Nonlin. Processes Geophys., 21, 939–953, 2014
www.nonlin-processes-geophys.net/21/939/2014/
doi:10.5194/npg-21-939-2014

Nonlinear Processes in Geophysics
Open Access

Logit-normal mixed model for Indian monsoon precipitation

L. R. Dietz and S. Chatterjee

University of Minnesota, School of Statistics, Minneapolis, Minnesota, USA

Correspondence to: L. R. Dietz (diet0146@umn.edu)

Received: 8 February 2014 – Published in Nonlin. Processes Geophys. Discuss.: 13 March 2014
Revised: 17 July 2014 – Accepted: 14 August 2014 – Published: 12 September 2014

Abstract. Describing the nature and variability of Indian monsoon precipitation is a topic of much debate in the current literature. We suggest the use of a *generalized linear mixed model* (GLMM), specifically, the logit-normal mixed model, to describe the underlying structure of this complex climatic event. Four GLMM algorithms are described and simulations are performed to vet these algorithms before applying them to the Indian precipitation data. The logit-normal model was applied to light, moderate, and extreme rainfall. Findings indicated that physical constructs were preserved by the models, and random effects were significant in many cases. We also found GLMM estimation methods were sensitive to tuning parameters and assumptions and therefore, recommend use of multiple methods in applications. This work provides a novel use of GLMM and promotes its addition to the gamut of tools for analysis in studying climate phenomena.

1 Introduction

Explanation of Indian monsoon precipitation has been a challenging problem in physics as well as data analysis. In this paper, we focus on statistical analysis of the summer monsoon precipitation data, to provide insight symbiotic with deterministic physics modeling. Previous statistical analysis studies regarding precipitation in Indian monsoons have explored two main areas – identifying methodology of data analysis and covariate selection.

The establishment of appropriate statistical methodology for explanation and prediction of precipitation, while simultaneously capturing underlying variability, is paramount. These methods are used in identification of trends for prediction, however, trends tend to be inconsistent across studies and may relate to linked variability on different temporal and spatial scales as noted by Turner and Annamalai (2012).

For instance, Goswami et al. (2006) used daily central Indian rainfall and found rising trends in frequency and magnitude of extreme rain events along with decreasing light and moderate rainfall. While validating their 2006 study, Ghosh et al. (2012) indicated increasing spatial variability in observed Indian rainfall extremes. They also found that moderate rainfall increased in central India despite a decreasing trend in occurrence of moderate rainfall. For high and extremely high rainfall, they noted a few locations experienced a significant upward or downward trend, however, most grid boxes showed a lack of trend.

A similar study conducted by Ghosh et al. (2009) used a finer spatial scale and indicated a mixture of increases and decreases of extreme rainfall events dependent on location. An increasing trend in exceedances of 99th (extreme) percentile daily rainfall was discovered by Krishnamurty et al. (2009). On the other hand, they stated many parts of India exhibited a decreasing trend for exceedances of the 90th (moderate to extreme) percentile. Increases in the frequency of both light and moderate to extreme rainfall events were observed in Singh et al. (2014), along with decreasing probability of regional rainfall events and higher variability in the intensity of these events.

These studies utilized parametric – regression, extreme value theory, time series methods – and nonparametric statistical techniques, yet their lack of unanimity suggests important properties of the Indian monsoon remain partially misunderstood.

In view of the above, we propose adding the *generalized linear mixed model* (GLMM) as a potential framework for analysis of Indian monsoon precipitation data. A GLMM is a broader framework compared to the standard (linear, log-linear, logistic, or other) regression in that there are *random*

effects involved. This implies part of the signal is random, and changes from one set of circumstances to another. In the current context, a GLMM may be suitable for capturing local, instantaneous variability. Such local variability may arise from cloud and other physical micro-properties. When there is no such local variability, an appropriate variance component in the GLMM would be zero, thus, recovering the true underlying "fixed-effects" regression model.

The second principal focus of literature has been identifying relevant covariates for study of Indian monsoon precipitation. Certain oscillations are commonly useful predictors for precipitation. For instance, the synoptic activity index (SAI) developed in Ajayamohan et al. (2008) correlated strongly with frequency of extreme rainfall. The Indian Ocean dipole (IOD) studied in Rajeevan et al. (2008) was shown to modulate inter-annual, inter-decadal and long-term trends of extreme rainfall events. Most commonly, the El Niño-Southern Oscillation (ENSO) (Kumar et al., 1999; Li and Yanai, 1996; Prell and Kutzback, 1992; Turner and Annamalai, 2012) is cited as a driver of the monsoon.

Several other climatic predictors of monsoons have been proposed in the literature including Himalayan/Eurasian snow extent (Kumar et al., 1999), Pacific trade winds (Li and Yanai, 1996), atmospheric CO_2 concentration (Prell and Kutzback, 1992), and tropospheric temperature difference (Xavier et al., 2007). Unfortunately, none have been conclusively attributed for the monsoon rainfall which suggests an intricate relationship between some or all of these factors.

