The Annals of Applied Statistics
2017, Vol. 11, No. 1, 393–426
DOI: 10.1214/16-AOAS1009

BAYESIAN NONHOMOGENEOUS MARKOV MODELS VIA PÓLYA-GAMMA DATA AUGMENTATION WITH APPLICATIONS TO RAINFALL MODELING[1]

BY TRACY HOLSCLAW*, ARTHUR M. GREENE†, ANDREW W. ROBERTSON† AND PADHRAIC SMYTH*

*University of California, Irvine** *and Columbia University*†

Discrete-time hidden Markov models are a broadly useful class of latent-variable models with applications in areas such as speech recognition, bioinformatics, and climate data analysis. It is common in practice to introduce temporal nonhomogeneity into such models by making the transition probabilities dependent on time-varying exogenous input variables via a multinomial logistic parametrization. We extend such models to introduce additional nonhomogeneity into the emission distribution using a generalized linear model (GLM), with data augmentation for sampling-based inference. However, the presence of the logistic function in the state transition model significantly complicates parameter inference for the overall model, particularly in a Bayesian context. To address this, we extend the recently-proposed Pólya-Gamma data augmentation approach to handle nonhomogeneous hidden Markov models (NHMMs), allowing the development of an efficient Markov chain Monte Carlo (MCMC) sampling scheme. We apply our model and inference scheme to 30 years of daily rainfall in India, leading to a number of insights into rainfall-related phenomena in the region. Our proposed approach allows for fully Bayesian analysis of relatively complex NHMMs on a scale that was not possible with previous methods. Software implementing the methods described in the paper is available via the R package NHMM.

1. Introduction. Consider the problem of modeling the dynamics of a multivariate discrete time series $\boldsymbol{y}_t$, with component measurements y_{ts}, $s = 1, \ldots, S$, and a discrete-time index $t = 1, \ldots, T$. A useful modeling approach in this context is the hidden Markov model (HMM), where the observed $\boldsymbol{y}_t$'s are assumed to be a stochastic function of a (hidden) finite-state Markov process $\boldsymbol{z}$, with components $z_t \in \{1, \ldots, K\}$, and where each vector $\boldsymbol{y}_t$ is assumed to be conditionally independent of all other $\boldsymbol{y}_{t'}$ vectors and state variables $z_{t'}$, $t' \neq t$, given state z_t [Zucchini, MacDonald and Langrock (2016)]. The conditional distribution of the $\boldsymbol{y}_t$ vectors at time t given the state z_t is often assumed to be time homogeneous, defined by so-

Received June 2016; revised December 2016.

[1]Supported in part by the U.S. Department of Energy, Office of Science, Grant DE-SC0006616 and by NASA Grant NNX15AQ06A.

Key words and phrases. Nonhomogenous hidden Markov model, multivariate time series, Pólya-Gamma latent variables, probit and logit link.

called emission distributions,[2] $f(\boldsymbol{y}_t|z_t = k, \boldsymbol{\theta}), k \in \{1, \ldots, K\}$, where $\boldsymbol{\theta}$ represents the emission distribution parameters. The distributional choice for f will depend on the particular characteristics of the $\boldsymbol{y}_t$ measurements for a given application.

HMMs are appealing for problems where the dynamics of $\boldsymbol{y}_t$ are too complex to be directly modeled (e.g., for high-dimensional problems where S is large), but can instead be approximated via a discrete-state hidden Markov process $\boldsymbol{z}$. For example, a common assumption in practice [and one that is used in this paper—see also Zucchini, MacDonald and Langrock (2016, page 140), and the discussion of contemporaneous conditional independence] is to assume that the components of $\boldsymbol{y}_t$ are conditionally independent given the state, that is, that $f(\boldsymbol{y}_t|z_t = k, \boldsymbol{\theta}) = \prod_{s=1}^{S} f_s(y_{ts}|z_t = k, \boldsymbol{\theta})$, where $f_s(\cdot)$ denotes the conditional distribution of component s of the observed vector $\boldsymbol{y}_t$. HMMs can also be used to produce a time-dependent clustering of the observations $\boldsymbol{y}_t$, where the state variables z_t are interpreted as indicators of cluster memberships, with the Markov dependence providing temporal dependence (in contrast to mixture model clustering, for example, where the cluster memberships are modeled as being independent). Using HMMs for clustering in this manner can be useful in econometric, ecological, or other scientific time-series applications [e.g., MacDonald and Zucchini (1997), Raphael (1999), Siepel and Haussler (2004), Mamon and Elliott (2007), Patterson et al. (2016)]. The goal is often to try to gain insight into possible latent processes that might be giving rise to the observed $\boldsymbol{y}_t$ data, for example, by analyzing and interpreting differences among the emission distributions $f(\boldsymbol{y}_t|z_t = k, \boldsymbol{\theta})$ across states.

The time homogeneity of the standard HMM (at the parameter level, as described above) can be limiting in practice, for example, if $\boldsymbol{y}_t$ has seasonal dependence or is nonstationary. One approach to relaxing this assumption is to allow the $K \times K$ transition matrix probabilities to be dependent on an exogenous time-series $\boldsymbol{x}_t$, resulting in a nonhomogeneous hidden Markov model (NHMM) [e.g., Hughes and Guttorp (1994), Diebold and Lee (1994), Hughes, Guttorp and Charles (1999), Kirshner, Smyth and Robertson (2004), Kim, Piger and Startz (2008), Paroli and Spezia (2008), Meligkotsidou and Dellaportas (2011), Rajagopalan, Lall and Tarboton (1996)]. A natural parametrization is to model each of the K rows of the transition matrix via a multinomial logistic function, with K possible outcomes (the K possible states at time $t + 1$ given the current state z_t). Temporal inhomogeneity can also be introduced in the emission component of the model, for example, by allowing the parameters of the emission distributions $f(\boldsymbol{y}_t|z_t = k, \boldsymbol{\theta})$ to vary with time t and location s as a function of another exogenous process $\boldsymbol{w}_{ts}$ [e.g., Holsclaw et al. (2016)].

[2]Here we use the term "emission distributions," widely used in speech recognition and language modeling [e.g., Jurafsky and Martin (2014)]—these are also referred to as "state-dependent distributions" [e.g., Zucchini, MacDonald and Langrock (2016)].

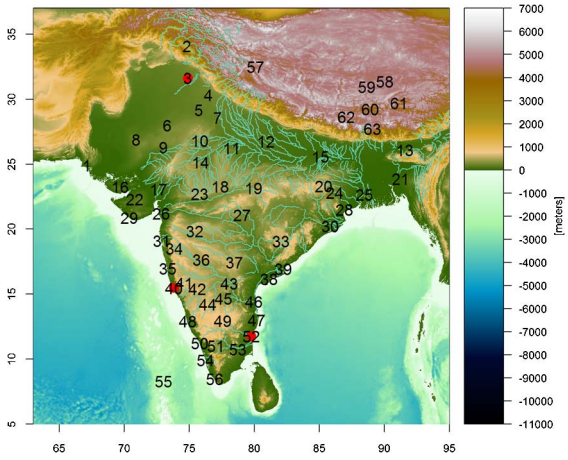


FIG. 1. *Locations of the* 63 *rain gauge stations, showing the topography of South Asia. Stations* 3 (31.63°*N*, 74.87°*E*), 40 (15.48°*N*, 73.82°*E*), *and* 52 (11.77°*N*, 79.77°*E*) *are each marked with a dot; these diverse locations will be used in subsequent plots as examples.*

As a motivating example we consider the problem of modeling and simulating daily station rainfall data over India where the observations y_{ts} correspond to the amount of rain that has fallen on day t at weather station s. The data we analyze has been collected daily for 30 years at 63 rain gauge stations across India, totaling well over half a million observations (6.9×10^5). The geographical area of interest contains diverse subregions where the rainfall varies greatly in seasonal timing and amount. As shown in Figure 1, some stations lie in the Himalayas, while others are located variously in coastal, monsoon, or desert regions; these data are not isotropic in nature. Figure 2 shows the rainy days in lighter shades, indicating amounts and dry days in dark shades for three contrasting stations [see Holsclaw et al. (2017) for additional stations].

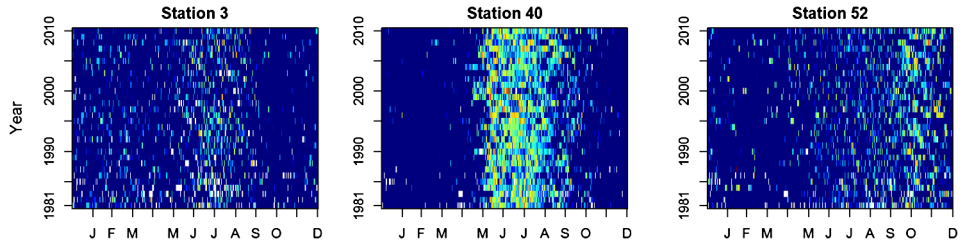


FIG. 2. *Daily rainfall data* (*log of the amount in mm*) *with the x-axis being the day of the year and the y-axis depicting* 30 *years. The left panel shows the relatively dry, Station* 3 *located in NW India, the middle panel shows Station* 40 *on the west coast, strongly impacted by the summer monsoon, and the right panel shows Station* 52 *on the SW coast that is influenced by the winter monsoon, peaking in October–December. Darker colors indicate lower daily log rainfall amounts, and lighter colors indicate higher daily values; white is for missing observations.*

Accurately modeling and simulating rainfall on a daily timescale is important across a number of diverse applications, such as crop modeling, flood risk assessment, and water policy decisions [Hansen et al. (2006), Challinor et al. (2009), Piani et al. (2010)]. Multivariate HMMs have been successfully applied to this modeling problem in the past, where the hidden variables z_t can be interpreted as weather states exhibiting persistence at daily timescales, and the emission distributions $f(\boldsymbol{y}_t|z_t = k, \boldsymbol{\theta})$ capture the spatial and distributional characteristics of observed rainfall for each state [Zucchini and Guttorp (1991), Hughes and Guttorp (1994), Kirshner (2010), Greene, Robertson and Kirshner (2008), Zucchini, MacDonald and Langrock (2016)]. Of direct interest to climatologists is the situation where the rainfall in a given region is being influenced or driven by time-varying atmospheric variables $\boldsymbol{x}_t$ such as pressure differentials at large spatial scales. Relating these large-scale variables to local rainfall characteristics at particular station locations s is known as *downscaling*. NHMMs have been found to be broadly useful in this context where the $\boldsymbol{x}_t$ variables act as "drivers" for the Markov transition matrix as described earlier [Hughes, Guttorp and Charles (1999), Bellone, Hughes and Guttorp (2000), Charles et al. (2004), Robertson (2009), Germain (2010), Carey-Smith, Sansom and Thomson (2014), Heaps, Boys and Farrow (2015)]. Other work in a downscaling context, such as that of Berrocal, Gelfand and Holland (2010) and Fuentes and Raftery (2005) for ozone and airborne particulates, focuses on the use of Gaussian models—these models are not appropriate here given the nonnegativity and nonnormality of precipitation data.

There are a multitude of other modeling approaches that could be used in this context. In particular, dynamic spatio-temporal models provide a rich framework for modeling spatial and temporal dependencies. These models often use continuous latent-space representation (in contrast to the discrete state representation of the HMM approach) and are often parametrized in a manner that can incorporate relevant scientific knowledge, for example, in the form of differential equations [see Hooten and Wikle (2010) for a review]. Such approaches can provide richer representations for spatial structure that go beyond the conditional independence assumption that has often been used when NHMMs are applied to precipitation modeling (and that we use here in this paper). In Section 5.4 and in Holsclaw et al. (2017) we examine the model's ability to capture spatial dependence across stations and conclude that while the conditional independence approach tends to underestimate the true spatial dependence, that the model nonetheless is capturing much of the dependence that is empirically observed. For applications where spatial dependence is of critical importance, additional spatial dependence could be incorporated in the emission component of our proposed model at the cost of additional complexity and computational effort.

Many of the early applications of HMMs and NHMMs, to climate data as well as to other problems, have relied on point estimates of model parameters, often using the Expectation-Maximization (EM) algorithm for parameter estimation

[Dempster, Laird and Rubin (1977)]. There is, however, a growing need for efficient Bayesian methods for assessing uncertainty in these types of models [e.g., Rydén (2008)]. For example, in the context of climate data, modeling the uncertainty in rainfall amounts is important in both seasonal forecasting and climate change downscaling applications [Maraun et al. (2010), Vermeulen et al. (2013)], and Bayesian simulations are better suited to characterizing such uncertainty than point-estimate approaches.

While there has been extensive development of Bayesian methods for HMMs [Scott (2002), Frühwirth-Schnatter (2006), Rydén (2008), Patterson et al. (2016)], there has been little work on Bayesian estimation of NHMMs. Prior work has typically focused on analysis of small univariate data sets due to the complexity and computational expense of the Metropolis–Hastings MCMC schemes used for inference [e.g., Filardo and Gordon (1998), Spezia et al. (2014)]. Meligkotsidou and Dellaportas (2011) apply the Bayesian multinomial logit regression (MNL) latent variable technique developed by Holmes and Held (2006a, 2006b) to the NHMM, illustrating the approach using a relatively small univariate financial econometrics data set with monthly observations over 38 years. For many applications, however, we need methods that scale up efficiently to much larger data sets. The rainfall data set we analyze later in the paper consists of a 63-dimensional time series with $T \approx 30 \times 365 = 10{,}950$ observations per time series.

