Copula-based Correlation Structure for Multivariate Emission Distributions in Hidden Markov Models

Reetam Majumder* Amita Mehta † Nagaraj K. Neerchal‡ §

Abstract

Hidden Markov models (HMM) for multi-site daily precipitation usually assume that precipitation at each location is independently distributed conditional on the daily state; correlation in precipitation at different locations is induced by the state process. In practice, however, spatial correlations are underestimated especially when working with remote sensing data. This results in simulated data which cannot recreate the spatiotemporal patterns of the historical data. We construct a daily precipitation generator based on a hidden Markov model with Gaussian copulas (HMM-GC) using GPM-IMERG (Integrated Multi-satellitE Retrievals for Global Precipitation Measurement) remote sensing data for the Chesapeake Bay watershed on the East Coast of the USA. Daily precipitation from 2000–2019 for the wet season months of July to September is modeled using a 6-state HMM. Positive precipitation at each location is given by a two-part distribution with a delta function at zero and a mixture of two Gamma distributions; Gaussian copulas are used to accommodate the correlation in precipitation at different locations. Based on 20 years of synthetic data simulated from an HMM and an HMM-GC, we conclude that the HMM-GC captures key statistical properties of IMERG precipitation better than the HMM.

Key Words: Hidden Markov models, Stochastic simulations, Gaussian copula, Spatiotemporal analysis, Geostatistics.

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 dependent on precipitation as a primary water source. To this end, precipitation generators are constructed to produce time series of synthetic data representative of the general rainfall patterns within the region. In particular, stochastic precipitation generators aim to replicate key statistical properties of the historical data like dry and wet stretches, spatial correlations, and extreme weather events. Not only are these models used to downscale numerical weather models, synthetic data from these models find use in climate projections, impact assessments of extreme weather events, water resources and agricultural management, and for public and veterinary health [1].

Our region of interest is the Chesapeake Bay watershed which includes parts of six states and nine major river systems on the East Coast of the USA. Figure 1 shows the Chesapeake Bay watershed and the different land cover classes within it. The watershed has a diverse, interconnected ecosystem which is affected by extreme weather potentially related to climate change [8], and has been targeted for restoration as an integrated watershed and ecosystem. Rainfall within the watershed and resulting runoff into the rivers and the bay bring substantial amounts of sediments and nutrients to the bay and impact the water quality of the bay. Therefore, understanding and forecasting rainfall patterns and temporal variability, particularly extreme rainfall events in the Chesapeake Bay watershed are crucial for monitoring and managing water quality in the Bay. We use daily data from

*University of Maryland, Baltimore County, USA
†Joint Center for Earth Systems Technology, UMBC, USA
‡University of Maryland, Baltimore County, USA
§Chinmaya Vishwavidyapeeth, Kerala, India

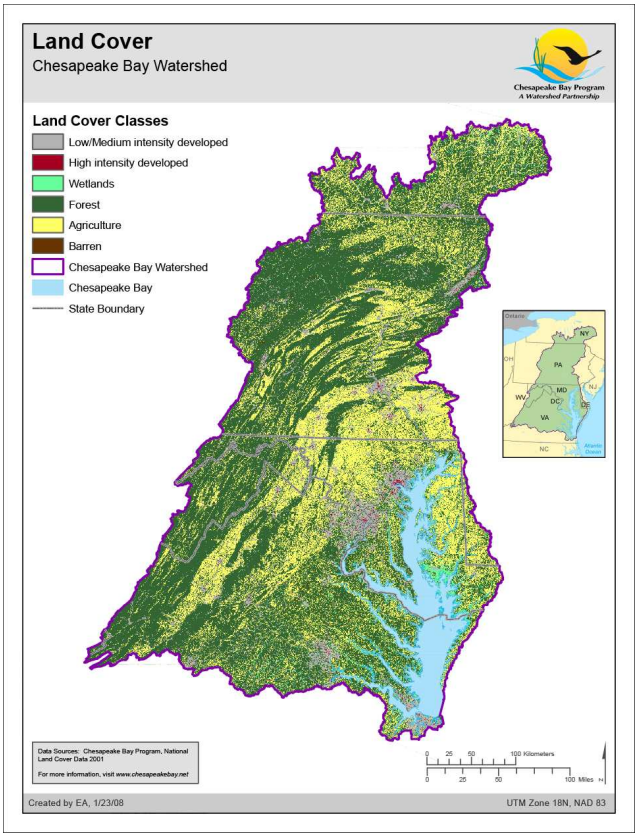

Figure 1: Land cover classes within the Chesapeake Bay watershed.

the GPM-IMERG dataset [2] for the months of July to September from 2000–2019. With a $0.1^\circ \times 0.1^\circ$ spatial resolution, The IMERG dataset covers the 64,000 square mile watershed with 1927 grid points. A hidden Markov model with Gaussian copulas (HMM-GC) is used for multi-site daily precipitation generation within the watershed.