Because explicit attribution to covariates may not be possible, GLMM is a logical model for Indian monsoon precipitation. It allows underlying randomness to drive observed data in a particular hierarchy while still accounting for hypothesized drivers of rainfall.

This paper provides an introduction on extending GLMMs to climate applications. Three paradigms of estimation – approximate likelihood, method of moments, and Bayesian – were tested using four separate algorithmic implementations. The methods of estimating GLMMs were penalized quasi-likelihood, penalized iteratively reweighted least squares, method of simulated moments, and data cloning. The theory and limitations of these estimates are described in detail, then utilized in simulations to test the validity of the methodology.

Simulation findings showed that penalized quasi-likelihood was not accurate for the given application, thus, the three remaining methods were used to fit logit-normal models with random intercepts by weather station for Indian summer rainfall data in light, moderate, and extreme rainfall classifications. Maximum temperature and elevation were consistently significant in the models aligning with the physics of precipitation. ΔTT – tropospheric temperature difference – was also significant for many of the models. The most meaningful finding was a random effect by weather station was non-negligible in many of the models. This provides further credibility to the methodology in applications to climate. Overall, we feel GLMMs could be a significant addition to data analytics in climate applications.

The rest of the paper is organized as follows. Section 2 gives a short background on GLMMs and in particular, elucidates the logit-normal model. The theory of the chosen estimation methods for GLMMs are discussed in Sect. 3. Section 4 furnishes the results of several simulations using these existing methods. Section 5 applies these methods to monsoon precipitation data from India. Finally, Section 6 presents conclusions and future work in this area.

2 Overview of generalized linear mixed models

2.1 Exponential families

Before discussing GLMMs, we provide preliminaries on the key component use of an exponential family for the observed data. An exponential family probability mass/density function (pmf/pdf) has several unique properties conducive to modeling. For further discussion of these properties, refer to Ch. 2 of Keener (2010). For simplicity, consider a univariate random variable Y distributed as an exponential family. The canonical form of the pmf/pdf then can be written as follows:

$$f(y|\eta) = \exp\left\{\frac{y\cdot\eta - c(\eta)}{a(\phi)} - r(y,\phi)\right\}. \quad (1)$$

Note $a(\cdot)$ is a function of a dispersion parameter ϕ, $r(\cdot,\cdot)$ is a function of data and the dispersion parameter, and $c(\cdot)$ is a function of parameters and is known as the cumulant function. The statistic y is complete and sufficient; it is known as the canonical statistic with corresponding canonical parameter η. An exponential family can be written in a more general fashion compared to Eq. (1), but will not be discussed here for simplicity.

2.2 Model description

A GLMM is a probability model with a hierarchical structure. Given the latent unobservable second layer, known as random effects, the top layer has a pdf/pmf following an exponential family distribution. Assume the observed data are independent conditional on the random effects, and that we have $i = 1, \ldots, N$ observations. Define the ith response as Y_i.

Then, we can write a GLMM as

Level 1:

$$Y_i|\boldsymbol{u} \overset{\text{ind.}}{\sim} f(y_i|\boldsymbol{u}) = \exp\left\{\frac{y_i\cdot\eta_i - b(\eta_i)}{a(\phi)} - r(y_i,\phi)\right\}, \quad (2)$$

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{u}, \quad (3)$$

Level 2:

$$\boldsymbol{U} \sim \mathcal{N}_M(\boldsymbol{0}, \Sigma). \quad (4)$$

The p-components of the vector $\boldsymbol{\beta}$ are called *fixed effects*. The random effects covariance Σ is a function of the q-dimensional $(\sigma_1, \ldots, \sigma_q)$ known as *variance components*.

Fixed covariates are represented by the $p \times 1$ vector $\boldsymbol{x}_i$. Random covariate vectors for the ith data point and rth variance component can be denoted by a $m_r \times 1$ vector $\boldsymbol{z}_{ir}$. We combine the vectors for each variance component to form the random covariate vector $\boldsymbol{z}_i = (z_{i1}^T, \ldots, z_{iq}^T)^T$ of length $M = \sum_{r=1}^{q} m_r$. The random effects vector, $\boldsymbol{U}$, follows an M-dimensional normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ.

In Eq. (2), $b(\cdot)$ is a function of only the canonical parameter η_i. The "linear" part of GLMM comes from the fact that η_i can be represented as a linear function of the fixed and random parameters. $r(\cdot, \cdot)$ is a function of the data and ϕ. As before, $a(\cdot)$ is a dispersion function.

To illustrate this form, we consider a commonly used version of this model, the logit-normal GLMM. The random-intercept form of this model is

$$\text{Level 1:} Y_i|\boldsymbol{u} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\theta_i), \tag{5}$$

$$\text{logit}(\theta_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i, \tag{6}$$

$$\text{Level 2:} U_i \overset{\text{ind.}}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{7}$$

Notice that in this model, $\boldsymbol{z}_i$ is a vector with a 1 in the ith position and 0's in all other positions.