The development of an efficient Bayesian sampling scheme to handle logistic transition matrices in NHMMs is a problem that has proven challenging in the past because of the lack of conjugacy that arises due to the logistic functional form. With scalability in mind, we adopt the Pólya-Gamma latent variable method previously used for sampling in a multinomial logistic (MNL) regression framework [Polson, Scott and Windle (2013)] and extend it to the NHMM in this paper. We are motivated by the results in Polson, Scott and Windle (2013) which showed that the Pólya-Gamma latent variable method is significantly faster than alternative sampling schemes such as those of Holmes and Held (2006a) and Frühwirth-Schnatter and Frühwirth (2007). Furthermore, because the Pólya-Gamma method uses only Gibbs sampling steps, this obviates the need for extensive parameter-tuning of the sampling algorithm, leading to a significantly simpler implementation in software compared to methods based on Metropolis–Hastings steps, for example. We have implemented the algorithm proposed in this paper and made this model available in the NHMM R package on the Comprehensive R Archive Network (CRAN).

The contributions of our paper are as follows. We propose a novel hidden Markov model with inhomogeneity in both the transition and emission state-dependent distributions. This model generalizes earlier NHMMs that contained either transition or emission inhomogeneity but not both. An additional significant contribution of the paper is the development of a fully Bayesian estimation scheme for this class of models. In particular, we develop a scalable Bayesian sampling scheme for the logistic transition component of the NHMM, enabling these methods to be applied to much larger data sets than in prior work. Finally, we

demonstrate the application of the model and the Bayesian inference algorithms to a large-scale multi-decadal precipitation data set.

Section 2 lays out the proposed Bayesian multivariate NHMM. A description of the Bayesian implementation of the MCMC algorithm and the handling of missing data, predictive simulations, and forecasting are discussed in Section 3; further modeling considerations such as variable selection and model choice for the NHMM are included in Appendix A. Local rainfall amounts for 63 stations in and around India and the exogenous variables to be downscaled are described in Section 4; specific details pertaining to the exogenous variables can be found in Appendix B. Section 4 also provides a brief summary of rainfall modeling. Section 5 includes the analysis and results of the NHMM when applied to the Indian rainfall data. Our findings and general conclusions are summarized in Section 6.

2. Bayesian multivariate nonhomogeneous Markov model.

2.1. *The NHMM and the likelihood.* The observed multivariate time-series $y_{ts}, s = 1, \ldots, S$, with discrete-time index $t = 1, \ldots, T$, is modeled using an NHMM. A general way to introduce exogenous dependence into the transition matrix is to allow each transition to have its own set of logistic coefficients (or at least $K - 1$ of them, subject to identifiability), implying $O(K^2 B)$ coefficients in total for B exogenous variables. This was the approach taken in Meligkotsidou and Dellaportas (2011) for $K = 2$. However, this model requires a large number of parameters as K grows. A more parsimonious approach (and the one we follow in this paper) is to have one set of regression coefficients for each state, with $O(KB)$ coefficients in total, allowing the probability of entering each state j to be modulated by a set of K weighted regressors (via the logistic link), but where the modulation is independent of the previous state i [e.g., Kirshner, Smyth and Robertson (2004)]. The intuitive interpretation is that the exogenous variables control how likely the Markov chain is to enter each state j (via the logistic link and regression coefficients), and thus, as the exogenous variables change over time, so do the probabilities of being in each state. As a simple example, if one of the exogenous variables reflects seasonality (time of year), this allows the model to visit hidden states in a seasonal fashion. The hidden process $\boldsymbol{z}$ is Markov with an inhomogeneous $K \times K$ transition matrix $\boldsymbol{Q}_t$ with components q_{ijt}, where $i, j \in \{1, \ldots, K\}$. The transition probability entries at time t are modeled via a multinomial logistic link function:

$$q_{ijt} = P(z_t = j | z_{t-1} = i, \boldsymbol{x}_t, \boldsymbol{\zeta}) = \frac{\exp(\xi_{ij} + \boldsymbol{x}_t' \boldsymbol{\rho}_j)}{\sum_{m=1}^{K} \exp(\xi_{im} + \boldsymbol{x}_t' \boldsymbol{\rho}_m)}, \tag{1}$$

where $\boldsymbol{x}_t$ is a B-dimensional exogenous covariate time series, $t = 1, \ldots, T$, and $\boldsymbol{\rho}_j$ is a B-dimensional vector of coefficients corresponding to the B components of $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{Bt})$. For notational convenience let $\boldsymbol{\zeta} = \zeta_{ij} = (\boldsymbol{\rho}_j, \xi_{ij})$ for all $i, j \in \{1, \ldots, K\}$. We assign one of the $\boldsymbol{\zeta}_{\cdot j}$ to zero for some value of j (one $\boldsymbol{\rho}_j$ and

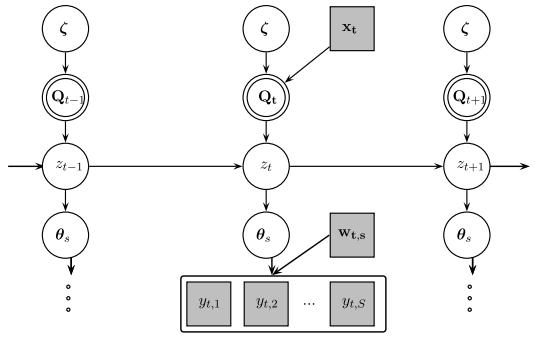


FIG. 3. *Graphical model*: *The observed values* $(y_{ts}, \boldsymbol{w}_{ts}, \boldsymbol{x}_t)$ *are in gray boxes. The unknown parameters* ($\boldsymbol{\theta}$ *for the emission distribution and* $\boldsymbol{\zeta}$ *for the transition probabilities*) *and hidden states* (z_t) *are circles.* $\boldsymbol{Q}_t$ (*in double circles because they are directly calculated in contrast to sampled parameters*) *is a set of matrices that contain the transition probabilities arising from the Markov property of the hidden states and the exogenous variables* $\boldsymbol{x}_t$.

a vector of $\boldsymbol{\xi}_{\cdot j}$ for some j) for identifiability. The choice of the logistic function above is discussed further in Section 3.1.

The other main component of an NHMM is the set of state-dependent emission distributions $f(\boldsymbol{y}_t|z_t = k, \boldsymbol{\theta}), k = 1, \ldots, K$, and where $\boldsymbol{\theta}$ is the set of all parameters of the emission distribution. Each combination of state (k) and station (s) has its own emission distribution (for this particular application the emission distribution will be a zero-inflated mixture of exponential distributions). In general, these distributions can be specified to be inhomogeneous over time by allowing the parameters to depend on time-varying exogenous variables $\boldsymbol{w}_{ts}$, yielding $f(\boldsymbol{y}_t|z_t, \boldsymbol{\theta})$.

Figure 3 shows a graphical representation of the multivariate NHMM and how the two types of exogenous variables $(\boldsymbol{x}_t, \boldsymbol{w}_{ts})$ impact the model. If the values of the latent variables $\boldsymbol{z}$ are assumed to be known, the conditional likelihood for the model above can be expressed as

$$P(\boldsymbol{y}_t|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\zeta}, \boldsymbol{\theta}) = \prod_{t=1}^{T} f(\boldsymbol{y}_t|z_t, \boldsymbol{w}, \boldsymbol{\theta}) P(z_t|z_{t-1}, \boldsymbol{x}_t, \boldsymbol{\zeta}), \tag{2}$$

where $P(z_t|z_{t-1}, \boldsymbol{x}_t, \boldsymbol{\zeta})$ for $t = 1$ is defined via an initial state distribution $P(z_1)$ and where the $P(z_t|\ldots)$ transition probabilities are defined as in equation (1). When the latent variables are unknown, the likelihood $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\zeta}, \boldsymbol{\theta})$ can be computed by marginalizing over the unknown $\boldsymbol{z}$ values in the usual recursive manner for HMMs [e.g., see Scott (2002)]. Priors and inference procedures for the unknown parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\theta}$ are described in the next section.

3. Bayesian inference and MCMC algorithm. We describe below how to perform inference in a Bayesian framework for the model in the preceding section using a Markov chain Monte Carlo (MCMC) algorithm. Posterior full conditional

distributions can be computed for each of z, $\boldsymbol{\zeta}$, and $\boldsymbol{\theta}$ independently, such that each step of the MCMC algorithm focuses on only one set of parameters at a time. Our primary emphasis below is on the development of a sampling method for the transition matrix parameters $\boldsymbol{\zeta}$ since this has traditionally presented difficulties in the context of Bayesian analysis of NHMMs and has effectively limited the sizes of data sets that can be analyzed in past studies.

If the posterior full conditional distributions are known in closed form, then the parameters can be sampled by Gibbs steps within the MCMC. For problems where the posterior distributions are not conjugate, it is sometimes possible to have auxiliary variable methods facilitate rendering full posterior conditional distributions in a form that can be sampled from. For the NHMM described above, two sets of latent variables can be added to the model: one set for sampling the coefficients $\boldsymbol{\zeta}$ associating with the transition probabilities of the hidden states [Polson, Scott and Windle (2013)] and another set associated with parameters $\boldsymbol{\theta}$ of the emission distributions [Albert and Chib (1993)]. Using auxiliary variables (and the resulting Gibbs sampling algorithm) in this manner can be more efficient compared to alternative approaches such as Metropolis–Hastings, for example, leading in some cases to better mixing (and thus less thinning and fewer iterations) as well as having the advantage of not requiring tuning parameters for the sampler (i.e., which results in a user friendly R Package).

3.1. *Sampling the $\boldsymbol{\zeta}_k$ coefficients.* In this NHMM, there are $K-1$ coefficients ($\boldsymbol{\zeta}$) associated with each of the B observed daily variables $\boldsymbol{x}_t$. These coefficient parameters ($\boldsymbol{\zeta}$) are related to the transition probabilities associated with the hidden states through a link function. There are two standard link functions that are typically used in this context: the logistic multinomial (MNL) and the multinomial probit (MNP) [Riihimaki, Jylanki and Vehtari (2013), Neal (1997)], both of which are commonly used in regression modeling of polychotomous response variables. Although there has been relatively little literature on Bayesian inference with MNP or MNL link functions for NHMMs, Bayesian implementations in the context of regression modeling are well studied [e.g., see Albert and Chib (1993), Aitchison and Bennett (1970), Chib and Greenburg (1998), Imai and van Dyk (2005), McCulloch, Polson and Rossi (2000), Johndrow, Lum and Dunson (2013), Zhang, Boscardin and Belin (2008) for MNP regression and Holmes and Held (2006a, 2006b) Scott (2011), Polson, Scott and Windle (2013), Frühwirth-Schnatter (1994), O'Brien and Dunson (2004) for MNL regression]. Although mathematically quite similar [Paap and Frances (2000)], MNL and MNP require quite different Bayesian sampling algorithms. The Bayesian implementation of the MNP regression model usually samples the coefficients using latent variables; this is quite efficient and therefore works well for large data sets [Albert and Chib (1993)]. However, unlike the regression case, the NHMM has the additional need to calculate the transition probabilities for which there is no analytic solution in the case of the MNP. For this reason the MNL has tended to be the link function

of choice for NHMM modeling. But the Bayesian implementation of MNL tends to be slow, requiring multiple tuning parameters and long sampling runs. For this reason, it is often only used with relatively small data sets and small numbers of coefficients [Scott (2011), Frühwirth-Schnatter (1994), O'Brien and Dunson (2004)]. For example, Meligkotsidou and Dellaportas (2011) construct a Bayesian NHMM inference procedure by drawing from the MNL regression method of Holmes and Held (2006a, 2006b) using a relatively complex slice sampler to analyze a small univariate time series.

