Daily Precipitation Generation using a Hidden Markov Model with Correlated Emissions for the Potomac River Basin

Gerson C. Kroiz[1], **Reetam Majumder**[1,*], **Matthias K. Gobbert**[1], **Nagaraj K. Neerchal**[1,2], **Kel Markert**[3], and **Amita Mehta**[4]

[1] Department of Mathematics and Statistics, University of Maryland, Baltimore County, USA
[2] Chinmaya Vishwavidyapeeth, Kerala, India
[3] University of Alabama in Huntsville, USA
[4] Joint Center for Earth Systems Technology (JCET), University of Maryland, Baltimore County, USA

A daily precipitation generator based on a hidden Markov model with Gaussian copulas (HMM-GC) is constructed using remote sensing data from GPM-IMERG for the Potomac river basin on the East Coast of the USA. Daily precipitation over the basin from 2001–2018 for the wet season months of July to September is modeled using a 4-state HMM, and correlated precipitation amounts are generated from a mixture of Gamma distributions using Gaussian copulas for each state. Synthetic data from a model using a mixture of two Gamma distributions for the non-zero precipitation is shown to replicate the historical data better than a model using a single Gamma distribution.

1 Introduction

The modeling and forecasting of seasonal and inter-annual variations in precipitation is used to determine water allocation and resource management for regions that are dependent on precipitation as a primary water source. To this end, statistical and dynamical precipitation generators can be constructed to produce time series of synthetic data representative of the general rainfall patterns within the region. In particular, stochastic weather generators aim to replicate key statistical properties of the historical data like dry and wet stretches, spatial correlations, and extreme weather events.

For multi-site daily precipitation generation, we consider a hidden Markov model with Gaussian copulas (HMM-GC) over the Potomac river basin located on the East Coast of the USA. The basin is the primary source of water for the region and receives large portions of its water supply from rainfall. We use daily data from the GPM-IMERG dataset [1] for the months of July to September from 2001–2018. With a $0.1^\circ \times 0.1^\circ$ spatial resolution, this results in 387 grid points across the basin.

2 The Hidden Markov Model with Gaussian Copulas

Let $\mathbf{R}_{1:T} = \{\mathbf{R}_1, \ldots, \mathbf{R}_t, \ldots, \mathbf{R}_T\}$ be the $M \times T$ matrix of precipitation amounts for a network of M grid points over T days, where $\mathbf{R}_t^{'} = (R_t^1, \ldots, R_t^M)$. Let $S_{1:T} = \{S_1, \ldots, S_t, \ldots, S_T\}$ be the set of hidden (unobserved) weather states, where $S_t \in \{1, \ldots, J\}$. At each location,

$$p[R_t^m = r | S_t = j] = \begin{cases} p_{jm0} & \text{if } r = 0 \\ \sum_{c=1}^{C} p_{jmc} f(r|\alpha_{jmc}, \beta_{jmc}) & \text{if } r > 0 \end{cases} \tag{1}$$

with $p_{jmc} \geq 0$ and $\sum_{c=0}^{C} p_{jmc} = 1$ for all $m = 1, \ldots, M$ and $j = 1, \ldots, J$; $f(\cdot|\alpha, \beta)$ is the density function of a Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. The states arise from a stationary, first-order Markov process. Spatial dependence is captured implicitly by the Markov chain $\{S_t\}$, and precipitation at the M locations for every $\mathbf{R}_t$ are independent given S_t. Details of the model formulation, estimation and simulation when using Exponential distributions can be found in [2, 3].

We fit a 4-state model with two Gamma distributions for the Potomac river basin. We have previously noted [4] that this model underestimates the spatial correlations between grid points due to the densely gridded nature of our remote sensing dataset. This results in unrealistic spatial patterns over the basin which also affects the daily mean and maximum precipitation estimates. We proposed a copula based generator for the precipitation; instead of generating precipitation independently conditional on the state, a Gaussian copula was used to generate correlated amounts. The performance of this HMM-GC and its comparison with the model from Equation 1 can be found in [4]. Further, in [4], we chose a model with two Gamma distributions instead of a single Gamma distribution even though the latter is more commonly used [5, 6] and had a lower Bayesian Information Criterion (BIC) score; this is because the BIC relies upon assumptions that do not hold for the order selection problem [5], and cannot always identify the correct model for a small number of days and a large number of states [7]. Here we discuss the effects of using an HMM-GC with a single Gamma distribution.

* Corresponding author: e-mail reetam1@umbc.edu

Table 1: Statistics based on an HMM-GC with one and two Gamma distributions.