Returning to the generic form of the GLMM, the assumption of conditional independence among observations implies the density of $\boldsymbol{Y}|\boldsymbol{u}$ is

$$f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{u}, \boldsymbol{\beta}), \tag{8}$$

and that the joint density of $(\mathbf{Y}, \mathbf{U})$ is

$$f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\beta}, \Sigma) = f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}) f(\boldsymbol{u}|\Sigma). \tag{9}$$

However, since random effects are unobserved, in order to utilize the observed data likelihood, one must find the marginal distribution with respect to the observed data $\mathbf{Y}$ only. The log-likelihood is then

$$\ell(\boldsymbol{\beta}, \Sigma|\boldsymbol{Y}) = \log \int f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\beta}, \Sigma) d\boldsymbol{u}. \tag{10}$$

This integral is rarely analytically tractable. Thus, maximum likelihood estimation (which is usually preferable when possible) is very difficult. Many methods for inference have been proposed. Variants of the most popular methods are examined in Sect. 3.

3 Methods for estimating in GLMM

3.1 Likelihood approximation methods

Both methods discussed in the following sections make use of a technique known as Laplace approximation to approximate an integral by a normal distribution (Tierney and Kadane, 1986). Then Eq. (10) can be written as follows:

$$\log \int f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}) f(\boldsymbol{u}|\Sigma) d\boldsymbol{u} = \tag{11}$$

$$\log \int \exp\{\log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}) + \log f(\boldsymbol{u}|\Sigma)\} d\boldsymbol{u}. \tag{12}$$

Let

$$h(\boldsymbol{u}) = \log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}) + \log f(\boldsymbol{u}|\Sigma). \tag{13}$$

Then, we can express the log likelihood as follows:

$$\ell(\boldsymbol{\beta}, \Sigma; \boldsymbol{y}) = \log \int e^{h(\boldsymbol{u})} d\boldsymbol{u}. \tag{14}$$

This expression can now be approximated. To use the approximation, one first needs the maximizer of the integrand. Let $\boldsymbol{u}_0$ be the maximizer of $e^{h(\boldsymbol{u})}$. Then a Taylor expansion around $\boldsymbol{u}_0$ yields the approximation to the log-likelihood,

$$\ell(\boldsymbol{\beta}, \Sigma; \boldsymbol{y}) \approx h(\boldsymbol{u}_0) + \frac{q}{2} \log 2\pi - \frac{1}{2} \log \left| -\frac{\partial^2 h(\boldsymbol{u})}{\partial \boldsymbol{u}, \partial \boldsymbol{u}^T} \right|. \tag{15}$$

3.1.1 Penalized quasi-likelihood

Penalized quasi-likelihood (PQL) was proposed by Breslow and Clayton (1993) to approximate the high-dimensional integral using Laplace approximation as a method for obtaining $\boldsymbol{u}_0$ and $\partial^2 h(\boldsymbol{u}) / \partial \boldsymbol{u}\, \partial \boldsymbol{u}^T$. Filling in the details of Eq. (13),

$$h(\boldsymbol{u}) = \log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}) - \frac{1}{2} \boldsymbol{u}^T \Sigma \boldsymbol{u} + \frac{q}{2} \log 2\pi - \frac{1}{2} \log |\Sigma|. \tag{16}$$

This equation is differentiated with respect to $\boldsymbol{u}$ and $\boldsymbol{\beta}$ respectively. Further approximations are made within the derivatives because Σ is also unknown. The approximate derivatives are used to form estimating equations for the mean parameters. For more detailed discussion of these approximations, please refer to McCulloch and Searle (2010). The same estimating equations arise from jointly maximizing

$$\log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}) - \frac{1}{2} \boldsymbol{u}^T \Sigma \boldsymbol{u}, \tag{17}$$

with respect to $\boldsymbol{u}$ and $\boldsymbol{\beta}$.

These equations are solved by using Fisher scoring as an iterated reweighted least squares (IRLS) problem. The quasi-likelihood, $\log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta})$, is optimized taking into account the penalty, $\frac{1}{2} \boldsymbol{u}^T \Sigma \boldsymbol{u}$. This penalty term has a shrinkage effect, i.e. forces values of $\boldsymbol{u}$ to be closer to zero.

Variance components in Σ are subsequently estimated using a restricted maximum likelihood approach. Further details on the estimation algorithm are found in Sect. 2 of Breslow and Clayton (1993).

The function in R which computes PQL estimates is `glmmPQL{MASS}`. PQL is reasonably accurate when data

are approximately normal and can be very fast. However, Lin and Breslow (1996) and others have criticized this method for its bias in highly non-normal data. It is especially bad in binomial data with a small sample size or true probabilities near zero or one. Reliance on the quadratic expansion of the log-likelihood is appropriate with normal random effects, yet it is very difficult to assess normality of these unobserved effects.

