A Comparison of Stochastic Precipitation Generation Models for the Potomac River Basin

Author: Gerson C. Kroiz, Mentor: Matthias K. Gobbert

Department of Mathematics and Statistics, UMBC

Senior Thesis, Spring 2020

Abstract

Weather ensembles are an integral part of weather forecasting and can also be used to test the sensitivity and performance of climate models. Among meteorological variables, simultaneous simulation of precipitation at multiple sites presents unique challenges since precipitation has a semi-continuous distribution. We compare Robertson's hidden Markov model setup with Wilks' Multivariate Markov Chain based generator to see how well they recreate the spatiotemporal characteristic of gridded satellite precipitation estimates. Our results show that the Wilks method does a better job of capturing spatial correlations, while the HMM model can estimate and simulate longer durations of time.

1 Introduction

The modeling and forecasting of precipitation play a significant role in determining water allocation and resource management for regions that depend on precipitation as a primary water source. It is important to model both its seasonal and inter-annual variations, and one method of achieving this is through precipitation generators. Using a dynamical or statistical model created from historical precipitation data, these generators can produce time series of multivariate synthetic data representative of the general rainfall patterns of the region and allow us to study its statistical properties. The synthetic data aims to replicate pairwise spatial correlations and extreme weather events over long periods of time that might show up in the historical data. These models are used at different spatial and temporal resolutions, and our interest lies in modeling multi-site daily precipitation. For the precipitation generation, this work considers daily precipitation over the Potomac river basin. 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. The data source used for this study was the GPM-IMERG L3 1 Day VO6 [1] dataset for the months July to September from 2001–2018, resulting in 1656 days worth of data. With a $0.1^\circ \times 0.1^\circ$ resolution, the GPM-IMERG data set divides the basin into a 387 point grid space. Using the daily GPM-IMERG data, we aim to simulate synthetic rainfall data over each location of the 387 point grid.

The first precipitation generator for satellite precipitation estimates of the Potomac River Basin that we used is a hidden Markov model (HMM). HMMs are a flexible class of models which have been widely used for simulating daily rainfall based on observed weather station data by structuring the model to have precipitation amounts dependent on a finite number of hidden weather states [2, 6]. The state process is assumed to be a Markov chain, and conditional on the state, the emission process is modeled at each location using a mixture of a delta function at zero, and one or more Exponential or Gamma distributions.

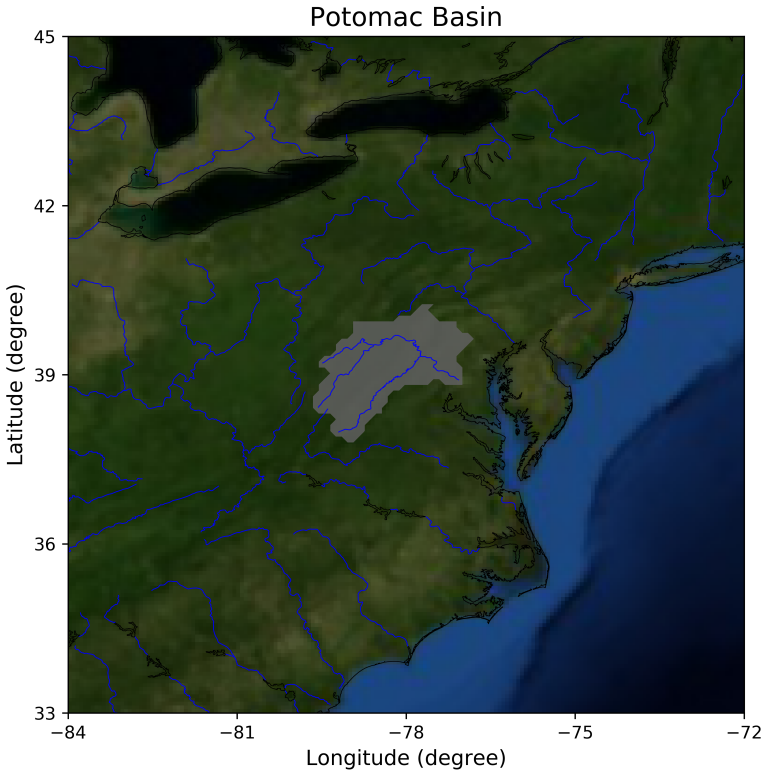


Figure 1: Extent of the Potomac river basin on the East Coast of the USA indicated by the gray shape; rivers are represented by blue lines. The Potomac originates at two separate sources in West Virginia and Virginia, and drains into the Chesapeake Bay which connects to the Atlantic ocean on the Eastern side of the map.

