Final Technical Report

DOE Award SC0006619 on Stochastic Simulation of Precipitation Data, Sept 2011-August 2015
PI: Padhraic Smyth, University of California, Irvine

a) Methodological development

We developed a framework for a new type of hidden Markov model (HMM), including the statistical and mathematical foundations of the model as well as implementing in software in the R software package and making it publicly available. In this new modeling framework, an HMM is coupled with a generalized linear model (GLM) within a Bayesian framework to downscale exogenous input variables to spatio-temporal daily rainfall. This model (which we refer to as a GLM-HMM) is an extension of the non-homogeneous hidden Markov model (NHMM) to include uncertainty quantification in a Bayesian framework. The GLM-HMM also allows the exogenous inputs to directly modulate the emission distribution (governing the occurrence and amounts at each station) in a more flexible manner than a traditional NHMM, allowing for modeling of non-stationarity (for both seasonal and inter-annual variation).

The GLM-HMM includes hidden weather states for each day which have Markov transition dynamics that capture the time dependence and spatial relationship of the rainfall much like a traditional NHMM. Each day is modeled by one of the K hidden weather states, where each state has a distinct spatial pattern across stations in terms of distribution on precipitation occurrence and amount. The Markov state transitions are modeled through a set of transition probability matrices that vary over time, allowing the capture of seasonal variation in the relative frequencies of different states.

The distributions of daily rainfall depend on both the daily state variable as well as the exogenous input variables (which can for example be a large-scale variable reflecting inter-annual variability). The precipitation distribution for each station is a mixture of a delta function at zero and two gamma distributions (one for lower rainfall amounts and one for higher amounts). The exogenous variables influence which of the three component distributions is chosen given a particular daily weather state. As well as large scale annual exogenous input variables, seasonal or daily inputs can also be included in the model. The Bayesian framework allows for uncertainty estimation of the dependence of daily rainfall on particular exogenous input variables.

b) Algorithm and Software Development

In our work we have found that Bayesian implementations of NHMM models are computationally expensive and can be difficult to tune. By incorporating the exogenous variables into a different portion of the model (via the new GLM-HMM approach described above) we can retain the general form and functionality of the NHMM while removing the need for extensive tuning of algorithm parameters. Latent variables are added to ensure the mathematical conjugate properties are available to use Gibbs steps throughout the Markov chain Monte Carlo (MCMC) algorithm. The model is implemented in R with the aid of the Rcpp package to construct compilable modules in C++ for faster run time. This reduced the run-time 100-fold due to the compiled nature of the Rcpp modules. N=2000 iterations can be run in 4 hours for a field of 23 stations and 31 monsoon seasons (June, July, Aug, Sept) without the need for specialized user knowledge of MCMC tuning methods. In addition, the R code has been parallelized and, using the snowfall package, is run on a multi-core Linux machine to reduce computational time.

In more recent work we have developed a further extension of our Bayesian inference algorithms that allow us to fit these types of models to much larger data set, e.g., to almost 700,000 daily measurements of precipitation over 30 years across India. This new extension is based on a technique known as Polya-Gamma data augmentation, which has been proposed and applied successfully recently to multivariate data – our work extends this approach to handle temporal data using our GLM-HMM approach.

We implemented or developed metrics to assess the fit and predictive performance of the GLM-HMM model. Given particular settings of the exogenous variables, we collect N=500 simulated daily downscale runs from the model for each station. Assessment of the spatial features of the model is done through comparison of pair-wise station correlations of the simulated time series. The skill of the model to capture the temporal aspects of the data is assessed by run lengths of wet and dry spells of these 500 simulated runs. The main difference in the GLM-HMM and a traditional NHMM is in how the exogenous variables handle station level seasonality and downscale at a daily level; seasonal plots with uncertainty bands were compared to the raw data. The NHMM can have difficulty because it extends one seasonal input to the entire field of stations, whereas the GLM-HMM can have a seasonal trend for each station. The GLM-HMM allows the emission distribution of a state to change throughout the season, instead of being stationary. These trends were apparent in our model assessment plots. Additionally, larger model fit assessment was done through the computation of the recursive log-likelihood and annual metrics to measure the influence of different exogenous variables on both training and test sets of data. 6-fold cross validation is used to split the data into six portions for training and test (holding out 5 year portions of data at a time).

The resulting software has been publicly released as an R software package, NHMM: Bayesian NHMM Modeling, https://cran.r-project.org/web/packages/NHMM/index.html

c) Application of NHMMs to downscaling of rainfall projections in India and China

Downscaling rainfall using the GLM-HMM has been applied to a region in India covering the Upper Indus basin that feeds into the reservoir system, and in China for the Upper Yangtze basin. The impact of exogenous variables such as the El Nino-Southern Oscillation (ENSO), standardized anomaly index (SAI), and wind shear (WSI) have been assessed by this model for their skill in out-of-sample prediction of downscaling the monsoon season for these regions. Thirty years of daily rainfall data is available for training and testing the model for these regions. Test data is held out to assess the ability of the model to forecast or hindcast given the exogenous variables.

d) Publications:

T. Hoslclaw, A. Greene, A. R. Robertson, P. Smyth, A Bayesian hidden Markov model of daily precipitation over South and East Asia, Journal of Hydrometereorology, doi:10.1175/JHM-D-14-0142.1, 17(1):3--25, 2016.

A. Greene, T. Holsclaw, A. Robertson, P. Smyth, A Bayesian multivariate nonhomogeneous Markov model, in Machine Learning and Data Mining Approaches to Climate Science, Springer, pp.61--69, 2015.

P. Arnesen, T. Holsclaw, P. Smyth, `Bayesian detection of changepoints in finite-state Markov chains for multiple sequences, in press, 2016.

T. Holsclaw, A. W. Robertson, A. M. Greene, and P. Smyth, Bayesian non-homogeneous Markov models via Polya-Gamma data augmentation with applications to precipitation modeling, in preparation, 2016.