3.1.2 Penalized iteratively reweighted least squares

Another approach to likelihood approximation is presented by Bates (2010). The main difference from PQL is that it attempts to approximate the true likelihood rather than the quasi-likelihood.

To understand the approach, first, let $\boldsymbol{U} \sim \mathcal{N}_M(\boldsymbol{0},\ \Sigma)$. Consider the decomposition of the random effects covariance matrix $\Sigma = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$. Then, $\boldsymbol{U} = \boldsymbol{\Gamma}\boldsymbol{V}$ where $\boldsymbol{V} \sim \mathcal{N}_M(\boldsymbol{0},\ I_m)$. This implies that the canonical parameter in Eq. (3) can be written as

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{\Gamma}\boldsymbol{v}. \tag{18}$$

Substituting in $\boldsymbol{v}$ to Eq. (9), we note that $f(\boldsymbol{y},\ \boldsymbol{v})$ is proportional to $f(\boldsymbol{v}|\boldsymbol{y})$. Thus, $\boldsymbol{v}_0$ is found to maximize $f(\boldsymbol{v}|\boldsymbol{y})$.

The penalized iteratively reweighted least squares (PIRLS) algorithm is as follows.

1. Given starting values for $\boldsymbol{\beta}$, Σ, and $\boldsymbol{v}_0$, evaluate $\boldsymbol{\eta}$, $\mu_{Y|v}$, and $\mathrm{var}_{Y|v}$. Let $\boldsymbol{W} = \mathrm{diag}\ \mathrm{var}_{Y|v}^{-1}$.
2. Use a Gauss–Newton algorithm to solve

$$\mu_{V|y} = \arg\min_{v}\left(||\boldsymbol{W}^{1/2}\left(\boldsymbol{y} - \mu_{Y|v}\right)||^2 + ||\boldsymbol{v}||^2\right). \tag{19}$$

3. Update the weights, $\boldsymbol{W}$, and check for convergence. If not converged, go to step 2.

Once the conditional mode $\tilde{\boldsymbol{v}}$ is determined, a Laplace approximation to the deviance ($-2 \times$ log-likelihood) is evaluated at $\tilde{\boldsymbol{v}}$. This evaluation may alternatively be done by the Gauss–Hermite quadrature which is discussed further in Bolker et al. (2009). The function in R used to compute estimates is `glmer{lme4}`. This method can experience similar problems to PQL in cases where the random effects are non-normal. The Gauss–Hermite quadrature can allay some of these issues, but is only computationally feasible for small numbers of random effects.

3.2 Method of simulated moments

Jiang (1998) describes methodology known as the *method of simulated moments* (MSIM). The method first derives a set of sufficient statistics. Estimating equations are then obtained by equating sample moments of sufficient statistics to their expectations.

Referring to the model elucidated in Eqs. (5)–(7), let the dispersion function be $a(\phi) = \frac{w_i}{\phi}$ where w_i is a weight depending on the exponential family of the response. Let $\boldsymbol{\theta} = (\boldsymbol{\beta},\ \sigma_1,\ \ldots,\ \sigma_q)$. Restrict all elements of the $\boldsymbol{z}_i$ to be either 0 or 1. Represent

$$\boldsymbol{z}_i^T\boldsymbol{u} = \left(\boldsymbol{z}_{i1}^T\boldsymbol{u}_1, \ldots, \boldsymbol{z}_{iq}^T\boldsymbol{u}_q\right) \tag{20}$$

$$= \left(\sigma_1\boldsymbol{z}_{i1}^T\boldsymbol{v}_1, \ldots, \sigma_q\boldsymbol{z}_{iq}^T\boldsymbol{v}_q\right), \tag{21}$$

where $\boldsymbol{V}_r \sim \mathcal{N}_{m_r}(0,\ I_{m_r})$.

Then,

$$f(y_i|\boldsymbol{v}) = C(y_i, \boldsymbol{\theta}, \phi)^*$$
$$\exp\left\{\left(\sum_{i=1}^{N} w_i\boldsymbol{x}_i y_i\right)^T\frac{\boldsymbol{\beta}}{\phi} + \sum_{r=1}^{q}\frac{\sigma_r}{\phi}\sum_{i=1}^{N} w_i y_i \boldsymbol{z}_{ir}^T\boldsymbol{v}_r\right\},$$

where $C(\cdot,\ \cdot,\ \cdot)$ represents the other portion of the function.