Polson, Scott and Windle (2013) has recently introduced a new MNL method using Pólya-Gamma latent variables, providing an algorithm that is more efficient (both in terms of time per run and needing no tuning parameters), which opens up the possibility of handling much larger data sets with these models. This provides the motivation to apply the Polson, Scott and Windle (2013) Pólya-Gamma MNL latent variable method to the NHMM. There are a number of aspects of the MNL regression method that are altered in the extension to the NHMM case [refer to Section 5 of Polson, Scott and Windle (2013) for details for sampling $\boldsymbol{\zeta}$ of the MNL regression]. In the NHMM, there is no observed multinomial data as in the MNL regression. Instead, the sampled hidden states z_t are set up in matrix form to conform to the MNL regression method. $\boldsymbol{Z}$ is a T by K matrix with entries Z_{tk}, where the columns contain the binary representation of the hidden states (a 1 in the column of the z_t and 0 elsewhere) and are updated during each of the iterations of the MCMC sampler. The exogenous variables and the Markov dependence (from z_{t-1} for $t = 2, \ldots, T$) are included in the matrix $\boldsymbol{X}$ which has dimension T by $K + B$. The first K columns encode the information of the Markov property (z_{t-1}) in a binary form followed by B columns for the exogenous variables ($x_{b,t}$ for all b and t). $\boldsymbol{\zeta}$ is a K by $K + B$ matrix of coefficients, indexed by where $k = 1, \ldots, K$ and $h = 1, \ldots, K + B$. One of the rows of $\boldsymbol{\zeta}$ is set to zero for identifiability (the first K by K entries are the $\boldsymbol{\rho}$'s and the next B columns are the $\boldsymbol{\xi}$'s). The full conditional posterior distribution for the ζ_{kh}'s allows them to be drawn conditioned on the current draw of hidden states and other variables. The likelihood for ζ_{kh} is given by

$$
\begin{aligned}
l(\zeta_{k,h}|\zeta_{-k,h}) &= \prod_{t=1}^{T} \left(\frac{e^{\eta_{tkh}}}{1+e^{\eta_{tkh}}}\right)^{Z_{tk}} \left(\frac{e^{\eta_{tkh}}}{1+e^{\eta_{tkh}}}\right)^{1-Z_{tk}} \\
&= \prod_{t=1}^{T} e^{(Z_{tk}-1/2)\eta_{tkh}} e^{-\eta_{tkh}^2/2} \omega_{tkh}\, \mathrm{PG}(\omega_{tkh}|1,0),
\end{aligned} \tag{3}
$$

where $\eta_{tkh} = X_{th}\zeta_{kh} - C_{tkh}$ with $C_{tkh} = \log \sum_{i \neq k} \exp X_{th}\zeta_{ih}$ (which is needed for the multinomial logistic form). $\boldsymbol{\omega}$ is a set of latent variables with components ω_{tkh}. At each time step there is only one observation of the hidden state, and so in terms of the MNL regression the observation count is one. The full conditional posteriors are given by

$$
\zeta_{kh}|\boldsymbol{\Omega}_{kh} \sim N(m_{kh}, V_{kh}) \quad \text{and} \quad \omega_{tkh}|\zeta_{kh} \sim \mathrm{PG}(1, \eta_{tkh}),
$$

where scalars $V_{kh} = (\boldsymbol{X}_h'\boldsymbol{\Omega}_{kh}\boldsymbol{X}_h + b_{kh}^{-1})^{-1}$ and $m_{kh} = V_{kh}(\boldsymbol{X}_h'((\boldsymbol{Z}_k - 1/2) - \boldsymbol{\Omega}_{kh}\boldsymbol{C}_{kh}) + b_{kh}^{-1}a_{kh})$. $\boldsymbol{\Omega}_{kh}$ is a T by T diagonal matrix containing $\boldsymbol{\omega}_{kh}$ along the diagonal. a_{kh} and b_{kh} are parameters of the conjugate prior; the implementation in our R package allows for a conjugate prior of the form $\boldsymbol{\zeta}_{kh} \sim N(a_{kh}, b_{kh})$. If a noninformative prior is desirable, then we can let a_{kh} and b_{kh}^{-1} be zero as we do in our rainfall example later in the paper.

The transition matrix is a necessary part of the NHMM not typically used in MNL/MNP regression. Once the coefficients ($\boldsymbol{\zeta}$) are sampled, then the transition probabilities can be easily obtained through the logistic relationship given in equation (1). This leads to a K by K transition matrix for time t:

$$\boldsymbol{Q}_t = \begin{bmatrix} q_{11t} & q_{12t} & \cdots & q_{1Kt} \\ q_{21t} & q_{22t} & \cdots & q_{2Kt} \\ \vdots & \vdots & & \vdots \\ q_{K1t} & q_{K2t} & \cdots & q_{KKt} \end{bmatrix},$$

where each row of $\boldsymbol{Q}_t$ sums to one.

3.2. *Sampling the hidden states, conditioned on parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\theta}$.* Conditioned on sampled values of the parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\theta}$, and given the observed data $\boldsymbol{y}, \boldsymbol{x}$, and $\boldsymbol{w}$, the posterior full conditional distribution of the hidden state z_t^n at the nth sampling iteration is as follows (dropping the third subscript t from the q variables for clarity):

$$z_t^n | \boldsymbol{\zeta}, \boldsymbol{\theta}, \cdots \sim \text{Multi}\left(\frac{q_{z_{t-1}^n,1} q_{1,z_{t+1}^{n-1}} f_1(\cdot)}{\sum_{k=1}^K q_{z_{t-1}^n,k} q_{k,z_{t+1}^{n-1}} f_k(\cdot)}, \ldots, \frac{q_{z_{t-1}^n,K} q_{K,z_{t+1}^{n-1}} f_K(\cdot)}{\sum_{k=1}^K q_{z_{t-1}^n,k} q_{k,z_{t+1}^{n-1}} f_k(\cdot)} \right),$$

where $f_{z_t}(\cdot) = f(\boldsymbol{y}_t | z_t = k, \boldsymbol{\theta})$ is the emission distribution for state $k = 1, \ldots, K$. Each of the z_t are sampled in succession for all $t = 2, \ldots, T$ at each of the $n = 1, \ldots, N$ iterations of the larger MCMC algorithm. Without loss of generality, we assign the first hidden state, associated with day one of the time series, to state one: $\Pr(z_{t=1} = 1) = 1$.

We can sample the hidden states $\boldsymbol{z}$ using well-known efficient recursive techniques. For example, Scott (2002) describes two Bayesian algorithms for sampling the hidden state of an HMM: a forward-backward (FB) recursive algorithm and a direct Gibbs (DG) sampler. The FB method mixes more rapidly but takes more computational effort. We use the DG method, which can require more iterations (for better mixing) but is less expensive per iteration.

Finite mixture models, including NHMMs, can suffer from the issue of nonidentifiability of the hidden states [Jasra, Holmes and Stephens (2005), Spezia (2009)]. Any pair of states could swap labels and the likelihood would remain invariant, leading to identical marginal posterior densities; see Frühwirth-Schnatter (2006) for a full discussion. Both Scott (2002) and Meligkotsidou and Dellaportas (2011)

discuss this issue for similar HMM and NHMM models, respectively. However, NHMMs are less likely to suffer from label switching compared to HMMs or finite mixtures due to the dependence of the latent states on fixed covariates, which effectively makes label-switching less likely for the states. In particular, for the model we propose in this paper, both the state transitions and the emission distribution parameters are dependent on fixed covariate time series. In our experimental results with rainfall data (described in Section 5) we did not see any evidence of label switching.

3.3. *Sampling for the emission distribution parameters.* For this application we model daily rainfall amounts by a zero-inflated mixture of two exponential distributions, an approach that has been found most effective in past work [Woolhiser and Roldan (1982), Wilks (1998, 1999a, 1999b), Ailliot et al. (2015)]. Other possible modeling options include zero-inflated Gamma distributions or mixtures of exponential, Normal, or Poisson distributions [Hay et al. (1991), Hughes and Guttorp (1994), Charles, Bates and Hughes (1999), Bellone, Hughes and Guttorp (2000), Holsclaw et al. (2016)]. The zero-inflated mixture of two exponential distributions has a physical interpretation of its three components corresponding to no rain, light rain, and heavy rain. The delta function at zero (δ_0) allows for zero inflation for additional dry days, and the light rain and heavy rain each have an exponential distribution, where

$$y_{ts}|z_t, \boldsymbol{\theta} \sim p_{0ts}\delta_0 + p_{1ts}\,\mathrm{Exp}(\lambda_{1z_ts}) + p_{2ts}\,\mathrm{Exp}(\lambda_{2z_ts}), \tag{4}$$

where $z_t = k$ and for this application $\boldsymbol{\theta}$ denotes the mixing probability parameters and rate parameters of the emission distributions. The mixing probabilities $\boldsymbol{p} = (p_{0ts}, p_{1ts}, p_{2ts})$ are assumed to be dependent on the A exogenous variables $\boldsymbol{w}_{ts} = (w_{1ts}, \ldots, w_{Ats})$ and are modeled by a generalized linear model (GLM) through a probit link: $\boldsymbol{p}_{ts} = g^{-1}(\beta_{0z_ts} + \boldsymbol{w}'_{\cdot ts}\boldsymbol{\beta}_{1\cdot s})$ for all a; β_{0z_ts} provides the dependence on the K hidden states, with $z_t = k$ and $w_{ats}\beta_{1as}$ as the mean. Let $\boldsymbol{\beta} = (\beta_{0z_ts}, \beta_{1as})$ for $t \in T$, $a \in A$, and $s \in S$; let $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\beta})$ denote the parameters of the emission distributions. The β_{0z_ts} are state dependent and function like a random effect, whereas the β_{1as} are not state dependent, thus allowing significance testing of the exogenous variable per station.

The probit link for ordered multinomial categories allows for the sampling of the coefficients to be done through the standard Bayesian data augmentation approach [Cox (1970), McCullagh and Nelder (1989), Albert and Chib (1993)]. To allow for conjugate full conditional posterior distributions of the parameters of $\boldsymbol{\beta}$, we need to introduce two sets of latent variables ($\boldsymbol{L}$ and $\boldsymbol{M}$). The first set of latent variables $\boldsymbol{L}$ (with components L_{ts} taking values in the set $\{0, 1, 2\}$) facilitates calculations of $\boldsymbol{p}$. The emission distribution becomes

$$\begin{aligned} y_{ts}|\ldots &\sim p_{0ts}\delta_0 + p_{1ts}\,\mathrm{Exp}(\lambda_{1z_ts}) + p_{2ts}\,\mathrm{Exp}(\lambda_{2z_ts}) \\ &\sim [\delta_0 I_{L_{ts}=0}]\big[\mathrm{Exp}(\lambda_{1z_ts})I_{L_{ts}=1}\big]\big[\mathrm{Exp}(\lambda_{2z_ts})I_{L_{ts}=2}\big], \end{aligned}$$

where $z_t = k$. A second set of latent variables $\boldsymbol{M}$ with components $M_{ts} \sim N(\beta_{0z_t s} + \boldsymbol{w}'_{\cdot ts}\boldsymbol{\beta}_{1\cdot s}, 1)$ is introduced to enable Gibbs sampling of $\boldsymbol{\beta}$. The latent variables $\boldsymbol{L}$ are three ordered categories (no rain, light rain, and heavy rain) which, following the ordered multinomial probit algorithm in Albert and Chib (1993), require one fixed break point (set to zero) and one unknown break point (γ) (more categories would require more unknown breakpoints). The relationship between $\boldsymbol{L}$ and $\boldsymbol{M}$ is as follows:

$$L_{ts} = \begin{cases} 0, & M_{ts} < 0, \\ 1, & 0 < M_{ts} < \gamma, \\ 2, & \gamma < M_{ts}. \end{cases}$$

This results in posterior full conditional distributions as described in Holsclaw et al. (2016). For our rainfall modeling application the λ_{1ks} and λ_{2ks} parameters are each given a low-weight conjugate prior [$\Gamma(1, 1)$] [label switching does not occur because of the ordered nature of the latent variable method of Albert and Chib (1993)]. The $\boldsymbol{\beta}$ coefficients have noninformative priors as well [Albert and Chib (1993)]. This setup leads to the parameters $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\beta})$ having closed-form full conditional posterior distributions that can be sampled via Gibbs steps in the MCMC algorithm.

3.4. *Missing data imputation.* The missing data points can be treated as unknown random variables whose posterior distributions are inferred along with the other variables in the model. The posterior conditional distribution of each missing data point (y^o_{ts} at time t and station s) is given by $\boldsymbol{y}^o_t | z^o_t, \ldots \sim f(\boldsymbol{\theta}^o_{z^o_t}, \boldsymbol{w}_t)$. Data that is missing at random from the observed time series can be imputed as part of the MCMC algorithm. y^o_{ts} can be drawn at each iteration of the MCMC from this distribution, where $\boldsymbol{\theta}^o_{z^o_t}$ and z^o_t are also draws from their posterior full conditional distributions.

3.5. *Predictive conditional chains and forecasting.* New time series of length T can be simulated conditioned on the $\boldsymbol{x}$ and $\boldsymbol{w}$ inputs. In this paper we simulate these forecast chains conditioned on the exogenous variables for held-out years of inputs. First, the exogenous variables $\boldsymbol{x}$ and the sampled coefficients ($\boldsymbol{\zeta}^o$) are used to generate the transition probabilities ($\boldsymbol{q}^*$), and then chains of the hidden states ($\boldsymbol{z}^*$) are simulated. Unlike the scheme for imputing missing data described in Section 3.4, the predictive conditional chains require a predictive draw from the hidden states ($\boldsymbol{z}^*$). Because of the autoregressive nature of the states (the Markov property of the NHMM), the conditional predictive chains can be generated one day at a time, dependent on the previous day. The exogenous variables $\boldsymbol{w}$, their sampled coefficients ($\boldsymbol{\theta}^o$), and the newly generated chains of hidden states are then used to simulate from the emission distribution ($\boldsymbol{y}^*_r$). For a new time step r, this

process can be expressed as

$$
\begin{aligned}
q^*_{ijr} | \boldsymbol{X}_r, \boldsymbol{\zeta}^o &= g^{-1}(\boldsymbol{X}'_r \boldsymbol{\zeta}^o) \\
z^*_r | q^*_{z^*_{r-1} jr} &\sim \text{Multi}(q^*_{z^*_{r-1}1r}, \ldots, q^*_{z^*_{r-1}Kr}) \\
\boldsymbol{y}^*_r | z_r, \ldots &\sim f_k(\boldsymbol{\theta}^o_{z^*_r}, \boldsymbol{w}_r),
\end{aligned}
$$

where $z^*_r = k$ and q^*_{ijr}, z^*_r, and $\boldsymbol{y}^*_r$ are new predictive draws at time r.