The other method explored in this study is another widely used model called the Wilks approach [7]. Wilks modeled precipitation occurrences and precipitation amounts separately. The occurrences are modeled at each location as a first-order, two-state Markov process for dry and wet days, and amounts are modeled using a Gamma distribution. The Markov chain process at each location captures the temporal correlation, and the use of Gaussian copulas allows for explicit specification of spatial correlation for both the amounts and occurrences.

This work is organized as follows. Section 2 describes how both models were implemented into code. Using these models, Section 3 compares the performance of the synthetic data from the two methods against historical data. Section 4 draws conclusions about the models based on the results in the previous section.

2 Generating Synthetic Data

2.1 Hardware and Software

The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (`hpcf.umbc.edu`). The study used CPU nodes with two 18-core Intel Xeon Gold 6140 Skylake CPUs (2.3 GHz clock speed, 24.75 MB L3 cache, 6 memory channels)

and 384 GB memory. The nodes are connected by a network of four 36-port EDR (Enhanced Data Rate) InfiniBand switches with 100 Gb/s bandwidth and 90 ns latency.

The precipitation data has been preprocessed using Python scripts developed by Kel Markert for the NASA SERVIR training on hydrologic modeling using VIC. The code is available at `https://github.com/KMarkert/servir-vic-training` and was run using `Python 2.7.x` on `taki`. The software used for the majority of the hidden Markov model computations is the MVNHMM toolbox developed by Sergey Kirshner [3] and Padhraic Smyth and available at `http://www.sergeykirshner.com/software/mvnhmm`. The toolbox was developed for Linux and was installed and used on `taki`. Ancillary scripts written in `Python 3.6.x` and `R 3.6.x` were used extensively in the analysis.

2.2 HMM Code Development

Within the MVNHMM toolbox, the main functions used in this study were the Baum-Welch algorithm for estimation, the Viterbi algorithm for the generation of most likely estimated states, and the simulation algorithm. Calling any of these actions requires a specialized parameter file. We wrote scripts with `Python 3.6.4` for automating the generation of parameter files and running the models with relative ease using batch jobs, and also used the `mpi4py` library to parallelize the tasks of grid-search for optimum parameters and model estimation and simulation for different locations in parallel. Parallelization allows for the creation and execution of multiple parameter files simultaneously, leading to overall time reductions when running a study consisting of parameter variations. The number of desired parameter files can be requested via the batch scripts used to run the code.

2.3 Wilks Code Development

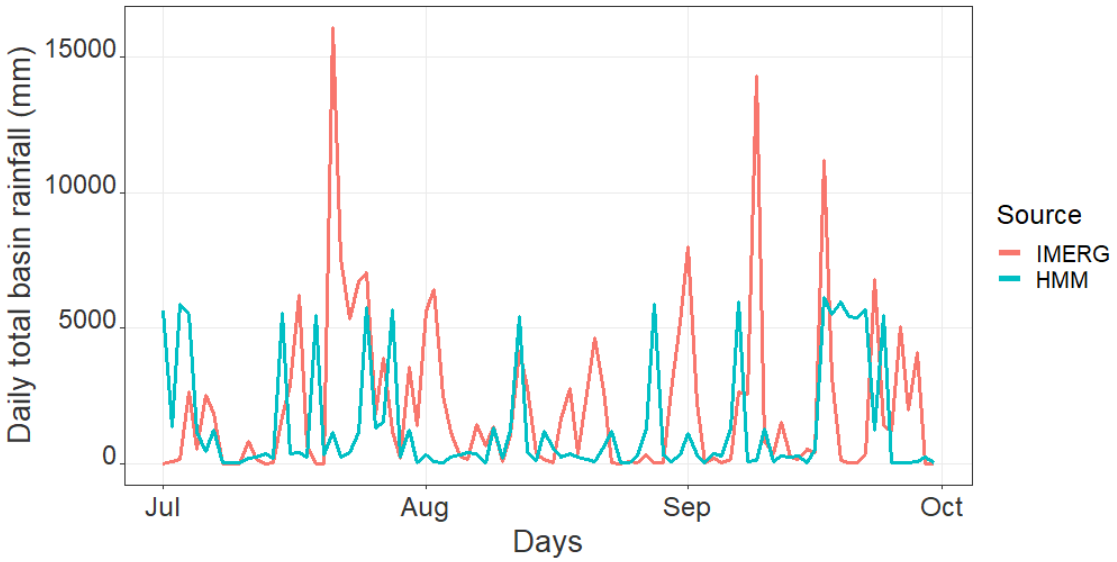
The code for the Wilks Method is written entirely in `R` and is a transcription of the steps described in Mhanna and Bauwens [5]. While the model is meant to run on each month separately and has scope for parallelization, we ran it on the entire three months for a more straightforward comparison with the HMM. The input is a text file with the IMERG data, from which the parameters of both the occurrence and amounts are estimated. Ancillary functions are defined and sourced from a secondary file. The code runs entirely in serial and generates a text file with simulated data for as many years as we require.