This yields canonical parameters $(\boldsymbol{\beta}/\phi,\ \sigma_1/\phi,\ \ldots,\ \sigma_q/\phi)$ with corresponding sufficient statistics

$$\left(\left(\sum_{i=1}^{N} w_i\boldsymbol{x}_i y_i\right)^T, \sum_{i=1}^{N} w_i y_i z_{i1}, \ldots, \sum_{i=1}^{N} w_i y_i z_{iq}\right).$$

Estimating equations are derived as

$$\sum_{i=1}^{N} w_i\boldsymbol{x}_i y_i \overset{\text{set}}{=} \sum_{i=1}^{N} w_i\boldsymbol{x}_i\mathbb{E}_\theta(y_i) \tag{22}$$

$$\sum_{l=1}^{m_r}\left(\sum_{i=1}^{N} w_i z_{irl} y_i\right)^2 \overset{\text{set}}{=} \sum_{l=1}^{m_r}\mathbb{E}_\theta\left(\sum_{i=1}^{N} w_i z_{irl} y_i\right)^2. \tag{23}$$

Note that the expectations on the right hand side are functions of the parameters while the formulae on the left hand sides are functions of data only. Since the expectations are not available, they must be estimated by Monte Carlo simulation. The system of equations can then be solved for the parameters using the Newton–Raphson algorithm.

We implemented this method in a newly created R program. As shown in Jiang (1998), this method is consistent and is potentially computationally less intensive than a Markov Chain Monte Carlo (MCMC) method.

3.3 Data cloning

GLMM estimates can be produced in a traditional Bayesian framework; one must choose priors for the parameters of interest and calculate the posterior distribution by multiplying the prior densities by the likelihood, $L(\boldsymbol{\beta},\ \Sigma|\boldsymbol{Y})$, corresponding Eq. (10). One may then use MCMC to generate a dependent sample from the posterior distribution from which estimates can be derived based on strong laws.

Lele et al. (2010) derived a method called *data cloning* to be used in conjunction with MCMC. The algorithm can

Table 1. MSIM simulation results: $\mu = 2, \sigma^2 = 1$.

Par.	# of Sub.	Obs. per subject 2	10	50	200
μ	10	17.41 (4.38)	2.11 (0.07)	2.05 (0.03)	2.00 (0.01)
	50	2.08 (0.05)	1.98 (0.02)	2.02 (0.01)	2.00 (0.00)
	200	2.01 (0.03)	1.98 (0.02)	1.99 (0.01)	1.99 (0.00)
	1000	2.00 (0.04)	1.99 (0.01)	2.01 (0.01)	2.00 (0.00)
σ^2	10	741.99 (302.51)	1.71 (0.33)	1.16 (0.09)	0.97 (0.04)
	50	1.02 (0.10)	0.98 (0.05)	0.98 (0.02)	0.99 (0.01)
	200	0.87 (0.05)	0.97 (0.03)	0.98 (0.02)	0.99 (0.01)
	1000	0.92 (0.06)	0.99 (0.02)	1.00 (0.02)	1.00 (0.01)
Loss	10	2885.74 (1016.73)	4.32 (0.71)	1.14 (0.33)	0.09 (0.01)
	50	3.29 (0.58)	0.19 (0.02)	0.04 (0.01)	0.01 (0.00)
	200	0.33 (0.04)	0.07 (0.01)	0.02 (0.00)	0.00 (0.00)
	1000	0.50 (0.16)	0.05 (0.00)	0.02 (0.00)	0.00 (0.00)

be summarized in the following three steps. First, create a k-cloned data set $\boldsymbol{y}_k = (\boldsymbol{y}, \boldsymbol{y}, \ldots, \boldsymbol{y})$ where the observed data vector is repeated k times. Choose a prior distribution $\pi(\boldsymbol{\beta}, \Sigma)$. Then, the posterior distribution, $\pi_k(\boldsymbol{\beta}, \Sigma|\boldsymbol{Y})$, which corresponds to the k-cloned data is

$$\pi_k(\boldsymbol{\beta}, \Sigma|\boldsymbol{Y}) = \frac{[L(\boldsymbol{\beta}, \Sigma|\boldsymbol{Y})]^k \pi(\boldsymbol{\beta}, \Sigma)}{\int\limits_{(\boldsymbol{\beta}, \Sigma)} [L(\boldsymbol{\beta}, \Sigma|\boldsymbol{Y})]^k \pi(\boldsymbol{\beta}, \Sigma) d(\boldsymbol{\beta}, \Sigma)}. \quad (24)$$

Under regularity conditions as $k \to \infty$,

$$\pi_k(\boldsymbol{\beta}, \Sigma|\boldsymbol{Y}) \to \mathcal{N}(\widehat{(\boldsymbol{\beta}, \Sigma)}, \frac{1}{k} S^{-1}(\widehat{\boldsymbol{\beta}, \Sigma})), \quad (25)$$

where $\widehat{(\boldsymbol{\beta}, \Sigma)}$ is the maximum likelihood estimate (MLE) of $(\boldsymbol{\beta}, \Sigma)$ and S is the Fisher information matrix of the original data. Thus, large k means the posterior distribution is nearly degenerate at the MLE.

