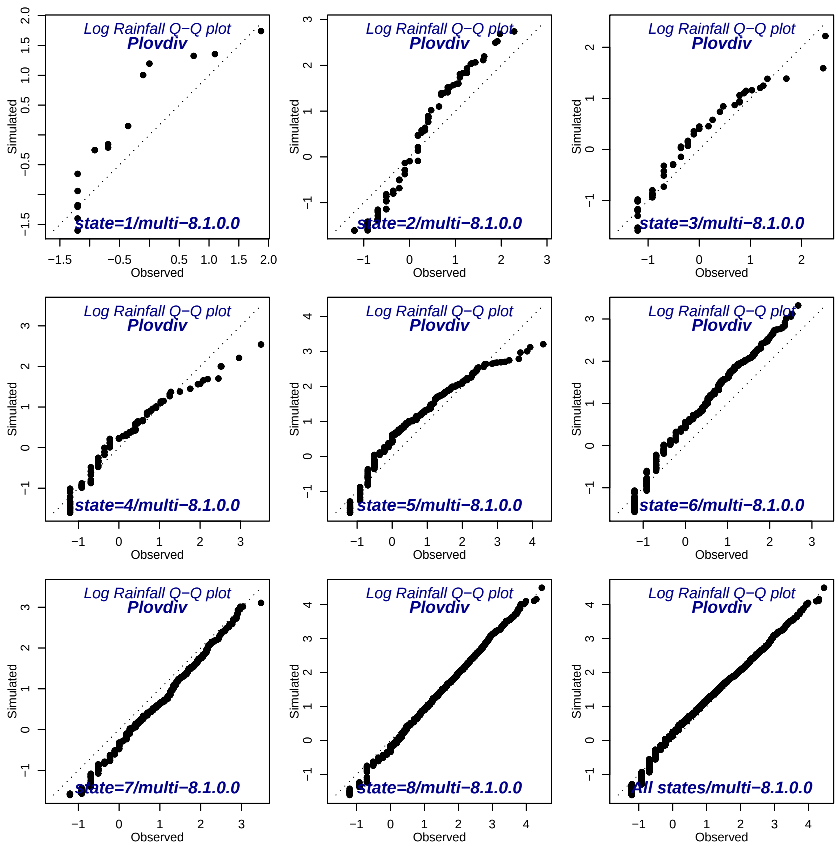

Figure 4: QQ-plot in logaritmic scale of observed vs simulated precipitation intensities for Plovdiv station for the historical period 1960-1990. The simulated precipitation intensity data are based on the parameter precipitation model estimate defined as a weighted linear combination of Plovdiv neighbouring gauge stations parameter estimates according to occurrence and intensity clusters due to the 8-state NHMM daily model calibrated for the 1960-1990 period on the whole year data. The artificial-based and historical precipitation intensities are calculated over all days classified under each weather state separately and totaly.

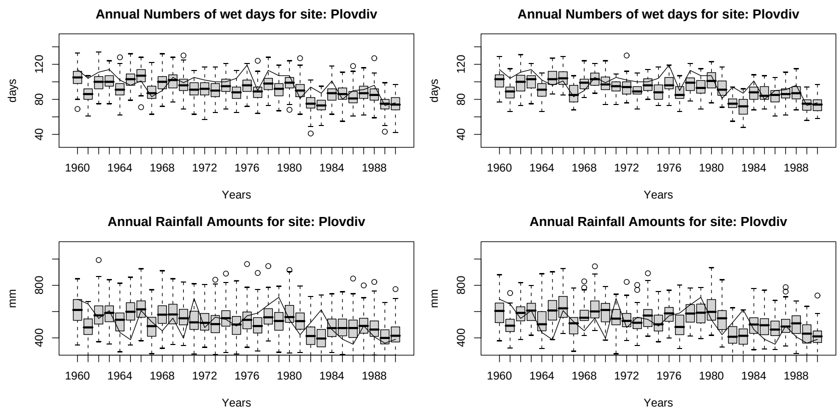

Figure 5: On the left-hand two plots are given the observed (solid line) vs downscaled (box-plots) 1960-1990 annual wet-day frequencies and amounts for Plovdiv. The box-plots depict the range of 500 simulation trials (the edges of the box represent the 25 percentile and the 75 percentile of the simulations) produced by the 8-state NHMM. The same info is presented on the right-hand two plots, however, the downscaled (box-plots) 1960-1990 annual wet-day frequencies and amounts data are based on the parameter precipitation model estimate defined as a weighted linear combination of Plovdiv neighboring gauge stations parameter estimates according to occurrence and intensity clusters due to the 8-state NHMM daily model calibrated for the 1960-1990 period on the whole year data.

tions (Sadovo, Chirpan, Ivailo) according to occurrence clusters due to the 8-state NHMM daily model calibrated for the 1960-1990 period on the whole year data. Results of similar standard are obtained about other stations using this technique.

Plots in the Figure 5 present conditional simulations, the range in simulated annual wet-day frequencies and total rainfall amounts about station Plovdiv. The visual assessment confirm that the approach does well at capturing the interannual variability.