2.4 Statistical Analysis

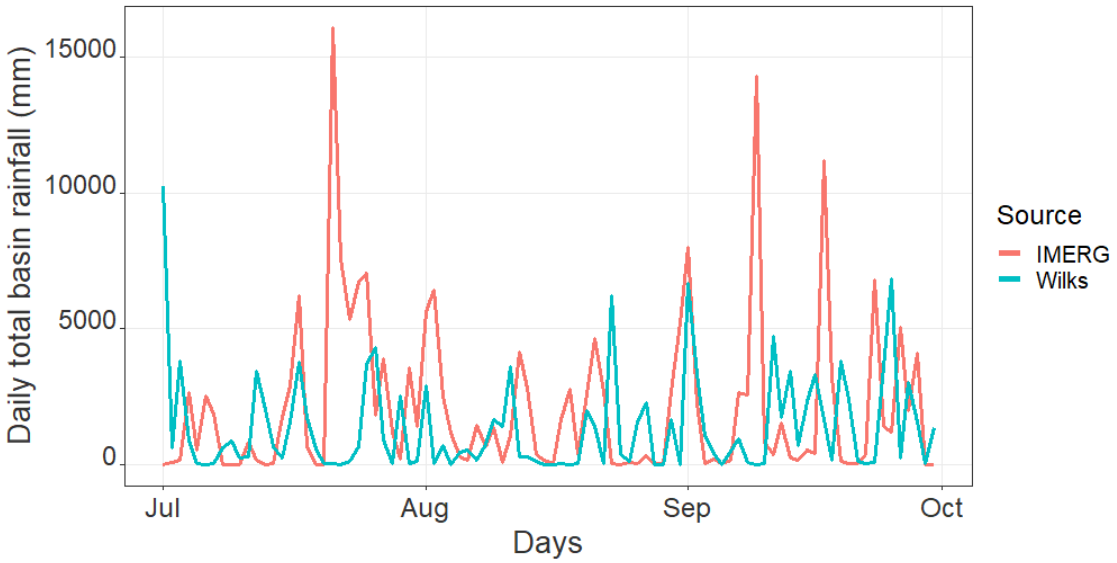
The bulk of the statistical analysis was carried out in `R 3.6.3`. The MVNHMM toolbox was still used to run the Baum-Welch and Viterbi algorithms since it is the only available software that can fit a mixture of emission distribution. For the HMM, a script was written in `Python` to extract information from the generated parameter files, which was then imported into `R` for the remaining part of the study. This was necessary since the toolbox does not have a mechanism to simulate correlated emissions or states. All plots for the two models were generated using the `ggplot2` library in `R`.

3 Results from the HMM and Wilks Model

3.1 Temporal Distribution of the Synthetic Data



(a)



(b)

Figure 2: Daily total precipitation over the basin for 2018 based on IMERG (red) and simulated data using (a) HMM (green) and (b) Wilks method (green).