Specifically, we use the first 27 years of data (1981–2007) to fit the model, and then generate predictive conditional chains for 2008–2010. These chains can then be compared to three years of held-out observed $\boldsymbol{y}$ data for the purposes of model selection and distributional checks [see Section 6 and also Holsclaw et al. (2017) for plots].

4. Analysis of daily rainfall in India. India has a large population that relies heavily on annual monsoonal rainfall patterns. Variations in rainfall occurrence and amounts can lead to floods and droughts with significant major impacts on food production, hydroelectricity production, and human safety. These variations can be better understood by studying the interactions of daily rainfall with large-scale and regional exogenous weather variables [Wilks and Wilby (1999), Immerzeel, van Beek and Bierkens (2010), Hansen et al. (2006)]. The daily timescale for rainfall modeling is of particular interest because of the effect of flooding, dry spell length, and soil moisture content on agriculture and food supply [Stern and Coe (1984)].

4.1. *Rainfall data.* The rainfall data used in this paper (as briefly described earlier in Section 1) corresponds to daily rainfall amounts[3] between the years of 1981–2010 for a diverse set of 63 weather stations in the Indian region (Figure 1). Stations were selected for inclusion in the data set if no more than 10% of the days for that station had missing observations. This resulted in a total of 689,850 observations over the 30-year period [with leap days removed as in Furrer and Katz (2007)], with 63 daily rainfall time-series $\boldsymbol{y}_{ts}$, $1 \le s \le 63$, $1 \le t \le 10{,}950$.

Figure 4 shows a plot of the seasonal cycle, where each line represents one of the 63 stations, illustrating the diversity of rainfall and its seasonality across the stations. Some stations have strong summer monsoonal maxima, while others are much dryer, and some peak toward the end of the calendar year.

4.2. *Covariate climate indices.* The roles that remote climate "drivers" play in Indian rainfall variability are not fully understood, especially at regional scales, and the potential for prediction remains a topic of active research [e.g., Moron, Robertson and Ghil (2012)]. In this context, we chose six established climate

[3]Data obtained from the U.S. National Centers for Environmental Prediction (NCEP) Climate Prediction Center (CPC) Global Summary of the Day (GSOD) Observations.

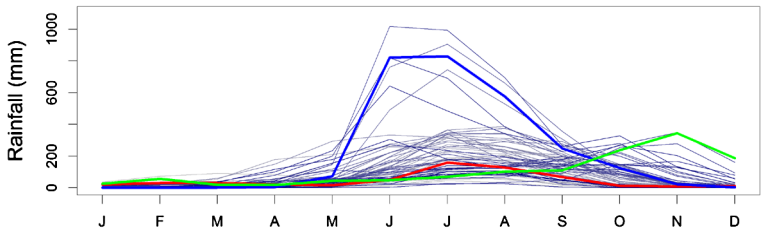


FIG. 4. *Monthly rainfall* (*mm*) *averaged over all years*, *one line for each of the* 63 *stations*. *Stations* 3, 40, *and* 52 *are highlighted with bold lines*. *See also Figures* 1 *and* 2 *for context*.

indices for our model as exogenous variables. All have been shown in previous studies to be associated with rainfall variability over India on different timescales. The variables are Westerly wind Shear Index (WSI), El Niño/Southern Oscillation (ENSO), Indian Ocean Dipole (IOD), Pacific Decadal Oscillation (PDO), and two components of the boreal summer intraseasonal oscillation (BSISO1 and BSISO2). The WSI encodes year-to-year (interannual) changes in the strength of the summer monsoon winds which are closely related with interannual variations in the monsoon rainfall [Wang and Fan (1999), Greene et al. (2011)]. ENSO and IOD are known influences on rainfall on interannual timescales [Gadgil (2003)], whereas PDO has a less well-understood impact [Joseph et al. (2013)]. The monsoon tends to be stronger during the La Nina phase when this ENSO index is *negative* [Gadgil (2003)] and when IOD is positive [Gadgil (2003)]. These aforementioned three variables are closely related to monthly SST. On sub-seasonal timescales Indian monsoon rainfall is impacted by the boreal summer intraseasonal oscillation (BSISO) for which we use the two indices BSISO1 and BSISO2 defined by Lee et al. (2013). Figure 5 shows the six input time series for the years 2008–2010. For a more detailed explanation of each of these variables see Appendix B.

Understanding of these exogenous variables has been hindered by longer timescale nonstationarity, possibly associated with anthropogenic climate change, or the remote impacts of other ocean basins [Gershunov, Schneider and Barnet (2001)]. Our approach is thus to include all six indices as candidate covariates, where BSISO is given as daily values and the monthly series (ENSO, WSI, IOD, PDO) are interpolated linearly to daily values.

In our model, there are two ways exogenous variables can be included: a station-level A-dimensional time series $\boldsymbol{w}$ with components w_{ats} or a global B-dimensional time-series $\boldsymbol{x}$ with components x_{bt}. The station-dependent variables ($\boldsymbol{w}$) are *local* in nature and directly influence the mixing weights of the point mass at zero and mixture of exponential distributions of the emission distribution for each station individually. Lower frequency climatic drivers tend to impact the climatic background, and we thus introduce the impacts of the WSI, ENSO, IOD, and PDO climate drivers via $\boldsymbol{w}$, directly influencing the characteristics of the emission

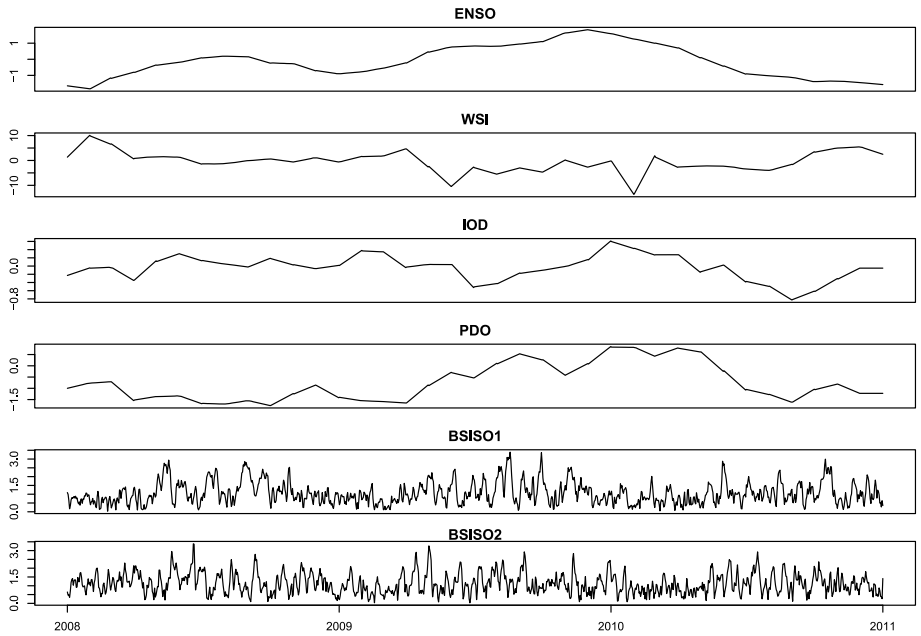


FIG. 5. *Exogenous variables*: *ENSO*, *WSI*, *IOD*, *PDO*, *BSISO*1, *and BSISO*2 *for three years*. *ENSO*, *WSI*, *IOD*, *and PDO are calculated by linearly interpolating monthly values to the daily timescale*. *BSISO*1 *and BSISO*2 *are available on a daily basis*.

distributions. In addition, a station-specific seasonal cycle (annual and biannual harmonic terms—four total terms) and a long-term drift term are included in $\boldsymbol{w}$.

In contrast, the large-scale time-dependent exogenous variables ($\boldsymbol{x}$) are not station-specific—they affect the whole region and influence the transition probabilities of the hidden states of the NHMM. Indian monsoon rainfall is mostly generated by local scale thunderstorm activity and monsoon depressions, while mid-latitude western disturbances are important over northern India, especially in winter. On sub-seasonal timescales the paths and intensities of these phenomena are controlled by large-scale atmospheric circulation patterns that can be naturally represented by a discrete set of weather states and the transitions between them [Ghil and Robertson (2002)]. These are modulated by the BSISO whose impacts are thus encoded in the model via the $\boldsymbol{x}$ variable influencing the transition matrix. Additionally, a general seasonal cycle is included for the state transitions in $\boldsymbol{x}$, which also has terms to fit annual and biannual harmonics due to seasonal cycles (i.e., there are a total of four sine and cosine terms in $\boldsymbol{x}$).

5. Results. In this section we assess the model's ability to capture distributional, temporal, and other aspects of rainfall, as well as investigating the effects of the exogenous climate variables through information gained from the parameter

uncertainty estimates. After fitting the model using 27 years of daily data (1981–2007), we simulated 1000 chains of length 27 years for the 63-station network, conditioned on the corresponding 27 years of exogenous variables $\boldsymbol{w}$ and $\boldsymbol{x}$, to produce the figures in this section. The last 3 years (2008–2010) of observed data ($\boldsymbol{w}$, $\boldsymbol{x}$, and $\boldsymbol{y}$) were held out. These 3 years of held-out data were used for model selection (Appendix A) and also to compare with predictive conditional chains [with plots shown in Holsclaw et al. (2017)]. For model selection, we use a combination of the Bayesian Information Criteria (BIC) and predictive log-probability scores (PLS) for selecting the number of hidden states and selecting among different combinations of exogenous variables. All of the results in the remainder of the paper are for the selected model with $K = 7$ states which was used to generate 1000 simulated chains of 27 years of data.

5.1. *Hidden states*. From the MCMC algorithm, we sample the hidden states (z) and marginalize over the iterations of the algorithm to find the most probable hidden state for each day (similar to a Viterbi sequence [Forney (1973)]). Figure 6 shows the mean daily rainfall amount at each station for each of the hidden states. The top of each pane indicates the number of days assigned to each state given the most probable state sequence (there are a total of $27 \times 365 = 9855$ days). State 1 represents largely dry days across the whole domain (some stations have little to no rainfall and have no dot), with moderate rainfall occurring in states 2 and 3. State 5 characterizes heavier rainfall over north-central India. The heaviest rainfall occurs along the western coasts in states 6 and 7, while state 4 is unique in representing rainfall over the southeast coast.

Figure 7 shows attributes of the daily sequences of states. The left panel shows year-long chains of most probable state sequences for each year of the data set,

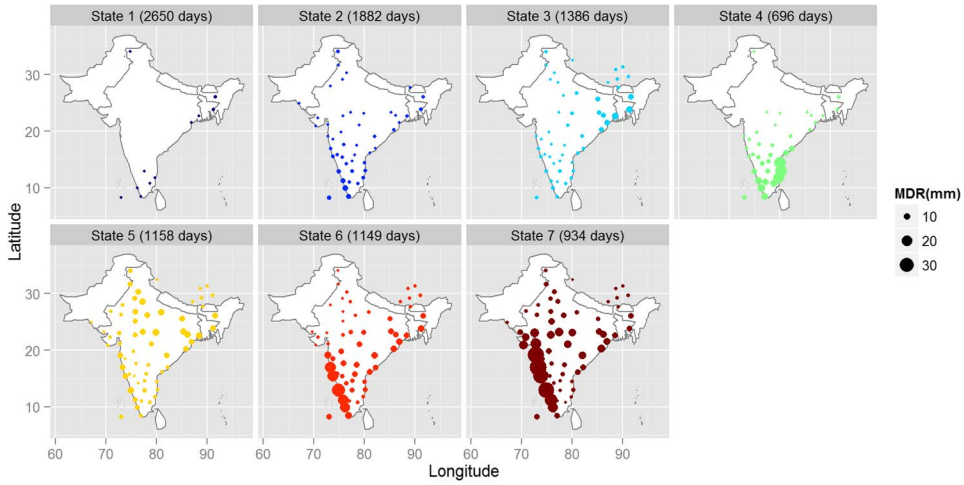


FIG. 6. *Mean daily rainfall amount for each of the hidden states, given by circle size.*

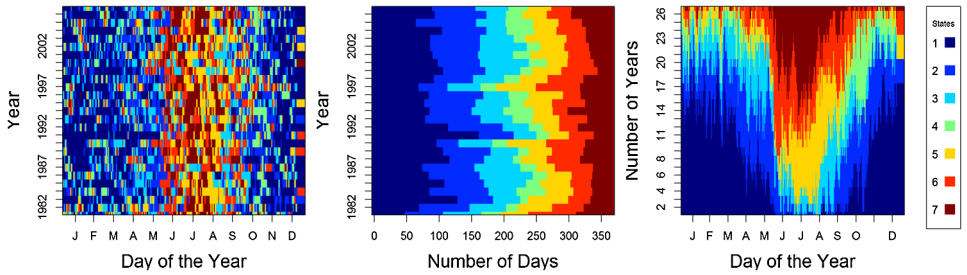


FIG. 7. *Most probable sequence of hidden states* (*left*), *together with the annual averages of number of days per state* (*middle*), *and averages across years for each calendar day* (*right*).