To generate a dependent sample from the posterior distribution $\pi_k(\boldsymbol{\beta}, \Sigma|\boldsymbol{Y})$, one may use an appropriate MCMC algorithm, such as a Gibbs sampler or Metropolis–Hastings algorithm.

Finally, one can calculate the sample means and variances of the components of $(\boldsymbol{\beta}, \Sigma)$. Estimates of MLEs for $(\boldsymbol{\beta}, \Sigma)$ correspond to these sample means and approximate variances of estimated MLEs correspond to k times the posterior variance of the original data as seen in Eq. (25).

This method was implemented using `dclone{dclone}` discussed in Solymos (2010) which relies on the well-reputed BUGS language for estimation of hierarchical models. The method is computationally intensive, and it may prove difficult to assess convergence as with any MCMC implementation.

Table 2. `dclone` simulation results: $\mu = 2, \sigma^2 = 1$.

Par.	# of Sub.	Obs. per subject 2	10	50	200
μ	10	13.18 (2.65)	2.12 (0.05)	2.03 (0.03)	2.02 (0.04)
	50	2.11 (0.07)	1.99 (0.02)	1.99 (0.02)	1.99 (0.01)
	200	2.05 (0.03)	2.02 (0.01)	1.99 (0.01)	2.01 (0.01)
	1000	2.01 (0.01)	2.00 (0.00)	1.99 (0.00)	2.00 (0.00)
σ^2	10	7.79 (1.79)	1.18 (0.11)	0.95 (0.06)	0.98 (0.05)
	50	1.67 (0.33)	1.00 (0.05)	0.99 (0.03)	1.00 (0.02)
	200	1.16 (0.08)	0.99 (0.02)	0.98 (0.01)	1.00 (0.01)
	1000	0.98 (0.04)	0.99 (0.01)	1.00 (0.01)	0.99 (0.00)
Loss	10	131.65 (86.2)	1.26 (0.14)	0.32 (0.04)	0.31 (0.05)
	50	1.58 (0.44)	0.19 (0.02)	0.06 (0.01)	0.04 (0.00)
	200	0.40 (0.06)	0.04 (0.00)	0.02 (0.00)	0.01 (0.00)
	1000	0.09 (0.01)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)

4 Logit-normal simulations

4.1 Simulation setup

For subject i where $i \in (1, \ldots, m)$ and observation j where $j \in (1, \ldots, n)$ and $(\mu, \sigma^2) = (2, 1)$, we simulated 100 different data sets from the model

$$\text{Level 1:} Y_{ij}|\boldsymbol{u} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\theta_{ij}\right) \quad (26)$$

$$\text{logit}\left(\theta_{ij}\right) = \eta_i = \mu + u_i \quad (27)$$

$$\text{Level 2:} U_i \overset{\text{ind.}}{\sim} \mathcal{N}\left(0, \sigma^2\right). \quad (28)$$

The number of subjects (m) was set at (10, 50, 200, 1000) and the observations per subject (n) was set at (2, 10, 50, 200). All methods were tested at each of these 16 settings. Means and standard errors over the 100 estimates at each setting were then calculated.

To quantitatively describe the estimation discrepancy between μ and $\hat{\mu}_{m,n}$, we used squared error loss,

$$Q\left(\hat{\mu}_{m,n}\right) = \left(\hat{\mu}_{m,n} - \mu\right)^2. \quad (29)$$

Because squared error loss is criticized for a bounded parameter space, we used Stein's loss,

$$S\left(\hat{\sigma}^2_{m,n}\right) = \frac{\hat{\sigma}^2_{m,n}}{\sigma^2} - 1 - \log \frac{\hat{\sigma}^2_{m,n}}{\sigma^2}, \quad (30)$$

to measure how well σ^2 was estimated by $\hat{\sigma}^2_{m,n}$. A combined loss was then calculated as

$$G\left(\hat{\mu}_{m,n}, \hat{\sigma}^2_{m,n}\right) = Q\left(\hat{\mu}_{m,n}\right) + S\left(\hat{\sigma}^2_{m,n}\right). \quad (31)$$

Ideally, as $m, n \to \infty$, $G(\hat{\mu}_{m,n}, \hat{\sigma}^2_{m,n}) \to 0$.

4.2 Simulation estimation analysis

The estimation results are displayed in Tables 1–4. All methods failed to reasonably estimate both μ and σ^2 in the smallest scenario with 10 subjects and two observations each.

Table 3. `glmer` simulation results: $\mu = 2, \sigma^2 = 1$.