Figure 2 displays the daily total precipitation amounts from July to October of 2018. In both plots, the red line shows the same IMERG data. These are compared to the green lines, from HMM in Figure 2 (a) and from the Wilks method in Figure 2 (b). From Figure 2 (a), we see that the HMM data is representative of the IMERG data, suggesting that the model does an adequate job of simulating average precipitation events. This conclusion is

determined by the similarities in the troughs and peaks of the graph. From Figure 2 (b), the Wilks model is also capable of simulating average precipitation events. When comparing the two figures, there are subtle differences. In Figure 2 (b), the Wilks data has a peak slightly above 10,000 mm. The HMM data in Figure 2 (a) does not display precipitation values past 6,000 mm. Based on this observation, the Wilks model does a better job of capturing extreme events, where the HMM data does not simulate extreme events well. However, the IMERG 2018 data displays extreme events over 15,000 mm, well above the maximum amounts of daily precipitation simulated from either model. These figures suggest that both methods are capable of capturing temporal correlation over three month periods but lose some degree of accuracy for simulating extreme precipitation events.

3.2 Extreme Precipitation Events of the Synthetic Data

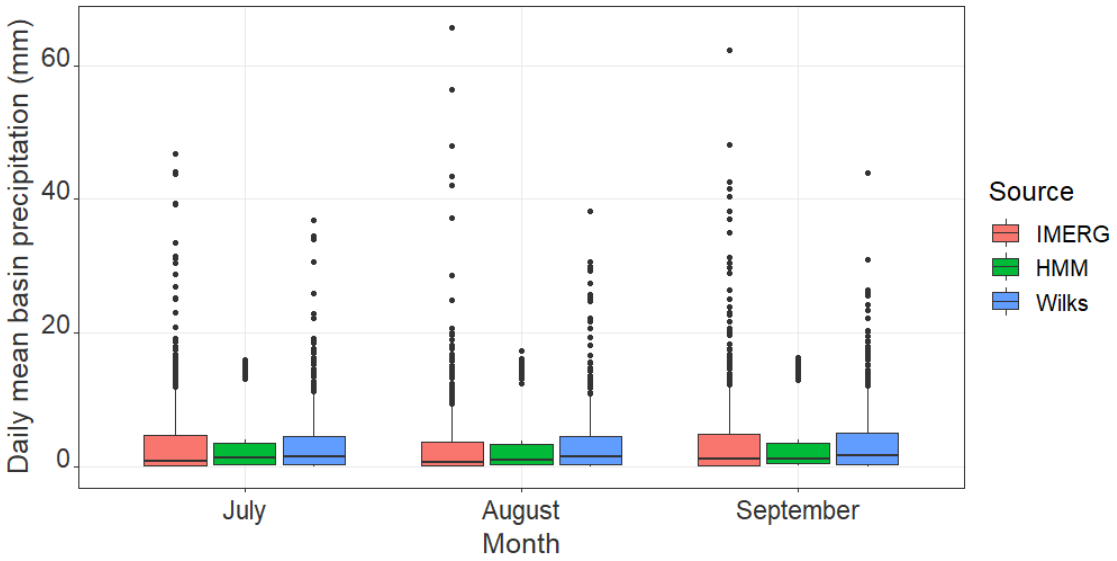


Figure 3: Distribution of daily **average** basin precipitation from IMERG data (red) from 2001–2018 and synthetic data from the HMM (green) and Wilks method (blue)

Figure 3 shows box plots of the daily average amounts of precipitation for the HMM, Wilks method, and the IMERG data. A box plot is a visual representation of the distribution of a quantitative variable in terms of a five point summary - the three horizontal lines represent the first quartile (lower boundary of the box), the median (middle of the box) and the third quartile (upper boundary of the box), whereas the lower and upper ends of the vertical lines (known as the whiskers) represent the minimum and maximum values. Points beyond the whiskers on either side are considered outliers. The similarities in the median values and interquartile range (values between first and third quartile) for each of the models and the IMERG data support the evidence that both models are capable of modeling average precipitation events. Similar to the previous plots, the short tails for the HMM data in Figure 3 suggests that the HMM struggles with capturing extreme precipitation events when compared to the synthetic data. The longer tails of the Wilks data shows that the Wilks

