Stochastic Precipitation Generation for the Potomac River Basin Using Hidden Markov Models

Gerson C. Kroiz¹, Jonathan Basalyga¹, Uchendu Uchendu²
RAs: Reetam Majumder¹, Carlos Barajas¹, Mentor: Matthias K. Gobbert¹
Collaborators: Kel Markert³, Amita Mehta⁴, Nagaraj K. Neerchal¹,⁵

¹Department of Mathematics and Statistics, UMBC
²Department of Information Systems, UMBC
³The University of Alabama in Huntsville / NASA-SERVIR
⁴Joint Center for Earth Systems Technology, UMBC
⁵Chinmaya Vishwavidyapeeth, Kerala, India

Cybertraining: Big Data + High Performance Computing + Atmospheric Sciences
University of Maryland, Baltimore County, Spring 2020
cybertraining.umbc.edu
Acknowledgments: NSF, UMBC, HPCF, CIRC, NASA-SEVRIR, JCET, Math & Stat, URA

05/15/2020
Outline

1 Background and Motivation

2 Methodology and Data
   - The Hidden Markov Model for Precipitation
   - Dataset and Model Parameters
   - Use of UMBC’s High-Performance Computing Facility

3 Results

4 Conclusions
The Potomac river basin is located on the central region of the East Coast and provides water to much of that region. Rainfall is one of the main sources of water for the basin. Due to seasonal and inter-annual variations, modeling rainfall adequately is crucial when planning water allocation within the basin.

Figure: Extent of the Potomac river basin indicated by the gray shape; rivers are represented by blue lines.
Generating ensembles for meteorological variables is common in climate studies and in the earth sciences, since physical based models are sensitive to initial conditions.

Most satellite data come with their share of measurement errors and errors stemming from data assimilation.

Spatio-temporal modeling and synthetic generation of precipitation faces additional challenges since precipitation has a semi-continuous distribution, with a point mass at zero and a continuous distribution on \((0, \infty)\).
Motivation
Synthetic Precipitation Generation under a Hidden Markov Model formulation

- **Precipitation Generators**: given existing data, they generate a synthetic time series of multi-site precipitation at a daily scale for long periods. Sites are often monitoring stations or points on a spatial grid.

- We focus on the **Hidden Markov Models (HMM)** approach for daily precipitation generation and discuss its performance when working with Satellite Precipitation Estimates (SPEs).

- The three features of the daily data we want to replicate are **long sequences, pairwise spatial correlations, and extreme events**.

- Usually, no model formulation can capture all features equally well.
Outline

1. Background and Motivation

2. Methodology and Data
   - The Hidden Markov Model for Precipitation
   - Dataset and Model Parameters
   - Use of UMBC’s High-Performance Computing Facility

3. Results

4. Conclusions
The Hidden Markov Model approach

The model follows Robertson et al. (2006)\(^2\), which in turn is based on the work of Hughes and Guttorp (1994)\(^3\)

- Precipitation is assumed to depend on a finite set of hidden (unobserved) weather states
- The hidden state model is first-order Markov, which captures temporal correlation
- Spatial correlation is also implicitly captured by the shared state
- Observed daily precipitation at each location is a "noisy" version of the hidden shared weather state
- Conditional on the daily state, precipitation amounts at each location are modeled as independent and identical observations from a mixture of exponential distributions


Performance on weather station data

- There are no studies on HMM performance for gridded SPEs, or comparing it with the other widely used precipitation generator which is based on the Wilks method\(^4\)
- We focus on modeling precipitation between July-September
- Our study uses daily IMERG V06\(^5\) data from 2001-2018 over the 387 grid points of the Potomac basin
- After using grid search to find optimum parameters, a 4-state model using a mixture of 2 Gamma distributions was chosen based on BIC scores

---


\(^5\)https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGDF_06/summary
Use of Taki Cluster

- Scripts for executing various elements of the HMM used mpi4py for parallelization in Python 3.6.4.
- The python scripts for running the HMM used multiple nodes on the HPCF 2018 CPU cluster.
- The bulk of the statistical analysis and data generation based on the Gaussian copula was carried out in R 3.6.3 using the markovchain package.
- Plots were produced via the ggplot2 package in R.
Outline

1. Background and Motivation

2. Methodology and Data
   - The Hidden Markov Model for Precipitation
   - Dataset and Model Parameters
   - Use of UMBC’s High-Performance Computing Facility

3. Results

4. Conclusions
Spatial correlation visualization

- Visualization comparing how both HMM and HMM-GC capture spatial correlation of the IMERG data
- HMM-GC displays significant improvement over HMM.
Comparision between HMM and HMM-GC
Maximum daily precipitation

- HMM overestimates maximum precipitation across the basin
- HMM-GC captures daily maximums much better than HMM
Both HMM and HMM-GC have similar distributions for mean daily precipitation.

The key difference is the tail values corresponding to high precipitation at all locations, where the HMM-GC does a much better job.
Scatterplot of the proportion of dry days per month

Figure: Scatterplot of the mean proportion of dry days per month at each grid point based on historical IMERG data (2001-2018) compared with means computed over 100 years of synthetic HMM-GC data

- Each point of synthetic represents an average taken over 100 years
- July is overestimated while August is underestimated.
- Low RMSE indicates that the model captures the number of monthly precipitation occurrences at each location
Figure: Scatterplot of the mean precipitation per month at each grid point based on historical IMERG data (2001-2018) compared with means computed over 100 years of synthetic HMM-GC data

- Each point representative of a 100 year mean worth of synthetic data at location
- August overestimated and July underestimated.
- Low RMSE indicates that the precipitation amounts are generally modeled by the HMM-GC
Conclusions

HMM
- Captures general precipitation events over long periods of time
- Fails to capture spatial correlation between locations adequately
- Can replicate extreme precipitation events at individual locations, but not spatially consistent

HMM-GC
- Captures general precipitation events over long periods of time
- Significantly improved spatial correlation in synthetic data
- Spatially consistent replication of heavy precipitation events

HMM-GC improves the HMM’s ability to capture long sequences, pairwise spatial correlations, and extreme events.
- The replication of spatial correlation can be further improved.
- The variation in the scatter plots signifies that there is information in the IMERG data that the HMM-GC fails to capture.