Par.	# of Sub.	Obs. per subjects 2	10	50	200
μ	10	6.02 (0.70)	2.77 (0.18)	2.33 (0.09)	2.10 (0.02)
	50	2.18(0.09)	1.99(0.02)	2.02(0.01)	2.00 (0.00)
	200	2.03(0.04)	1.99(0.02)	1.99(0.01)	1.99 (0.00)
	1000	2.02(0.04)	1.99(0.01)	2.01(0.01)	2.00 (0.00)
σ^2	10	198.73 (81.39)	7.48 (1.36)	3.07 (0.76)	1.19 (0.04)
	50	1.66 (0.54)	0.95 (0.05)	0.94 (0.02)	0.94 (0.01)
	200	0.93 (0.06)	0.97 (0.03)	0.97 (0.01)	0.98 (0.01)
	1000	0.97 (0.05)	1.00 (0.02)	1.00 (0.01)	0.99 (0.00)
Loss	10	270.68 (84.19)	13.35 (1.91)	3.79 (1.12)	0.11 (0.02)
	50	4.96 (1.15)	0.19 (0.02)	0.04 (0.00)	0.01 (0.00)
	200	0.32 (0.04)	0.06 (0.01)	0.02 (0.00)	0.00 (0.00)
	1000	0.47 (0.20)	0.04 (0.00)	0.01 (0.00)	0.00 (0.00)

Table 4. `glmmPQL` simulation results: $\mu = 2, \sigma^2 = 1$.

Par.	# of Sub.	Obs. per subjects 2	10	50	200
μ	10	3.10 (0.17)	1.92 (0.16)	1.34 (0.14)	0.68 (0.08)
	50	1.83 (0.06)	1.61 (0.03)	1.60 (0.02)	1.54 (0.01)
	200	1.81 (0.04)	1.71 (0.02)	1.73 (0.01)	1.72 (0.01)
	1000	1.81 (0.04)	1.81 (0.02)	1.81 (0.01)	1.79 (0.01)
σ^2	10	1.71 (0.13)	1.26 (0.11)	0.81 (0.11)	0.26 (0.06)
	50	0.52 (0.06)	0.25 (0.04)	0.15 (0.04)	0.01 (0.01)
	200	0.51 (0.04)	0.54 (0.03)	0.67 (0.02)	0.68 (0.01)
	1000	0.48 (0.04)	0.72 (0.03)	0.75 (0.01)	0.74 (0.01)
Loss	10	6.04(0.50)	5.95(0.47)	7.80(0.50)	10.21(0.47)
	50	4.81(0.44)	5.61(0.38)	6.26(0.31)	7.89(0.12)
	200	3.40(0.48)	1.77(0.33)	0.24(0.06)	0.15(0.01)
	1000	3.95(0.47)	0.75(0.25)	0.10(0.01)	0.10(0.00)

This was expected because there are not enough replications within the subject to get a meaningful estimate of a variance by subject.

All other settings for MSIM, `dclone`, and `glmer` estimated μ within 2 standard errors. These methods also provided reasonable estimates of σ^2 for settings other than those with 10 subjects. The combination of the loss for the two estimates went to 0 quickly for all three methods. In general, estimation by these three methods were unbiased.

The method `glmmPQL` did not converge to the true values of (μ, σ^2) as evidenced by combined loss greater than 0 for all settings. Further, this method displayed an underestimating bias in both parameters. Also, this function in R could not produce estimates for some of the 100 data sets in each setting.

4.3 Simulation speeds

Subsequently, we tested 4 of the 16 simulation settings to determine computing speed of the estimation methods. The settings used were combinations of (50, 200) subjects with (10, 200) observations. The `system.time()` command in R was used to record times. Simulations were independently run on four computers, and each estimation method was tested in sequence in one R script on a single core. Computer specifications can be found in the Appendix.

We implemented MSIM in two ways for the speed test. In the intercept-only model Eqs. (26)–(28), it is possible to use a simple algorithm for estimation. However, a more general form of the algorithm is needed for problems including fixed covariates. This form relies on matrices and does not work with large data sets at this time. These methods are referred to *MSIM fast* and *MSIM slow*, respectively.

Results were similar for each of the four computers, therefore, only one of the sets of results are shown in Table 5. The results indicated that `glmmPQL` was fastest in the 50 subject cases and `glmer` was fastest in the 200 subject cases. These two likelihood methods were the fastest due to the nature of the approximations that they make. The Bayesian method, `dclone`, was slower at about 4 to 25 min to produce estimates. The simple algorithm of MSIM fast was faster than `dclone` and slightly slower than approximation methods taking 3 to 6 s per run. The MSIM slow method was much slower ranging from 1.5 min to nearly 4 h. The case with 200 subjects 200 observations could not be handled by this method because the size of matrices and vectors exceeded the storage capacity allowed by R.

Table 5. Total system time (in seconds) results for Nokomis.

Method	(# of subjects, obs. per subject) (50, 10)	(50, 200)	(200, 10)	(200, 200)
glmer	0.089	0.048	0.080	0.071
PQL	0.286	0.234	0.384	0.394
MSIM fast	2.576	3.419	2.479	2.483
Dclone	10.028	11.355	38.069	40.004
MSIM slow	94.729	9363.849	1069.468	–

4.4 Simulation conclusions

An ideal method would provide high quality estimates in a short amount of time. The simulation indicated trade-offs between speed and accuracy in some of the methods.