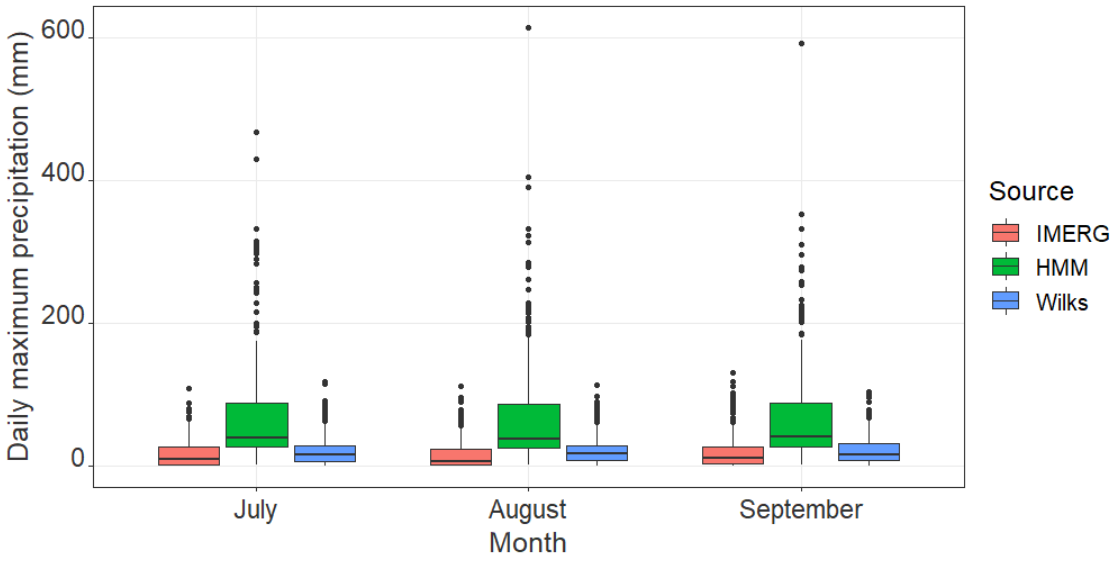


Figure 4: Distribution of daily **maximum** basin precipitation from IMERG data (red) from 2001–2018 and synthetic data from the HMM (green) and Wilks method (blue)

model has higher accuracy with simulating extreme precipitation amounts. While the HMM struggles with capturing extreme events over the entire basin, the high median value and long tail of extreme cases in Figure 4 shows that the HMM data at individual locations can replicate extreme values both within the range and well above the precipitation amounts of the IMERG data.

3.3 Spatial Correlation of the Synthetic Data

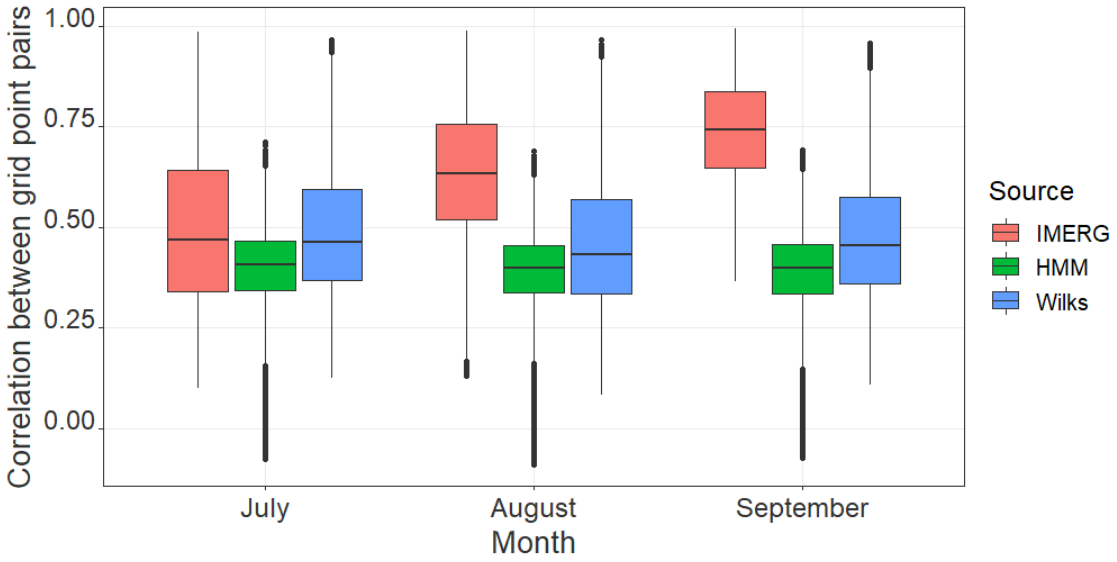


Figure 5: Pairwise spatial correlation between grid points for historical IMERG (red) compared with synthetic data from HMM (green) and Wilks model (blue) based on 18 years of data