The wet spells distributions of the Plovdiv stations are presented in the plots of Figure 6. It is seen that the temporal correlation in the historical data is also well captured by the proposed methods. Results of similar standard based are obtained for the remaining stations.

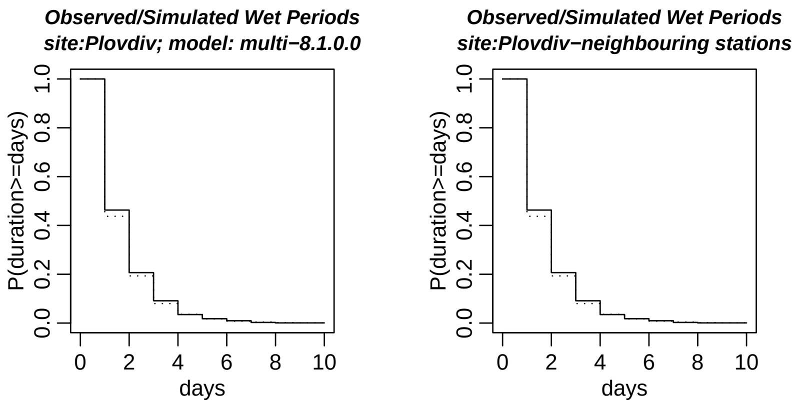

Figure 6: On the left-hand plot are given the observed and model-based wet spells distribution at Plovdiv station for the historical period 1960-1990. On the right-hand plot is given the the same info, however, the parameter precipitation model estimate is defined as a weighted linear combination respectively of Plovdiv neighbouring gauge stations parameter estimates according to occurrence and intensity clusters due to the 8-state NHMM daily model calibrated for the 1960-1990 period on the whole year data.

4 Conclusions

The space-time precipitation model can be used to generate simulations of precipitation amounts that incorporate synoptic atmospheric information. The hidden Markov model assumptions simplify the temporal and spatial structures to be parameterized, since the common weather state accounts for the temporal dependence and much of the spatial correlation between rain gauges. Several possible improvements to the model are currently under investigation, including more realistic spatial dependence structures and reduced parameterizations.

Models like the NHMM can be used to study the effect of climate variability. Repeated GCM simulations under current climate conditions can constitute different realizations of the atmospheric fields included in the model. The NHMM can be used to generate occurrences and amounts for each realization, thereby downscaling the effect of the variability in the synoptic-scale variables to precipitation. The output of GCM runs under altered climate conditions can serve as input into the downscaling model. Thus, the effects of the altered climate scenario could be downscaled to the local-scale precipitation processes by generating precipitation occurrences and amounts from the NHMM. Promising results in this direction are given in [10] and [11].

References

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

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

[3] A. Bardossy and E. J. Plate (1992) *Water Resour. Res.* **28** 1247.

[4] C. M. Goodess and P. D. Jones (2002) *Int. J. Climatol.* **22** 1593.

[5] C. M. Goodess and J. Palutikof (1998) *Int. J. Climatol.* **18** 1051.

[6] J. Stehlik and A. Bardossy (2002) *J. Hydrology* **256** 120.

[7] W. Zucchini and P. Guttorp (1991) *Water Resour. Res.* **27** 1917.

[8] I. L. MacDonald and W. Zucchini (1997) *Hidden Markov and Other Models for Discrete-valued Time Series,* Chapman & Hall.

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

[10] S. Charles, B. Bates, and P. Hughes (1999b) *J. Geoph. Res.* **104** 31657.

[11] S. Charles, B. Bates, P. Whetton, and P. Hughes (1999a) *Clim. Res.* **12** 1.

[12] E. Bellone, J. P. Hughes, and P. Guttorp (2000) *Clim. Res.* **15** 1.

[13] A. W. Robertson, S. Kirshner, and P. Smyth (2004) *J. Climate* **17** 4407.

[14] S. Charles, B. Bates, I. Smith, and P. Hughes (2004) *Hydrol. Process.* **18** 1373.

[15] A. W. Robertson, S. Kirshner, P. Smyth, and B. C. Bates (2005) *Q. J. R. Meteorol. Soc.* **131** 1.

[16] A. M. Greene, A. W. Robertson, and S. Kirshner (2008) *Q. J. R. Meteorol. Soc.* **134** 875.

[17] P. N. Neytchev, W. Zucchini, H. Hristov, and N. M. Neykov (2006) in *Proc. of the XXIIIrd conf. of Danubian countries on the hydrological forecasting and hydrological bases of water management,* Belgrade, Serbia, 28-31 August, ed. S. Bruk and T. Petkovic.

[18] N. M. Neykov, P. N. Neytchev, W. Zucchini, and H. Hristov (2012) *Environ. Ecol. Stat.* **19** 249.

[19] C. F. Bretherton, C. Smith and J. M. Wallace (1992) *J. Clim.* **5** 541.

[20] J. P. Hughes, P. Guttorp, and S. P. Charles (1999) *Appl. Stat.* **48** 15.

[21] M. Vrac and P. Naveau (2007) *Water Resour. Res.* **43** doi:10.1029/2006WR005308

[22] P. Ailliot, D. Allard, V. Monbet, and P. Naveau (2015) *Journal de la Société Française de Statistique* **156** 101.