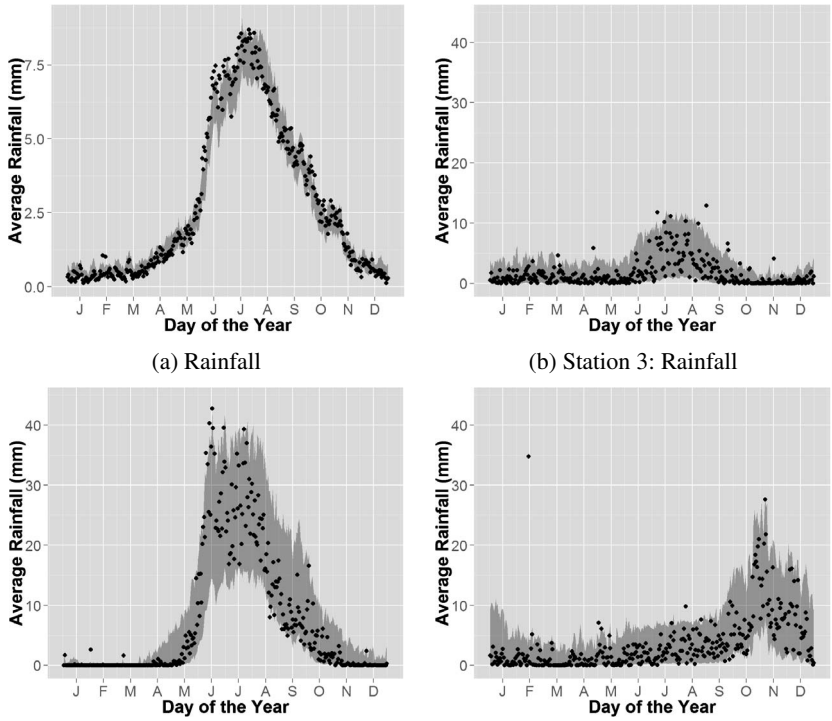
illustrating the dramatic seasonality of the summer monsoon together with a large amount of sub-seasonal and interannual variability with a stochastic character. The middle panel sums along the rows to depict variability in the annual counts of each state. The right panel sums by column to depict the seasonality and has the counts of each state per day of the year given that we observed 27 years; January 1st is on the left and December 31st is on the right. The state occurrence frequencies can be seen to follow distinct seasonal patterns.

The temporal distributions of each state can be naturally understood in terms of the rainfall climatology of India by referring to their temporal evolutions, shown in Figure 7. The wetter states occur more frequently during the summer monsoon season, while state 4 is characteristic of the winter monsoon over SE India, peaking in boreal autumn (right pane of Figure 7).

5.2. *Rainfall simulations*. In this section, we show both the seasonality and distribution of rainfall for the average across all stations as well as for three specific and diverse stations. While it is straightforward to capture the seasonality and rainfall distribution at a single station, doing it jointly across multiple stations is nontrivial [Charles, Bates and Hughes (1999), Bellone, Hughes and Guttorp (2000)]. The NHMM approach provides a useful mechanism for addressing this joint modeling problem by conditioning the stations on common shared state variables. The figures below are for the chains simulated on the first 27 years of data; see Holsclaw et al. (2017) for similar plots for the 3 held-out years of data.[4]

5.2.1. *Seasonality*. Figure 8 provides an indication of how well the model captures seasonality. The observed average rainfall per day over 27 years is shown by the black points in the figure, and the NHMM simulations correspond to the 95% probability interval (PI) bands in gray for the 1000 simulated sets. Figure 8(a)

[4]Holsclaw et al. (2017) also includes distributional plots showing the NHMM's ability to capture dry spell and wet spell lengths, inter-annual variability of mean rainfall, dry days, and heavy rainfall events.



(a) Rainfall

(b) Station 3: Rainfall

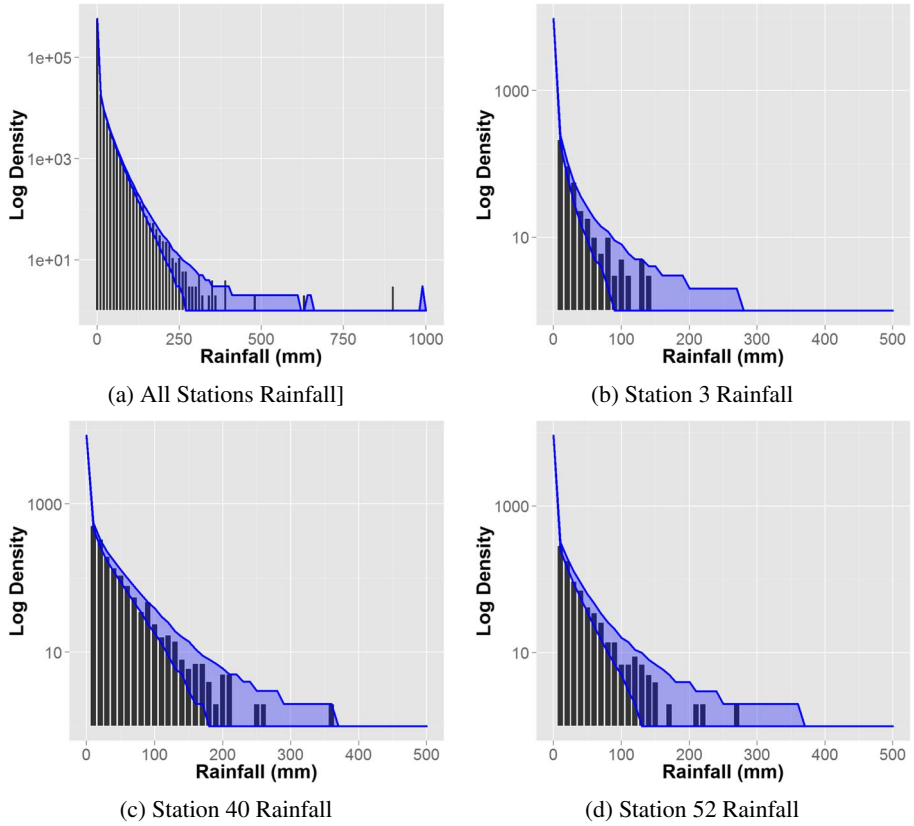
(c) Station 40: Rainfall

(d) Station 52: Rainfall

FIG. 8. *Observed data averaged over* 27 *years* (*black*) *and* 1000 *simulated sets and their* 95% *PI bands* (*gray*). (a) *Rainfall averaged over all stations*, (b) *Station* 3 (*dry*), (c) *Station* 40 (*wet summer*), *and* (d) *Station* 52 (*wet winter*).

shows that the seasonality of the simulated data from the model is similar to the observed data averaged over all stations. Figures 8(b), 8(c), and 8(d) show the same seasonal plots for rainfall but for the three contrasting Stations 3, 40, and 52. The seasonality of the simulated data is similar to the observed data, when averaged over all stations, as well as for the three individual stations with diverse climatologies. [Holsclaw et al. (2017) includes similar figures for all of the individual stations.]

5.2.2. *Distributional checks*. Also of interest is the NHMM's ability in capturing the large-scale distributional properties of the observed rainfall data. Figure 9 shows the observed 27 years of data in the gray histogram and the 95% PI bands for the simulated chains (densities are plotted on a logarithmic scale). Figure 9(a) is averaged over all stations; Figure 9(b), 9(c), and 9(d) show the data density for the same dry and wet stations as in Figure 8. These plots show that the distribution of the observed data is reasonably well represented by the model [see Holsclaw et al. (2017) for results pertaining to the dry spells for more details on the dry days



(a) All Stations Rainfall] (b) Station 3 Rainfall

(c) Station 40 Rainfall (d) Station 52 Rainfall

FIG. 9. *Observed data density on the log scale* (*gray*) *and* 1000 *simulated chains for* 27 *years given by the* 95% *PI bands*. (a) *Rainfall averaged over all stations*, (b) *Station* 3 *which is drier*, (c) *Station* 40 *which is wetter in summer, and* (d) *Station* 52 *which is wetter in winter.*

distribution]. Each of the diverse stations are well modeled, from wetter to drier locations.

5.3. *Model diagnostics of climate controls*. As described in Section 4.2, several covariates are included in the model. The BSISO1 and BSISO2 variables in the $\boldsymbol{x}$ vector impact the hidden state evolution on the daily timescale, and the ENSO, WSI, IOD, drift, and PDO in the $\boldsymbol{w}$ vector impact the mixing weights of the emission distributions on the monthly scale. There are $K-1$ coefficients for the transition probabilities and J coefficients for the emission distributions (one for each station). Figures 10 and 11 show the inferred values of the coefficients for the exogenous variables $\boldsymbol{x}$ and $\boldsymbol{w}$, respectively, and their 95% PIs [for the last 4000 draws from the posterior to ensure full convergence had happened, see Holsclaw et al. (2017) for trace plots]; they are considered to be statistically significant if the PIs do not contain zero (vertical dashed line). The Bayesian approach has made

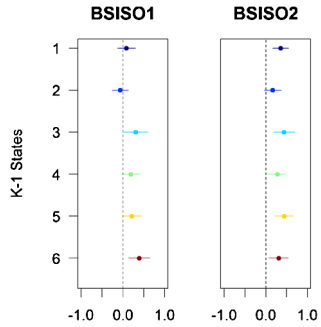


FIG. 10. *Coefficients for the exogenous variables* $\boldsymbol{x}$ *influencing the transition probabilities for each of* $K-1$ *states. There is one dot for each of the* $K-1$ *states (the* k*th state is set to zero). The* 95% *PI bands are given as a line around each dot.*

this type of significance testing of the exogenous variables possible; many other NHMM algorithms only find point estimates for parameters of interest.

Figure 10 shows the coefficients for the exogenous variables of $\boldsymbol{x}$, which affect the transition probabilities of the hidden state evolution. Of the $K-1$ coefficients for each exogenous variable, at least one is well away from zero in each set. In the case of BSISO2, all the coefficients are statistically significant, as their 95% PIs do not contain zero. The four seasonal harmonic input coefficients are not shown, but were also all significant. Figure 11 shows the coefficients for the exogenous variables ($\boldsymbol{w}$) for the emission distribution, one for each of the 63 stations (Station 1 at the bottom of the figure through Station 63 at the top). The harmonic terms representing rainfall seasonality are highly significant for most stations, consistent with Figure 4. Most of the other exogenous variables are significant at least at several stations, although their impacts are understandably much weaker than the seasonal

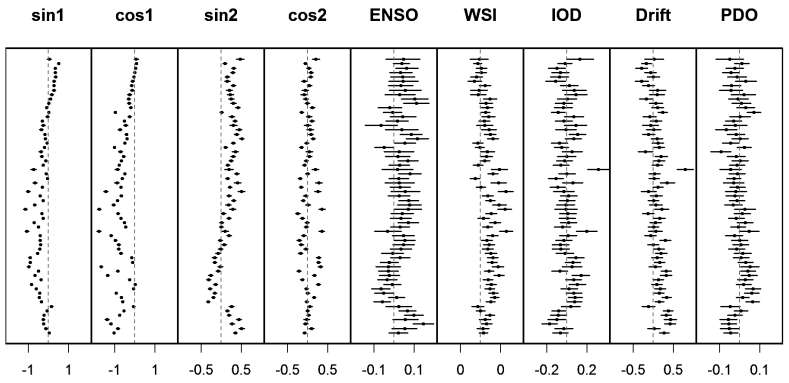


FIG. 11. *Coefficients for the exogenous variables* $\boldsymbol{w}$ *influencing the emission distributions for each of* J *stations, with Station* 1 *at the bottom through Station* 63 *at the top. The* 95% *PI bands are given as a line around each dot.*

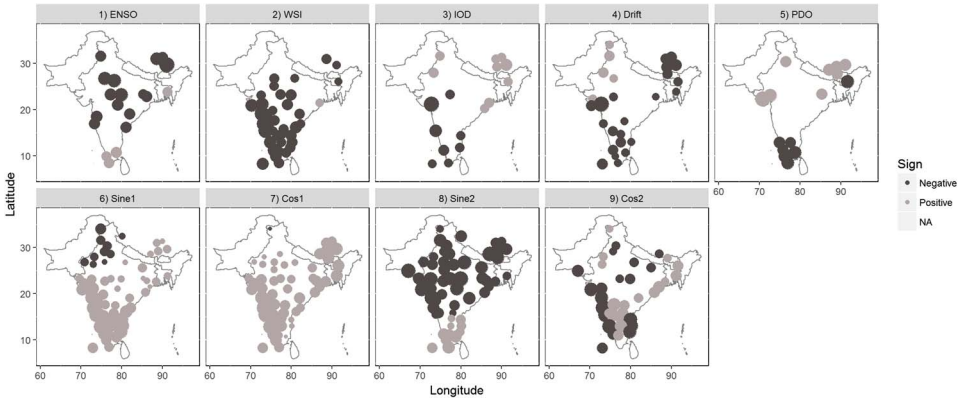


FIG. 12. *Coefficients for the exogenous variables* $\boldsymbol{w}$ *influencing the emission distributions for each of* J *stations*. 95% *PI for the coefficients having positive values* (*dark*) *and negative values* (*light*) (*coefficients containing zeros omitted*); *the magnitude of the coefficient is given by the relative size*.

modulation. Figure 12 shows the mean of the coefficient values for the climate covariates, plotted geographically to highlight any spatial coherency and regionality in the relationships (note that the coefficient magnitudes depend on the scale of the covariate and are thus not comparable between panels). There is some indication that certain subregions are affected preferentially by particular exogenous variables, with the circulation index WSI showing the broadest scale impact. This is consistent with the direct physical relationship between monsoon rainfall and winds, while the remote climate modes (ENSO, IOD, PDO) have weaker impacts [Gadgil (2003)].