Number of Gamma distributions	Log-likelihood for fitted model	Bayesian Information Criterion (BIC) score	RMSE for proportion of dry days	RMSE for mean monthly precipitation (mm)	Average absolute difference in spatial correlation
1	–9.78e+05	1.991e+06	0.056	29.62	0.166
2	–9.62e+05	1.994e+06	0.053	7.039	0.183

3 Comparison of HMM-GC models for the Potomac River Basin

In Table 1, we see that while the model with the single Gamma distribution (HMM-GC1) has a better (lower) BIC score, using two Gamma distributions (HMM-GC2) provides a better (higher) log-likelihood. The root mean square errors (RMSE) for the proportion of dry days is similar for both models due to a common estimation and generation process. However, the RMSE for the mean monthly precipitation at grid points is much higher for HMM-GC1 compared to HMM-GC2; this suggests that the HMM-GC2 provides more accurate monthly estimates. In Figure 1 which plots the distribution of pairwise spatial correlations between the grid points, we see that though both models underestimate spatial correlations, HMM-GC1 has a higher median as well as a wider range. However, HMM-GC1 also provides negative correlation estimates, whereas the original data has strictly positive correlations. HMM-GC2 does not suffer from this issue.

Finally, Figure 2 shows that HMM-GC1 overestimates the daily maximum basin precipitation, whereas HMM-GC2 has a distribution similar to the IMERG data. Along with our conclusions from Figure 1 and the RMSE values, it suggests that BIC values are not conclusive in selecting the best model in this case. Using a single Gamma distribution tends to produce unrealistic spatial correlations and tail values, which is mitigated by using a mixture of two Gamma distributions.

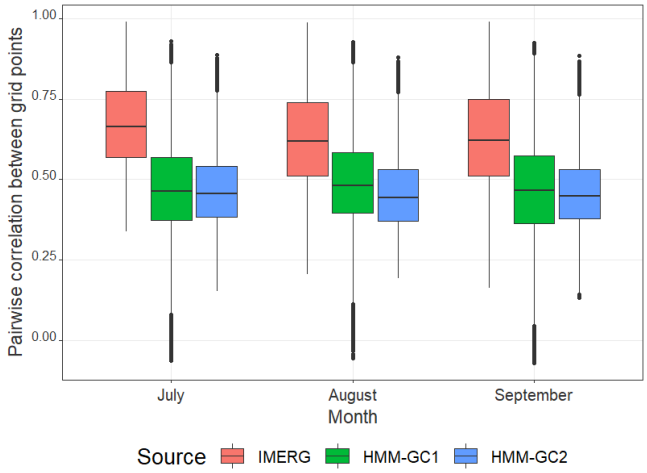

Fig. 1: Distribution of pairwise correlations between grid points based on IMERG data from 2001–2018 compared with 18 years of synthetic data from HMM-GC1 and HMM-GC2.

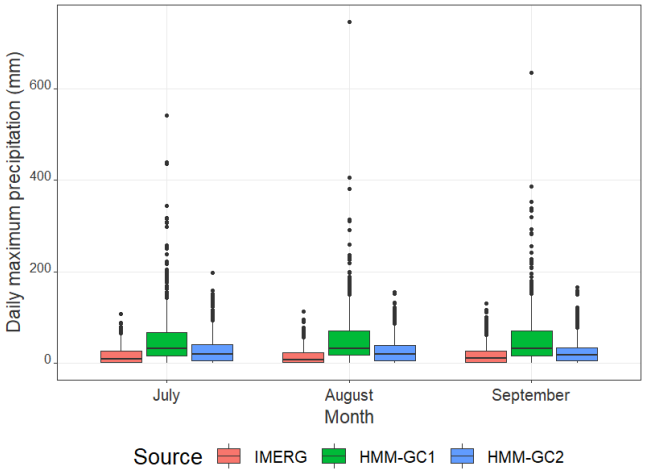

Fig. 2: Distribution of the maximum daily basin precipitation based on IMERG data from 2001–2018 compared with 18 years of synthetic data from HMM-GC1 and HMM-GC2.

Acknowledgements This work is supported in part by the U.S. National Science Foundation under the CyberTraining (OAC–1730250) and MRI (OAC–1726023) programs. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). Co-author Reetam Majumder was supported by JCET and as HPCF RA.

References

[1] G. Huffman, E. Stocker, D. Bolvin, E. Nelkin, and J. Tan, GPM IMERG Final Precipitation L3 1 day 0.1 degree × 0.1 degree V06, Edited by Andrey Savtchenko, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), 2019.
[2] J. P. Hughes and P. Guttorp, J. Appl. Meteorol. **33**, 1503–1515 (1994).
[3] A. W. Robertson, S. Kirshner, P. Smyth, S. P. Charles, and B. C. Bates, Q. J. Roy. Meteorol. Soc. **132**, 519–542 (2006).
[4] G. C. Kroiz, J. N. Basalyga, U. Uchendu, R. Majumder, C. A. Barajas, M. K. Gobbert, K. Markert, A. Mehta, and N. K. Neerchal, Stochastic Precipitation Generation for the Potomac River Basin using Hidden Markov Models, Tech. Rep. HPCF–2020–11, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2020. https://hpcf.umbc.edu.
[5] E. Bellone, J. Hughes, and P. Guttorp, Clim. Res. **15**(1), 1–12 (2000).
[6] M. Mhanna and W. Bauwens, Int. J. Climatol. **32**, 1098–1112 (2012).
[7] E. Bellone, Nonhomogeneous Hidden Markov Models for Downscaling Synoptic Atmospheric Patterns to Precipitation Amounts, Ph.D. Thesis, Department of Statistics, University of Washington, 2000.