2. Hidden Markov Model with Gaussian Copulas (HMM-GC)

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, with $\mathbf{R}_t^{'} = (R_{t1}, \ldots, R_{tM})$ representing precipitation on day t for all M locations. 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 m and day t,

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

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 induced by the Markov chain $\{S_t\}$, and precipitation at each of the M locations for every $\mathbf{R}_t$ is independent given S_t. Furthermore, daily precipitation depends only on the state on day t. Details of model formulation, estimation and simulation are described in [3, 10].

HMM parameters are usually estimated using the Baum-Welch (B-W) algorithm [9] which is a modification of the Expectation Maximization (EM) algorithm. Similarly, for the problem of estimating the most likely sequence of states given the data, the algorithm

proposed by Viterbi [11] is used, which maximizes the joint distribution of the observations and the model. We have previously found [5] that while the Baum-Welch algorithm estimates the marginal emission distribution parameters adequately, the assumption that the daily state can capture the spatial correlation of precipitation between grid points does not hold up very well for datasets with high spatial dimensions. For the purposes of simulating correlated multi-site daily precipitation, a Gaussian copula is employed to estimate the pairwise spatial correlations in our data.

For a grid of M locations, there are M(M-1)/2 pairs of grid points. Further, for each state, daily precipitation is independently and identically distributed at each location. For each state *j*, it is possible to construct an M-variate Gaussian copula $C_j(u_{1,j}, \ldots, u_{M,j})$ and generate daily correlated precipitation amounts $(r_{j1}, \ldots, r_{jM})$ using the correlation structure of C_j.

Thus, for the *j*-th state, we define the copula C_j as follows

$$C_j(u_{1,j}, \ldots, u_{M,j}) = \Phi_\Sigma(\Phi^{-1}(u_{1,j}), \ldots, \Phi^{-1}(u_{M,j})) \tag{2}$$

$$\text{and } u_{m,j} = G_{m,j}(r_{m,j}), \forall m = 1, \ldots, M \tag{3}$$

Here, $C_j(u_{1,j}, \ldots, u_{M,j})$ is an *M*-dimensional random vector with uniform marginals. Correspondingly, Φ_Σ is the cumulative distribution function (CDF) of an *M*-variate normal distribution with mean vector $\mathbf{0}$ and $M \times M$ correlation matrix Σ, Φ^{-1} is the inverse CDF of the univariate standard normal distribution, $r_{m,j}$ are the precipitation observations for state *j* at location *m*, and $G_{m,j}$ is the CDF of the mixture of the Gamma distributions modeling positive precipitation for state *j* at location *m*.

Following ideas discussed in [7], the Spearman rank correlation estimate $\hat{\rho}_j(m_1, m_2)$ from the historical data for state *j* between locations m_1 and m_2 allows us to estimate the corresponding Pearson correlation $\hat{\zeta}_j(m_1, m_2)$ using the relationship for a bivariate Normal distribution [6] given by

$$\zeta(m_1, m_2) = 2 \sin\left[\pi \frac{\rho(m_1, m_2)}{6}\right] \tag{4}$$

All marginal parameters used for the construction of the copula are estimated from the Baum-Welch algorithm, and the Viterbi algorithm provides the most likely sequence of states. The following procedure is used to construct a Gaussian copula for each state.

Algorithm 1: Algorithm to construct a Gaussian copula for each state.

for *states j in 1:J* **do**
- Subset the days corresponding to state *j*;
- Calculate the M(M-1)/2 estimates of $\rho_j(m_1, m_2)$ from Eqn. 4 ;
- Calculate pairwise correlations for the copula, $\zeta_j(m_1, m_2)$;
- Plug $\zeta_j(m_1, m_2)$ into the correlation matrix Σ;
- Set diagonal elements of the correlation matrix to be 1;
- Ensure that the resulting matrix is positive definite;

end

Note that the correlation matrix of the copula is also its covariance matrix. Positive definiteness is ensured by diagonalizing the matrix and replacing all negative eigenvalues with a small positive number, and recalculating the matrix [7].