Figure 13 shows mean rainfall amount over 1000 simulated chains versus day of the year, for the minimum (light) and maximum value (dark) of each of the exogenous variables (with all other inputs held at their mean values). The figure shows little difference in mean rainfall for the minimum and maximum values of the ENSO, PDO, and IOD covariates when averaged over all stations. However, there is a marked difference in the minimum and maximum WSI value on the average rainfall amount, consistent with the broad scale geographical impact seen in Figure 12. There is a slightly longer and heavier monsoon when the drift term is higher, indicating an upward trend in rainfall over the 27-year period. BSISO1 and BSISO2 amplitudes are the only inputs prescribed on a daily basis (whereas the other inputs are monthly). Smaller BSISO1 amplitudes are seen to be associated with longer and heavier monsoon seasons, while smaller BSISO2 amplitudes are associated with heavier monsoon seasons but of the same duration. Thus, the monsoon tends to be stronger when the intraseasonal oscillation is less active, which is physically consistent with fewer dry monsoon "breaks" in those years.

Figure 14 shows similar 1000-chain averaged annual simulations as Figure 13, but for the three selected stations. Years with strong monsoonal wind shear anomalies (WSI) are associated with a much longer summer monsoon rainfall season at

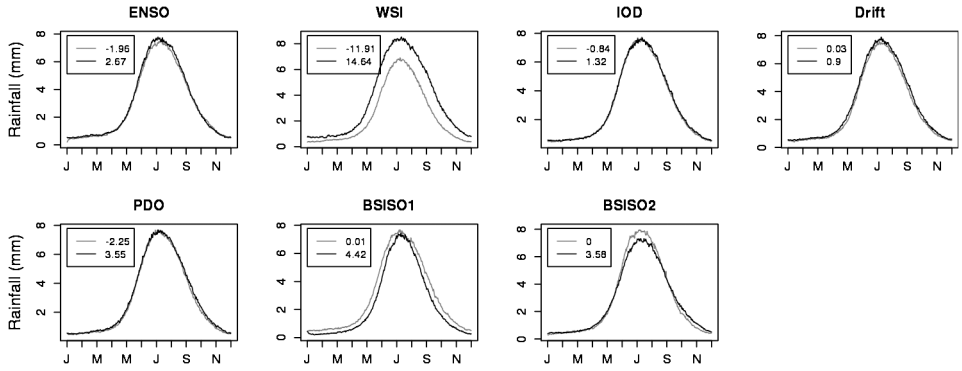
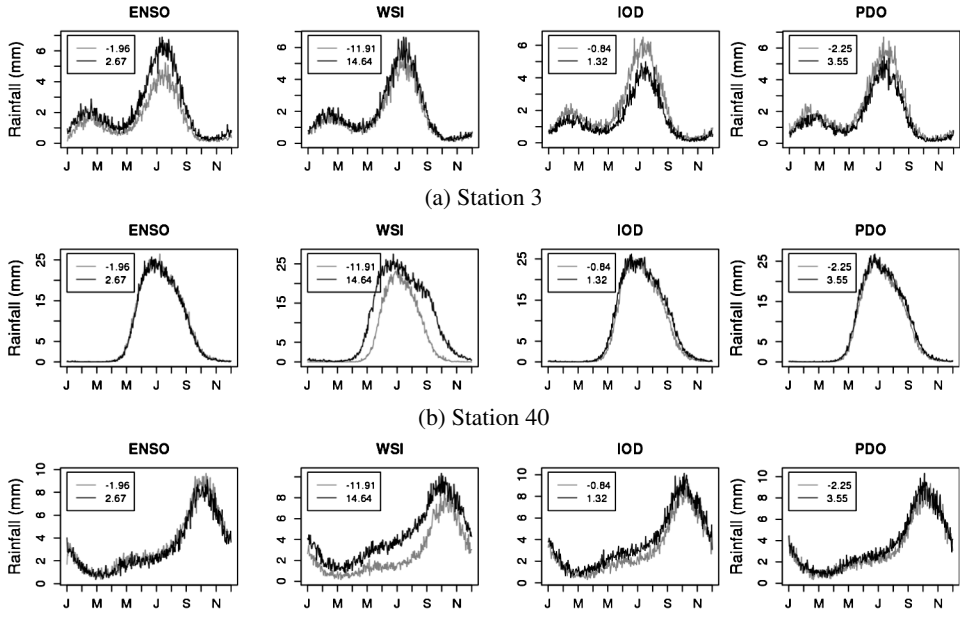


FIG. 13. *Maximum* (*dark*) *and minimum* (*light*) *for ENSO*, *WSI*, *IOD*, *drift*, *PDO*, *BSISO*1, *and BSISO*2 *for each day of the year versus mean rainfall over* 1000 *simulated chains averaged over all stations* [*see Holsclaw et al.* (2017) *for individual stations*].

the very wet station (Station 40) on the west coast, while the impact is on peak rainfall at the "dry" station in NW India (Station 3), not duration of the season. The SE India station (Station 52), while nominally in the fetch of the NE monsoon that peaks in autumn, nonetheless also feels the summer SW monsoon as well when the monsoonal circulation (given by WSI) is strong, resulting in an extended



(a) Station 3

(b) Station 40

(c) Station 52

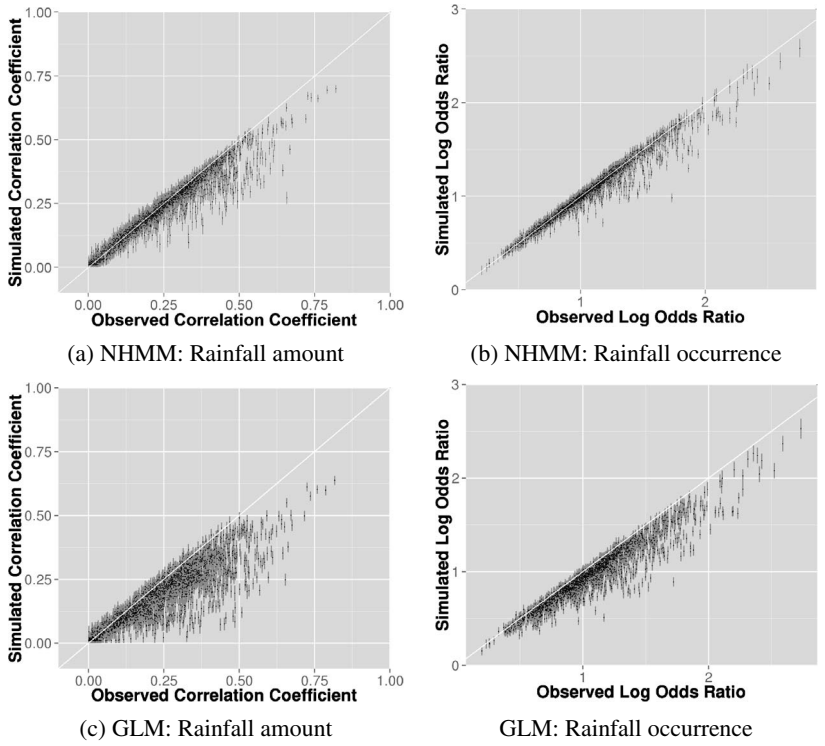
FIG. 14. *Three specific stations*: *maximum* (*dark*) *and minimum* (*light*) *for ENSO*, *WSI*, *IOD*, *and PDO for each day of the year versus mean rainfall over* 1000 *simulated chains*.

rainfall season from May–Jan. The indirect climate covariates again have smaller impacts, though they can be quite large at the individual stations. Their impacts are large during summer at Station 3 over inland NW India, although physical interpretation is not straightforward. [See Holsclaw et al. (2017) for additional stations and exogenous variable plots.]

5.4. *Spatial modeling*. Finally, we assess the model's ability to reproduce the observed spatial correlations of the rainfall patterns. Stations that are closer together tend naturally to have more highly correlated measurements, although this can be modulated somewhat by local topography. The two-way station correlations of daily rainfall, for amount and occurrence, range between (0.04, 0.41) and (0.04, 0.77), respectively. We use two measures of spatial correlation between each pair of stations, the log odds ratio for occurrence and the Spearman's rank correlation coefficient for rainfall amount. The log odds ratio is calculated for the occurrence (binary classification) as the log of the number of matched days between the two chains divided by the number of differences [Hughes, Guttorp and Charles (1999)].

We compare the observed empirical pairwise correlations to correlations obtained from the 1000 simulated chains of daily rainfall of the same length (each 27 years in length) in Figure 15. The x-axis corresponds to the observed correlations and the y-axis to the simulated correlations. 95% PI bands are included for the simulations. If the model were able to fully reflect the observed spatial correlations, then the points in the figure should lie around the diagonal line. The upper panel shows that the spatial correlations tend to be systematically underestimated by the NHMM, which is a known issue in rainfall modeling when assuming that the station variables are conditionally independent given the state variable [Hughes, Guttorp and Charles (1999), Kirshner (2010), Germain (2010)]. For comparison, the spatial correlations from an equivalent GLM model (with no state structure, which includes the $\boldsymbol{w}$ variables, but not $\boldsymbol{x}$ variables, as described in Appendix A as Model 3) are shown in the lower panel. The simulated GLM correlations are less accurate than those of the NHMM, indicating that the state variables are contributing to better modeling of spatial dependence. Further improvements could be made by going beyond the conditional independence assumptions within each state, for example, by using the type of tree-structured station dependence developed in Kirshner, Smyth and Robertson (2004). [See Holsclaw et al. (2017) for similar plots for the three held-out years of data.] For more analysis of isotropy and correlation see Holsclaw et al. (2017); these plots show that the NHMM is capturing most of the correlation from the data. Future work could include adding more complex spatial structure to the model, but this would have to consider the computational cost as some changes would be prohibitive.

6. Conclusion. We described a Bayesian implementation of the NHMM based on the Pólya-Gamma latent variable scheme, allowing for analysis of larger multivariate data sets than possible with prior approaches. The model allows for



(a) NHMM: Rainfall amount (b) NHMM: Rainfall occurrence

(c) GLM: Rainfall amount GLM: Rainfall occurrence

FIG. 15. *Pairwise station spatial correlation for the observed* (*x-axis*) *and the* 1000 *simulated chains* (*y-axis*). *The dot is the mean of the* 1000 *simulated chains*, *and* 95% *PI bands are in gray. Top*: *NHMM with state structure*. *Bottom*: *GLM model with no state structure*.

exogenous variables to influence both the transition probabilities of the hidden states and the emission distributions. The overall approach is flexible in that it can handle nonnormally distributed multiple time series of daily data such as rainfall. Sampling is done through a data augmentation approach which removes the need for tuning parameters and makes it nearly automatic.

We illustrated how the framework allows fitting a multivariate NHMM for daily rainfall simulation allowing for incorporation of covariate information of different forms, which is particularly attractive for downscaling of global climate model predictions and projections. In particular, we applied the approach to modeling of rainfall data over a large historical collection of daily weather station data across the Indian region. The general distributional properties of Indian rainfall, including seasonality, were shown to be well captured by the model, and spatial correlations between stations were adequately captured by the hidden states. The model was shown to provide particular meteorological insight into the roles of monsoon wind shear strength and intraseasonal wave activity on the seasonality of rainfall in the Indian region. In particular, it enables a novel analysis of rainfall variability integrated from daily-to-seasonal timescales and from local-to-subcontinental spatial

scales. This complements the methods traditionally used in monsoon diagnostic and predictability studies that often focus on correlation analysis between variables for a particular spatio-temporal scale, such as the predictability of seasonal averages of all India rainfall, for example [Shukla and Paolino (1983)]. More broadly, the Bayesian framework provides uncertainty estimates for all parameters of the model, allowing for assessment of the impact of each exogenous climate variable. The NHMM can be used to generate predictive chains and chains with different levels of the exogenous variables, providing both chains of the hidden states and emission distributions that can be used for model comparison and conditional simulation.

Future work could be aimed at adding a more complex spatial structure between stations to this Pólya-Gamma NHMM. Or the conditional independence assumption could be relaxed and subregions could be considered each with their own states. On smaller data sets, some have allowed K to be estimated; this could be added to the model. We could also change the modeling assumptions to estimate a coefficient for each of the exogenous variables for each of the states (not the same coefficient for all states). Other NHMM assumptions, like having individual coefficients for the transition probabilities, could include ρ_{ij} instead of just ρ_j in equation (1). These types of changes to the proposed model would require the consideration of trade-offs between model expressiveness and computational cost.

APPENDIX A: MODEL SELECTION

Model selection criteria. Model selection has two purposes: to choose the number of states (K) and select the variables in $\boldsymbol{x}_t$ and $\boldsymbol{w}_{ts}$. The number of hidden states needs to be determined for the model. More flexible models that allow for change in dimension tend to have complex algorithms that are computationally expensive (i.e., reversible-jump MCMC) [Green (1995), Robert, Rydén and Titterington (2000), Meligkotsidou and Dellaportas (2011)]. With such large data sets, we can run the NHMM with a few values of K (number of hidden states) and use a model selection criterion to choose an optimal number of states. The other model choice is the selection of meaningful exogenous variables. Many exogenous variables can be included in $\boldsymbol{w}$ and $\boldsymbol{x}$, but some may not be statistically significant.