Figure 5 displays the spatial correlation for locations within the data of the HMM, Wilks method, and IMERG. From Figure 5, the lower median and interquartile ranges of the two models suggest that they both struggle with capturing spatial correlation. As the spatial correlation of the IMERG data in the figure appears to increase for the later months, both models fail to capture this correlation. However, especially in July, the box plot of the Wilks data more accurately represents the spatial correlation of the IMERG data compared to the results from the HMM.

The generic HMM configuration used in this thesis fails to simulation spatial correlation between two grid points. However, further adjustments and the addition of Gaussian copulas to generate spatially correlated precipitation amounts presented in [4] significantly improved the spatial correlation. The report further shows that addition of Gaussian copulas improves the model's ability to simulate precipitation over long periods of time and extreme events.

4 Conclusions

In this thesis, we applied a hidden Markov model and the Wilks model to remote sensing data from GPM-IMERG over the Potomac river basin and examined how the synthetic data produced by the model compares to historical data described in Section 1. For spatiotemporal models of multi-site daily precipitation like the HMM and the Wilks model used in this work, there are three major data features we want to capture and replicate in synthetic data: spatial correlation between locations in the region, rainfall amounts and occurrences for long periods of time, and the extreme weather events of the region. In Section 3, we discussed model performance for the HMM and Wilks Model based on these metrics.

Temporal correlation shown in Figure 2 (a) and (b) displays that both models are capable of simulating average precipitation events. Figure 3 shows that Wilks model does a much better job of capturing extreme precipitation amounts for the entire basin. For spatial correlation, Figure 5 shows that both models can struggle with replicating the observed data, but the Wilks model is slightly more accurate. However, [4] shows further adjustments to the HMM that improve the model's simulation of spatial correlation.

The results gathered in this work suggest that the two models capture spatial correlations between locations in the region, the rainfall amounts and occurrences for long periods of time, and the extreme weather events of the region at different levels of accuracy. For capturing precipitation events for long periods of time, the Wilks method models each month separately as accuracy of estimation may decrease for longer periods of time. The HMM is capable of capturing precipitation events for longer durations. For simulating extreme precipitation events, the Wilks model displayed higher levels of accuracy compared to the HMM. However, as shown in Figure 4, the HMM is capable of capturing extreme precipitation amounts at individual locations. The Wilks method does a slightly better job of capturing spatial correlation, but both models could be improved to further the accuracy of this correlation.

Acknowledgments

Special thanks to Reetam Majumder, the research assistant for this work, for providing background information and aiding with any issues throughout the project. I would also like to acknowledge Dr. Gobbert for advising the research project and guiding the progress of the work. I also acknowledge funding via the grant CyberTraining: DSE: Cross-Training of Researchers in Computing, Applied Mathematics and Atmospheric Sciences using Advanced Cyberinfrastructure Resources from the National Science Foundation (grant no. OAC–1730250). I was also supported through an Undergraduate Research Award (URA) from UMBC. 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.

References

[1] G. J. Huffman, E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and Jackson Tan. GPM IMERG final precipitation L3 1 day 0.1 degree × 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 June 25, 2020.

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

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

[4] Gerson C. Kroiz, Jonathan N. Basalyga, Uchendu Uchendu, Reetam Majumder, Carlos A. Barajas, Matthias K. Gobbert, Kel Markert, Amita Mehta, and Nagaraj K. Neerchal. Stochastic precipitation generation for the potomac river basinusing hidden markov models. Technical Report HPCF–2020–11, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2020.

[5] 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.

[6] 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.

[7] D. S. Wilks. Multisite generalization of a daily stochastic precipitation generation model. *J. Hydrol.*, 210(1–4):178–191, 1998.