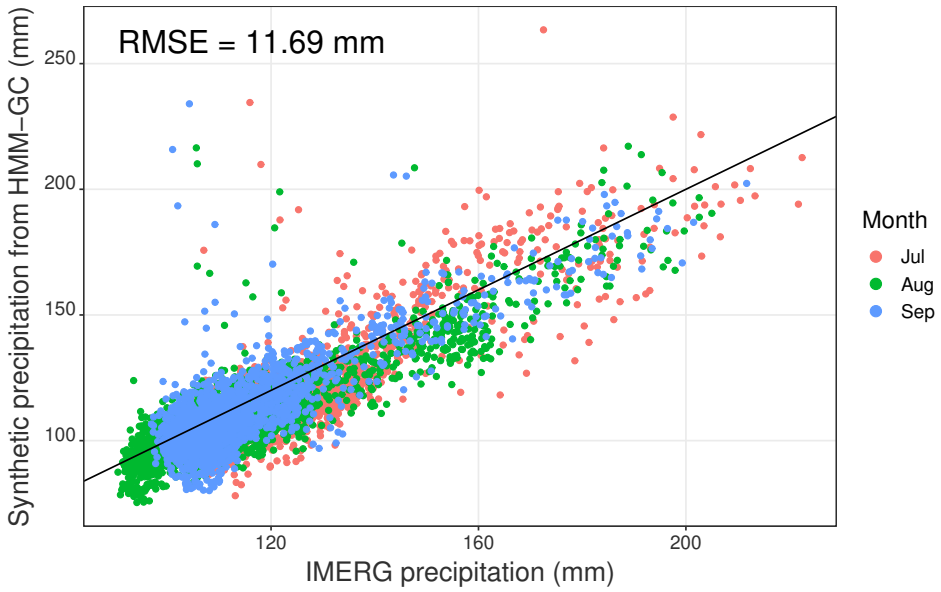

Figure 2: Scatterplot of the mean precipitation per month at each grid point based on historical IMERG data (2000–2019) compared with synthetic HMM-GC data.

3. HMM-GC for Daily Precipitation over the Chesapeake Bay Watershed

3.1 Model selection

Bayesian Information Criterion (BIC) scores have been used for model selection. The BIC is a model selection criterion calculated from the likelihood function; the model with the lowest BIC tends to be chosen. For each combination of parameters, we restarted the Baum-Welch algorithm 10 times, and the solution with the lowest BIC was used. We choose a 6-state HMM with 2 Gamma components (C=2) for our analysis which had a BIC score of 1.195e+07. Beyond this the improvements in BIC were marginal, and the model often failed to converge. Models with exponential distributions had higher BIC scores than their Gamma counterparts. Going forward, a Gaussian copula is constructed for each of the 6 states, and for each copula the marginal CDF comes from a mixture of 2 Gamma distributions.

3.2 Model fit and interpretation

Figure 2 compares monthly means computed from 20 years of synthetic HMM-GC data against IMERG data from 2000–2019. It plots the monthly precipitation at each grid point, with July–September means represented by red, green, and blue dots respectively. We see a linear relationship between the means computed from the synthetic HMM-GC data and those from IMERG, with positive bias at a few locations. The root mean square error (RMSE) of the synthetic data estimates is 11.69 mm. Similarly, the RMSE for the proportion of dry days is 2% for our data. Spatially, pairwise correlations between grid points for the IMERG data have a median of 0.395 and a maximum of 0.991, whereas the HMM-GC synthetic data has a median spatial correlation of 0.332 and a maximum of 0.925.

Table 1 lists the daily and seasonal precipitation statistics corresponding to the 6 states over all grid points for the Chesapeake Bay watershed calculated based on our 2000–2019 IMERG data. State 1 is the driest and occurs most often (23.26%), even during the wet season. State 3 is the wettest state with the highest mean and maximum precipitation and

Table 1: Precipitation statistics for the 6 HMM states over all 1927 locations within the Chesapeake Bay watershed.

State	Daily mean precipitation (mm)	Daily maximum precipitation (mm)	% of all days without rain	% of all days spent in state
1	0.12	2.24	18.32	23.26
2	5.08	16.61	1.24	16.63
3	13.84	51.97	0.22	13.97
4	4.47	13.78	1.50	13.64
5	0.91	4.91	6.30	15.76
6	1.97	13.26	6.82	16.74