There are two types of metrics we can consider for the model choice decision: model fit metrics (in-sample) or predictive metrics (out-of-sample). Standard in-sample model fit metrics, like Akaike information criterion (AIC), Bayesian information criterion (BIC), and deviance information criterion (DIC), account for the number of parameters used in the model [Akaike (1974), Schwarz (1978), Spiegelhalter et al. (2002)]. We found that in this model, because some of the assumptions of the DIC failed to hold, it tended to choose over-parametrized models. Bayes factors (BF) are not considered because the model has noninformative priors on many of the parameters [Kass and Raftery (1995), Dempster (1974)]. Following the recursive method given in Scott (2002) to calculate the log-likelihood, we report the BIC values which are calculated as negative two times the log-likelihood

plus a penalty for the number of parameters times the log number of observations [the parameters are counted: $S = 63$, $A = 9$, $B = 6$, $K \times (K - 1)$ for $\boldsymbol{\xi}$, $B \times (K - 1)$ for $\boldsymbol{\rho}$, $K \times S$ for $\boldsymbol{\beta}_0$, $A \times S$ for $\boldsymbol{\beta}_1$, $(K - 2) \times S$ for the $\boldsymbol{\gamma}$ cutpoints, $2 \times S \times K$ for $\boldsymbol{\lambda}$, and $\boldsymbol{z}$, $\boldsymbol{p}$, $\boldsymbol{L}$, $\boldsymbol{M}$ are latent variables that integrate out and are not counted]. The second type of model metric is a predictive measure (out-of-sample), and we consider it the preferable method in this situation. We plan to hold out the last three years of the time series and perform a predictive log score (PLS) [Gneiting and Raftery (2007), Meligkotsidou and Dellaportas (2011)]. After the last observation y_{Ts}, there are $r = 1, \ldots, R$ predictive time steps to the simulated chain ($R = 3 \times 365$ for three years of predictive chains); the PLS is given by $\sum_{r=1}^{R} \log(1/N \sum_{n=1}^{N} \prod_{s=1}^{S} f(y^*_{T+r,s}))$, where N is the number of MCMC iterations.

The PLS is calculated for the predictive ability of the model, and the BIC is calculated for the model fit. The BIC is calculated from the 27-years model fit (p is the parameter count of the model), and the PLS compares held-out three years to predictive conditional chains of three years generated from the model; see Table 2 and Table 1 for results. Table 1 shows several models for comparison with the NHMM and some different configurations of the exogenous variables. Table 2 shows several options of the number of hidden states for the preferred model from Table 1. Overall, the hidden states describe general spatial rainfall patterns; some states capture drier days, while others describe wetter weather patterns. This large region may require more states (e.g., seven to twelve) than a more local region, where three to six states might suffice.

Model selection results. We run a few baseline models to compare with the NHMM; see Table 1. Model 1 is set up with no states ($K = 0$) and no exogenous variables; this spreads rainfall homogeneously throughout the year. Model 2 is a standard weather state NHMM with all $\boldsymbol{x}$ inputs [i.e., BSISO1, BSISO2, seasonality (four harmonic components)]; the optimal number of states for this model is $K = 8$ using BIC. Model 3 treats each station independently ($K = 0$) with the GLM linking inputs in $\boldsymbol{w}$ [i.e., seasonality (four harmonic components), ENSO,WSI, IOD, Drift, PDO] with the mixing weights (there are no hidden states, thus no $\boldsymbol{x}$ inputs). This model is similar to Katz and Parlange (1995), Furrer and Katz (2007), Ailliot and Monbet (2012), where only a single station is modeled. Model 4 uses all inputs of $\boldsymbol{x}$ and $\boldsymbol{w}$; the placement of the inputs into either $\boldsymbol{x}$ or $\boldsymbol{w}$ was chosen by the climate scientists based on physical properties of the variables (i.e., larger regional variables and shorter timescale in $\boldsymbol{x}$). Other combinations and inputs were tested, but these were the ones that were significant to the model. One of the other models tested was one with only $\boldsymbol{x}$ inputs, but it did not perform as well as models including $\boldsymbol{w}$.

Two other models were also considered that had similar (slightly worse) BIC and PLS scores to Model 4. One was a GLM-HMM [Holsclaw et al. (2016), Heaps, Boys and Farrow (2015)] which includes all possible exogenous variables in $\boldsymbol{w}$.

TABLE 1
Comparing models for rainfall

No.	Model	States	p	BIC	PLS
1.	No inputs (no $\boldsymbol{x}$ or $\boldsymbol{w}$)	$K=0$	252	1,565,454	−85.7
2.	NHMM for $\boldsymbol{x}$ (no $\boldsymbol{w}$)	$K=8$	2079	1,360,954	−74.2
3.	Indep. GLM with $\boldsymbol{w}$ (no $\boldsymbol{x}$)	$K=0$	819	1,406,455	−77.0
4.	NHMM with $\boldsymbol{x}$ and $\boldsymbol{w}$ partial	$K=7$	2283	1,344,892*	−73.6*

∗ indicates the best value.

This model performed similarly in metrics but has some limitations. Because of climate change, it is of interest to forecast daily rainfall based on evolving $\boldsymbol{x}$ variables to modulate the hidden state distributions; the GLM-HMM type of model is stationary and has no mechanism for forecasting climate change like the NHMM. Additionally, another model was considered where all exogenous variables were included in both $\boldsymbol{x}$ and $\boldsymbol{w}$. This model performed similarly in metric scores to Model 4 (where variables were limited to being in either $\boldsymbol{x}$ or $\boldsymbol{w}$), but this model had far more parameters and suffered from lack of physical interpretability, as the coefficients of $\boldsymbol{x}$ and $\boldsymbol{w}$ were highly correlated. The most parsimonious and interpretable model is the NHMM with exogenous variables each included once, either in $\boldsymbol{x}$ or $\boldsymbol{w}$ based on their physical characteristics. Each algorithm was run 2000 iterations with an additional 10% burn-in; the samples of the parameters converged quickly to stationarity and the samples mixed well with no thinning [see Holsclaw et al. (2017) for run times and trace plots]. The final model was run 10,000 iterations with an additional 20% burn-in period.

Model 4 had preferable BIC and PLS metrics over all other models. Table 2 shows the the metrics for choosing the number of hidden states (K) for this model (other models had similar values of K). The table also shows the number of parameters (p); for parsimony we want to choose a model with maximum PLS, minimum BIC, and minimum p ($*$ denotes these values on the table). $K=1$ denotes a model with a single constant state (which is equivalent to having no states $K=0$). Table 2 shows that the BIC achieves local minima around seven to ten states. The PLS continues to improve with increased number of states. For parsimony it is sometimes best to choose the number of states where the PLS value is no longer improving as rapidly; this also happens around seven to ten states.

We compare the PLS scores of the NHMM (Model 4) to a baseline model with no states (Model 3). Model 3 fit independent GLM models to each station, whereas Model 4 includes the hidden states and spatial information. Model 3 has a PPL of -77.0 for the baseline model compared to the NHMM with $K=7$ states with a PPL of -73.6. The difference between the two log scores over the three forecast years (predictive conditional chains) is given by $\exp((-73.6-(-77.0))/3)=3.1$, which means the NHMM is 3.1 times better at annual predictive ability. Also, we

TABLE 2
Choosing the number of states for the rainfall model

K	p	BIC	PLS
0–1	693*	1,405,643	−77.1
2	953	1,376,646	−75.6
3	1215	1,366,881	−75.0
4	1479	1,356,233	−74.4
5	1745	1,351,150	−74.0
6	2013	1,346,841	−74.0
7	2283	1,344,892	−73.6
8	2555	1,342,660	−73.6
9	2829	1,342,309	−73.5
10	3105	1,341,327*	−73.3
11	3383	1,341,566	−73.2
12	3663	1,341,514	−73.1*
13	3945	1,341,798	−73.2
14	4229	1,342,378	−73.0*
15	4515	1,342,853	−73.0*

X and W have inputs based on physical properties of the region ($*$ desirable numerical score).

compare the PLS for Model 4 with $K = 7$ states and $K = 15$ states: $\exp((-73.0 - (-73.6))/3) = 1.2$, and find only a 1.2 times better annual predictive ability by including the eight additional states.

APPENDIX B: CLIMATE VARIABLE DETAILS

Six established climate indices related to rainfall in India are as follows: Westerly wind Shear Index (WSI), El Niño/Southern Oscillation (ENSO), Indian Ocean Dipole (IOD), Pacific Decadal Oscillation (PDO), and two components of the boreal summer intraseasonal oscillation (BSISO1 and BSISO2).

1. WSI: Year-to-year (interannual) changes in the strength of the summer monsoon winds are closely linked with monsoon rainfall variations, and we use the Westerly Shear Index (WSI), as defined in Wang and Fan (1999) as WSI1, to represent these. The WSI is defined by the vertical shear of the zonal wind (u850–u200) averaged over the box (5N–20N, 40E–80E), and was used in an NHMM for Indian rainfall by Greene et al. (2011). We use monthly averaged values, with the mean seasonal cycle subtracted, so as to focus on interannual variations in the monsoon circulation.

2–3. ENSO and IOD: ENSO and IOD indices were computed from the NOAA Extended Reconstructed Sea Surface Temperature Dataset, version 3b [Smith et al. (2008)], via the IRI Data Library (http://iri.columbia.edu). The El Niño/Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) are

TABLE 3
Two-way input correlations

	ENSO	WSI	IOD	PDO	BSISO1	BSISO2
ENSO	—	−0.51	0.28	0.41	−0.01	0.02
WSI		—	−0.19	−0.25	−0.01	−0.04
IOD			—	0.09	−0.03	−0.02
PDO				—	−0.08	0.01
BSISO1					—	0.07
BSISO2						—

known influences on rainfall on interannual timescales [Gadgil (2003)]. Monthly sea surface temperature (SST) in the Nino3.4 region (150W–90W, 5N–5S) are used to define the ENSO index; the monsoon tends to be stronger during the La Nina phase, when this ENSO index is *negative* [Gadgil (2003)]. The IOD index is defined by the difference in monthly SST anomalies in the western (50E–70E, 10N–10S) and eastern (90E–110E and 0S–10S) equatorial Indian Ocean; the monsoon tends to be stronger when IOD is positive [Gadgil (2003)].

4. PDO: While the Pacific Decadal Oscillation (PDO) has a less well-understood impact [Joseph et al. (2013)], the PDO index is defined by Zhang, Wallace and Battisti (1997) to be the leading PC of monthly SST anomalies in the North Pacific Ocean, poleward of 20N. The monthly mean global average SST anomalies are removed to separate this pattern of variability from any "global warming" signal that may be present in the data. This data set set is from the University of Washington (http://research.jisao.washington.edu/data_sets/pdo/).

5–6. BSISO1 and BSISO2: On sub-seasonal timescales Indian monsoon rainfall is impacted by the boreal summer intraseasonal oscillation (BSISO), data obtained from the APEC Climate Center (APCC, http://www.apcc21.org) [Lau and Chan (1986), Yoo, Robertson and Kang (2010)], for which we use the two indices BSISO1 and BSISO2 defined by [Lee et al. (2013)].

The cross-correlations between these daily series are given in Table 3 and are relatively low. Monsoon circulation anomalies (WSI) are quite strongly related to ENSO and PDO ($r = -0.51$ and 0.41 resp.), less strongly with IOD ($r = 0.28$), but not to the BSISO.

Acknowledgments. The rainfall data set was obtained from the Climate Prediction Center, National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, from the Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory (http://rda.ucar.edu/datasets/ds512.0).

SUPPLEMENTARY MATERIAL

Additional Results and Figures (DOI: 10.1214/16-AOAS1009SUPP; .pdf). The Supplemental Material includes figures for each individual station for many of the metrics and plots. A few additional results and metrics are also included.

REFERENCES

AILLIOT, P. and MONBET, V. (2012). Markov-switching autoregressive models for wind time series. *Environ. Model. Softw.* **30** 92–101.

AILLIOT, P., ALLARD, D., MONBET, V. and NAVEAU, P. (2015). Stochastic weather generators: An overview of weather type models. *J. SFdS* **156** 101–113. MR3338244

AITCHISON, J. and BENNETT, J. (1970). Polychotomous quantal response by maximum indicant. *Biometrika* **57** 253–262.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19** 716–723. MR0423716

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BELLONE, E., HUGHES, J. P. and GUTTORP, P. (2000). A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Clim. Res.* **15** 1–12.

BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A bivariate space–time downscaler under space and time misalignment. *Ann. Appl. Stat.* **4** 1942–1975. MR2829942

CAREY-SMITH, T., SANSOM, J. and THOMSON, P. (2014). A hidden seasonal switching model for multisite daily rainfall. *Water Resour. Res.* **50** 257–272.

CHALLINOR, A. J., EWERT, F., ARNOLD, S., SIMELTON, E. and FRASER, E. (2009). Crops and climate change: Progress, trends, and challenges in simulating impacts and informing adaptation. *J. Exp. Bot.* **60** 2775–2789.