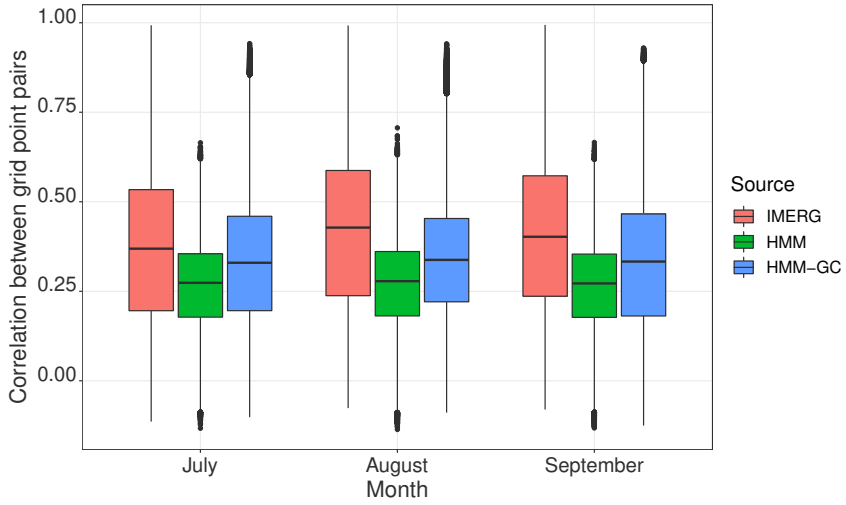

Figure 3: Pairwise spatial correlation between grid points for historical IMERG data (2000–2019) compared with synthetic data from HMM and HMM-GC models.

the smallest proportion of dry days (0.22%). State 6 is also a dry state in general but occasionally has heavy precipitation events. States 2, 4, and 5 have precipitation patterns which range between the other 3 states.

3.3 Comparison with classical HMM

Figure 3 shows box plots of the pairwise correlation between grid points based on IMERG data from 2000–2019, as well as synthetic data from HMM and HMM-GC for 20 years. The low median and interquartile ranges of HMM and HMM-GC compared to IMERG suggest that both models struggle with capturing spatial correlation to different degrees. We see that the classical HMM for precipitation tends to severely underestimate the correlations between precipitation amounts. The HMM-GC does a significantly better job of estimating the spatial correlations with its longer upper tail.

Figures 4 and 5 plot the daily total precipitation amounts over the basin from July to September of 2018. The cyan line in both plots represents daily IMERG precipitation from 2018 over the entire watershed. These are compared to synthetic precipitation represented by the red lines corresponding to the HMM in Figure 4 and the HMM-GC in Figure 5. The IMERG data in both figures contains low precipitation events, as well as high precipitation events of nearly 50000 mm. When comparing the HMM and HMM-GC data, we see that the HMM-GC simulates high precipitation events much better than the HMM, as

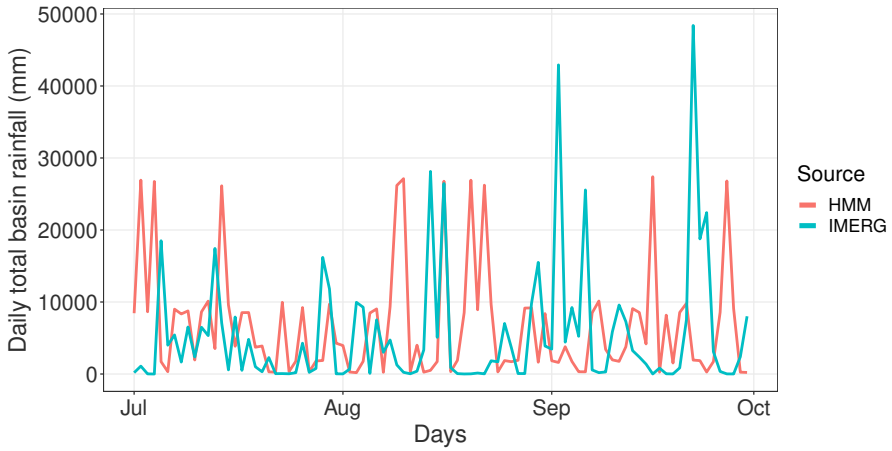

Figure 4: Time series of total daily rainfall over the basin in July to September 2018, compared against a single realization from the HMM.

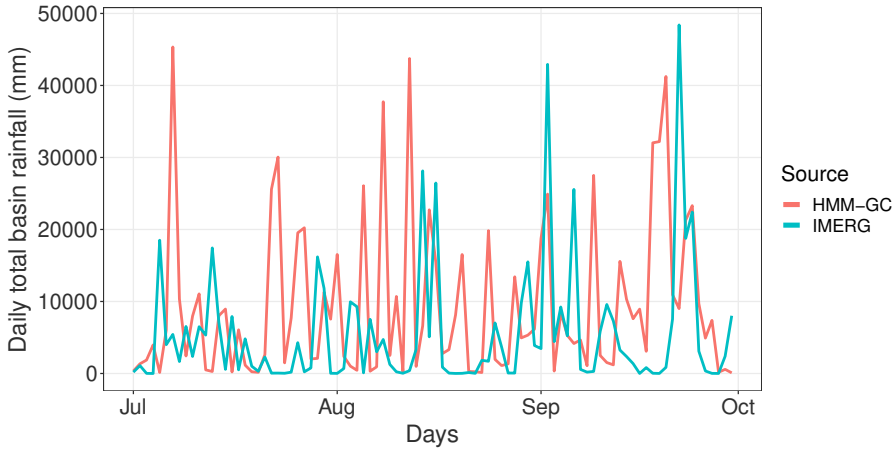

Figure 5: Time series of total daily rainfall over the basin in July to September 2018, compared against a single realization from the HMM-GC.