CHARLES, S. P., BATES, B. C. and HUGHES, J. P. (1999). A spatiotemporal model for downscaling precipitation occurrence and amounts. *J. Geophys. Res.* **104** 31657–31669.

CHARLES, S. P., BATES, B. C., SMITH, I. N. and HUGHES, J. P. (2004). Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrol. Process.* **18** 1373–1394.

CHIB, S. and GREENBURG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.

COX, D. R. (1970). *The Analysis of Binary Data*. Methuen & Co., Ltd., London. MR0282453

DEMPSTER, A. P. (1997). The direct use of likelihood for significance testing. *Stat. Comput.* **7** 247–252.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. R. (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38. MR0501537

DIEBOLD, F. X. and LEE, J. H. (1994). Regime switching with time-varying transition probabilities. In *Nonstationary Time Series Analysis and Cointegrations* (C. W. J. Granger and G. Mixon, eds.) 283–302. Oxford Univ. Press, London.

FILARDO, A. J. and GORDON, S. F. (1998). Business cycle durations. *J. Econometrics* **85** 99–123.

FORNEY, G. D. JR. (1973). The Viterbi algorithm. *Proc. IEEE* **61** 268–278. MR0439384

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal.* **15** 183–202. MR1263889

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Science & Business Media, Berlin.

FRÜHWIRTH-SCHNATTER, S. and FRÜHWIRTH, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Comput. Statist. Data Anal.* **51** 3509–3528.

FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45.

FURRER, E. M. and KATZ, R. W. (2007). Generalized linear modeling approach to stochastic weather generators. *Clim. Res.* **34** 129–144.

GADGIL, S. (2003). The Indian monsoon and its variability. *Annu. Rev. Earth Planet. Sci.* **31** 429–467.

GERMAIN, S. (2010). Bayesian spatio-temporal modelling of rainfall through non-homogenous hidden Markov models. Ph.D. thesis, Newcastle University, Newcastle, UK.

GERSHUNOV, A., SCHNEIDER, N. and BARNET, T. (2001). Low-frequency modulation of the ENSO-Indian monsoon rainfall relationship: Signal or noise? *J. Climate* **14** 2486–2492.

GHIL, M. and ROBERTSON, A. W. (2002). "Waves" vs. "particles" in the atmosphere's phase space: A pathway to long-range forecasting? *Proc. Natl. Acad. Sci. USA* **99** 2493–2500.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

GREENE, A. M., ROBERTSON, A. W. and KIRSHNER, S. (2008). Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time-scales using a hidden Markov model. *Q. J. R. Meteorol. Soc.* **134** 875–887.

GREENE, A. M., ROBERTSON, A. W., SMYTH, P. and TRIGLIA, S. (2011). Downscaling projectios of the Indian monsoon rainfall using a non-homogeneous hidden Markov model. *Q. J. R. Meteorol. Soc.* **137** 347–359.

HANSEN, J. W., CHALLINOR, A., INES, A., WHEELER, T. and MORON, V. (2006). Translating climate forecasts into agricultural terms: Advances and challenges. *Clim. Res.* **33** 27–41.

HAY, L. E., MCCABE, G. J., WOLOCK, D. M. and AYERS, M. A. (1991). Simulation of precipitation by weather type analysis. *Water Resour. Res.* **27** 493–501.

HEAPS, S. E., BOYS, R. J. and FARROW, M. (2015). Bayesian modelling of rainfall data by using non-homogeneous hidden Markov models and latent Gaussian variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 543–568.

HOLMES, C. C. and HELD, L. (2006a). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.* **1** 145–168. MR2227368

HOLMES, C. and HELD, L. (2006b). Response to van der Lans. *Bayesian Anal.* **6** 357–358.

HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2016). A Bayesian hidden Markov model of daily precipitation over South and East Asia. *J. Hydrometeorol.* **17** 3–25.

HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2017). Supplement to "Bayesian nonhomogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling." DOI:10.1214/16-AOAS1009SUPP.

HOOTEN, M. B. and WIKLE, C. K. (2010). Statistical agent-based models for discrete spatio-temporal systems. *J. Amer. Statist. Assoc.* **105** 236–248. MR2757201

HUGHES, J. P. and GUTTORP, P. (1994). A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.* **30** 1535–1546.

HUGHES, J. P., GUTTORP, P. and CHARLES, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **48** 15–30.

IMAI, K. and VAN DYK, D. A. (2005). MNP: R package for fitting the multinomial probit model. *J. Stat. Softw.* **14** 1–32.

IMMERZEEL, W. W., VAN BEEK, L. P. H. and BIERKENS, M. F. P. (2010). Climate change will affect the Asian water towers. *Science* **328** 1382–1385.

JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo and the label switching problem in Bayesian mixture modelling. *J. Statist. Plann. Inference* **20** 2305–2315.

JOHNDROW, J. E., LUM, K. and DUNSON, D. (2013). Diagonal orthant multinomial probit models. *J. Mach. Learn. Res. Workshop Conf. Proc.* **31** 29–38.

JOSEPH, P. V., GOKULAPALAN, B., NAIR, A. and WILSON, S. S. (2013). Variability of summer monsoon rainfall in India on inter-annual and decadal time scales. *Atmos. Ocean. Sci. Lett.* **6** 398–403.

JURAFSKY, D. and MARTIN, J. H. (2014). *Speech and Language Processing*. Prentice Hall, New York.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402

KATZ, R. and PARLANGE, M. (1995). Generalization of chain-dependent processes: Application to hourly precipitation. *Water Resour. Res.* **31** 1331–1341.

KIM, C.-J., PIGER, J. and STARTZ, R. (2008). Estimation of Markov regime-switching regression models with endogenous switching. *J. Econometrics* **143** 263–273. MR2423067

KIRSHNER, S. (2010). Modeling of multivariate time series using hidden Markov models. Ph.D. thesis, University of California, Irvine.

KIRSHNER, S., SMYTH, P. and ROBERTSON, A. W. (2004). Conditional Chow-Liu tree structures for modeling discrete-valued vector time series. In *Proc. 20th Conf. UAI* 317–324.

LAU, K.-M. and CHAN, P. H. (1986). Aspects of the 40–50 day oscillation during the northern summer as inferred from outgoing longwave radiation. *Mon. Weather Rev.* **114** 1354–1367.

LEE, J. Y., WANG, B., WHEELER, M. C., FU, X., WALISER, D. E. and KANG, I. S. (2013). Real-time multivariate indices for the boreal summer intraseasonal oscillation over the Asian summer monsoon region. *Clim. Dyn.* **40** 493–509.

MACDONALD, I. L. and ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. *Monographs on Statistics and Applied Probability* **70**. Chapman & Hall, London. MR1692202

MAMON, R. S. and ELLIOTT, R. J., eds. (2007). *Hidden Markov Models in Finance*. *International Series in Operations Research & Management Science* **104**. Springer, New York. MR2407726

MARAUN, D., WETTERHALL, F., IRESON, A. M., CHANDLER, R. E., KENDON, E. J., WIDMANN, M., BRIENEN, S., RUST, H. W., SAUTER, T., THEMESSL, M. et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **48** 1–34.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

MCCULLOCH, R., POLSON, N. G. and ROSSI, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econometrics* **99** 173–193.

MELIGKOTSIDOU, L. and DELLAPORTAS, P. (2011). Forecasting with non-homogeneous hidden Markov models. *Stat. Comput.* **21** 439–449.

MORON, V., ROBERTSON, A. W. and GHIL, M. (2012). Impact of the modulated annual cycle and intraseasonal oscillation on daily-to-interannual rainfall variability across monsoonal India. *Clim. Dyn.* **38** 2409–2435.

NEAL, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report No. 9702, Department of Statistics, University of Toronto.

O'BRIEN, S. M. and DUNSON, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics* **60** 739–746. MR2089450

PAAP, R. and FRANCES, P. H. (2000). A dynamic multinomial probit model for brand choices with different short-run effects of marketing mix variables. *J. Appl. Econometrics* **15** 717–744.

PAROLI, R. and SPEZIA, L. (2008). Bayesian inference in non-homogeneous Markov mixtures of periodic autoregressions with state-dependent exogenous variables. *Comput. Statist. Data Anal.* **52** 2311–2330.

PATTERSON, T. A., PARTON, A., LANGROCK, R., BLACKWELL, P. G., THOMAS, L. and KING, R. (2016). Statistical modelling of animal movement: A myopic review and a discussion of good practice. Available at http://arxiv.org/abs/0901.4804.

PIANI, C., WEEDON, G. P., BEST, M., GOMES, S. M., VITERBO, P., HAGEMANN, S. and HAERTER, J. O. (2010). Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. *J. Hydrol.* **395** 199–215.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

RAJAGOPALAN, B., LALL, U. and TARBOTON, D. G. (1996). Nonhomogeneous Markov model for daily precipitation. *J. Hydrol. Eng.* **1** 33–40.

RAPHAEL, C. (1999). Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Trans. Pattern Anal. Mach. Intell.* **21** 360–370.

RIIHIMAKI, J., JYLANKI, P. and VEHTARI, A. (2013). Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *J. Mach. Learn. Res.* **14** 75–109.

ROBERT, C. P., RYDÉN, T. and TITTERINGTON, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 57–75.

ROBERTSON, A. W. (2009). Seasonal predictability of daily rainfall statistics over indramayu district, Indonesia. *Int. J. Climatol.* **29** 1449–1462.

RYDÉN, T. (2008). EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Anal.* **3** 659–688. MR2469793

SCHWARZ, G. E. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SCOTT, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *J. Amer. Statist. Assoc.* **97** 337–351.

SCOTT, S. L. (2011). Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Statist. Papers* **52** 87–109. MR2777069

SHUKLA, J. and PAOLINO, D. A. (1983). The southern oscillation and long-range forecasting of the summer monsoon rainfall over India. *Mon. Weather Rev.* **111** 1830–1837.

SIEPEL, A. and HAUSSLER, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* **11** 413–428.

SMITH, T. M., REYNOLDS, R. W., PETERSON, T. C. and LAWRIMORE, J. (2008). Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). *J. Climate* **21** 2283–2296.

SPEZIA, L. (2009). Reversible jump and the label switching problem in hidden Markov models. *Statist. Sci.* **139** 50–67.

SPEZIA, L., COOKSLEY, S. L., BREWER, M. J., DONNELLY, D. and TREE, A. (2014). Modelling species abundance in a river by Negative Binomial hidden Markov models. *Comput. Statist. Data Anal.* **71** 599–614.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measure of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639.

STERN, R. D. and COE, R. (1984). A model fitting analysis of daily rainfall data. *J. Roy. Statist. Soc. Ser. A* **147** 1–34.

VERMEULEN, S. J., CHALLINOR, A. J., THORNTON, P. K., CAMPBELL, B. M., ERIYAGAMA, N., VERVOORT, J. M., KINYANGI, J., JARVIS, A., LÄDERACH, P., RAMIREZ-VILLEGAS, J. et al. (2013). Addressing uncertainty in adaptation planning for agriculture. *Proc. Natl. Acad. Sci. USA* **110** 8357–8362.

WANG, B. and FAN, Z. (1999). Choice of South Asian summer monsoon indices. *Bull. Am. Meteorol. Soc.* **80** 629–638.

WILKS, D. S. (1998). Multisite generalization of a daily stochastic precipitation generation model. *J. Hydrol.* **210** 178–191.

WILKS, D. S. (1999a). Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agric. For. Meteorol.* **93** 153–170.

WILKS, D. S. (1999b). Multisite downscaling of daily precipitation with a stochastic weather generator. *Clim. Res.* **11** 125–136.

WILKS, D. S. and WILBY, R. L. (1999). The weather generation game: A review of stochastic weather models. *Prog. Phys. Geogr.* **23** 329–357.

WOOLHISER, D. A. and ROLDAN, J. (1982). Stochastic daily precipitation models 2. A comparison of distributions of amounts. *Water Resour. Res.* **18** 1461–1468.

YOO, J. H., ROBERTSON, A. W. and KANG, I.-S. (2010). Analysis of intraseasonal and interannual variability of the Asian summer monsoon using a hidden Markov model. *J. Climate* **23** 5498–5516.

ZHANG, X., BOSCARDIN, W. J. and BELIN, T. R. (2008). Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Comput. Statist. Data Anal.* **52** 3697–3708. MR2427374

ZHANG, Y., WALLACE, J. M. and BATTISTI, D. S. (1997). ENSO-like interdecadal variability: 1900–93. *J. Climate* **10** 1004–1020.

ZUCCHINI, W. and GUTTORP, P. (1991). A hidden Markov model for space–time precipitation. *Water Resour. Res.* **27** 1917–1923.

ZUCCHINI, W., MACDONALD, I. and LANGROCK, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall, Boca Raton.

T. HOLSCLAW
P. SMYTH
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF CALIFORNIA, IRVINE
IRVINE, CALIFORNIA 92697
USA
E-MAIL: tholscla@ams.ucsc.edu
smyth@ics.uci.edu

A. M. GREENE
A. W. ROBERTSON
INTERNATIONAL RESEARCH INSTITUTE
FOR CLIMATE AND SOCIETY
THE EARTH INSTITUTE AT COLUMBIA
UNIVERSITY
230 MONELL, 61 ROUTE 9W, PO BOX 1000
PALISADES, NEW YORK 10964-8000
USA
E-MAIL: amg@iri.columbia.edu
awr@iri.columbia.edu