evidenced by the peaks in both graphs. The failure to simulate extreme precipitation events in the classical HMM formulation may be attributed to the underestimation of spatial correlations. Both methods perform well in simulating low precipitation events. In general, the HMM-GC outperforms the HMM both spatially and temporally with some additional computational cost for estimation and simulation.

4. Discussion

Based on the synthetic data, both the HMM and the HMM-GC have similar monthly precipitation statistics at individual locations; this is because the marginal parameters are the same for both models. However, when we start looking at precipitation over the entire watershed, the HMM shows unrealistic spatial patterns. Due to the underestimated spatial correlations, simulated data from the HMM can have high precipitation at specific locations but low precipitation values at nearby grid points. So while both models can simulate low precipitation events well, the HMM fails to simulate high precipitation events simultaneously across large areas of the watershed adequately. The HMM-GC alleviates this to a certain extent, as evidenced by the higher peaks in the simulated daily data.

However, simulated data from the HMM-GC still underestimates spatial correlations compared to historical data. One of the reasons could be that the Gamma CDFs that make up our copula are estimated from the data as well. There also arises another issue where several pairs of locations do not have any days with rainfall at both locations. This prevents us from calculating correlation based on just positive precipitation. We have tried to address this by using all available data to estimate the copula, not just the days with positive rainfall. However, while generating simulated data, rainfall occurrence is determined using the marginal estimates derived from the Baum-Welch algorithm, while the intensity is determined by the values generated from the copula. While this allows us to have a larger sample size for estimating the correlations and produces higher correlation estimates compared to using just the positive precipitation data, we are looking at ways to simulate both occurrence and intensity of daily rainfall from the copula.

Acknowledgements

The hardware in the UMBC High Performance Computing Facility (HPCF) is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS–0821258, CNS–1228778, and OAC–1726023) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See `hpcf.umbc.edu` for more information on HPCF and the projects using its resources. HMM estimation has been carried out using the MVNHMM toolbox (`http://www.sergeykirshner.com/software/mvnhmm`) developed by Sergey Kirshner [4]; ancillary scripts developed by Jonathan N. Basalyga and Gerson C. Kroiz [5] have also been used. Reetam Majumder was supported by JCET and as HPCF RA.

References

[1] K. Breinl, G. Di Baldassarre, M. Girons Lopez, M. Hagenlocher, G. Vico, and A. Rutgersson. Can weather generation capture precipitation patterns across different climates, spatial scales and under data scarcity? *Sci. Rep.-UK*, 7, 2017.

[2] G. J. Huffman, E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and J. Tan. GPM IMERG final precipitation L3 1 day 0.1 degree $\times$ 0.1 degree V06, 2019. Edited by Andrey Savtchenko, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), `https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGDF_06/summary`, accessed on Aug 28, 2020.

[3] J. P. Hughes and P. Guttorp. Incorporating spatial dependence and atmospheric data in a model of precipitation. *J. Appl. Meteorol.*, 33:1503–1515, 1994.

[4] S. Kirshner. Modeling of multivariate time series using hidden Markov models. Ph.D. Thesis, University of California, Irvine, 2005.

[5] 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. Technical Report HPCF–2020–11, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2020.

[6] W. H. Kruskal. Ordinal measures of association. *J. Am. Stat. Assoc.*, 53(284):814–861, 1958.

[7] M. Mhanna and W. Bauwens. A stochastic space-time model for the generation of daily rainfall in the Gaza Strip. *Int. J. Climatol.*, 32:1098–1112, 2012.

[8] C. B. Program. Climate change, 2012. Accessed September 16, 2020, from `https://www.chesapeakebay.net/issues/climate_change`.

[9] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[10] A. W. Robertson, S. Kirshner, P. Smyth, S. P. Charles, and B. C. Bates. Subseasonal-to-interdecadal variability of the Australian monsoon over North Queensland. *Q. J. Roy. Meteor. Soc.*, 132:519–542, 2006.

[11] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE T. Inform. Theory*, 13(2):260–269, 1967.