The `glmer` and `glmmPQL` methods were the fastest, but `glmmPQL` estimates were biased. Accurate estimates were produced by `dclone` despite being much slower than the approximation methods.

In the intercept-only implementation, MSIM provided fast, accurate estimation. However, it was much slower than other methods in its matrix version and failed when too many observations were used. It should be noted that tuning parameters within each of the methods, such as convergence criteria for MSIM or number of MCMC samples in `dclone` may impact computing time significantly.

Based on the output of these simulations, the `glmmPQL` method was not consistent. The other three methods – `glmer`, `dclone`, and MSIM – provided estimates with reasonable accuracy. Therefore, these estimation methods are used to fit models for Indian monsoon precipitation data.

5 Application in Indian monsoon precipitation

5.1 Data

Multiple data sets have been used to study Indian monsoon precipitation in the literature of different temporal and spatial granularity. However, the initial goal was to develop and test the methods on widely and freely available data sets for the purpose of understanding the usefulness of GLMM in this context. This led to our selection of the data sources described below.

We chose the National Climatic Data Center (NCDC)[1] in the National Oceanic and Atmospheric Administration (NOAA) to gather latitude (°), longitude (°), elevation (m), and daily minimum and maximum temperatures (°C). These data were collected from 1 January 1973 to 31 December 2013. Data were queried for all available Indian stations in the database. This data source was developed for a wide variety of potential applications, including climate analysis and monitoring studies that require data at a daily time resolution. Quality assurance checks are routinely applied to the full data set according to Menne et al. (2012).

We note this data had a large amount of missing observations, therefore, only stations with at least five observations were included in analysis. One year in particular, 1975, did not contain enough data to be included in the analysis. To elucidate this missingness, on 25 August 2012, there were 33 stations with missing (NA) values, 12 stations with precipitation of 0 mm, and 31 stations with greater than 0 mm precipitation. This implies several stations were not included in the data for this day and in general, stations included change over time. Figure 1 illuminates the rainfall on this date.

We also included several other covariates of interest. The first was tropospheric temperature difference (ΔTT); the air temperature averaged between the levels 600 and 200 hPa. The hypothesis that Indian ocean warming leads to reduction in ΔTT which in turn reduces monsoon circulation is noted in Xavier et al. (2007). Thus, the inclusion of this covariate in the models was relevant. Data were collected from the National Centers for Environmental Prediction (NCEP) Reanalysis site[2].

As stated in Wang (2006) and other literature, Indian rainfall is strongly associated with ENSO, and onset of discharge in Niño-3.4 region can lead to drought in India. The occurrences of precipitation extremes are thought to be fewer in drought years. The Niño-3.4 monthly anomaly series was gathered for inclusion in the models from the NCEP site sponsored by NOAA [3].

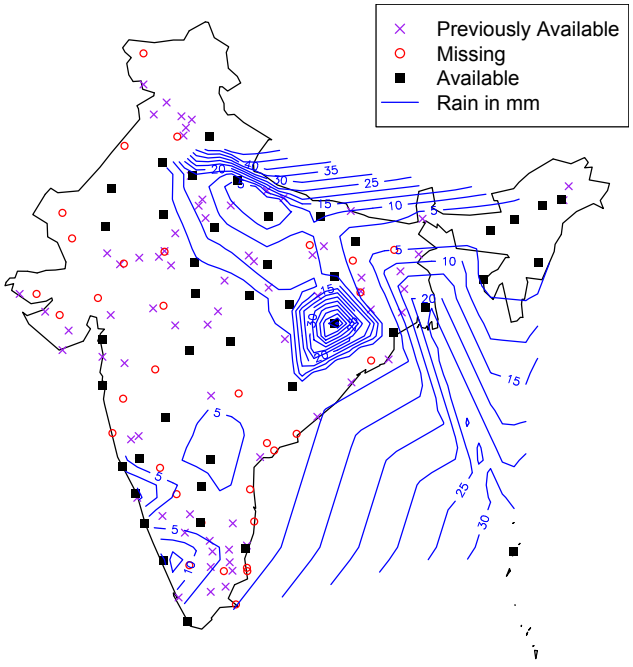

Figure 1. Observed Indian rainfall (in mm) on 25 August 2012, shown in contours. Markers indicate NCDC NOAA data status of individual stations.

The Indian Ocean Dipole (IOD) is an irregular oscillation occurring in the Indian Ocean. It is commonly measured by the Indian Dipole Mode Index (DMI) which takes the difference between sea surface temperature (SST) anomalies in the western and eastern Indian Ocean. Non-ENSO drought years are associated with DMI thus, this is a relevant covariate for inclusion in modeling. This index was only available for 1973–2010 models and data were procured from the Japan Agency for Marine-Earth Science and Technology (JAMSTEC) site[4].

We note that analysis of monsoon precipitation using thresholds was previously done in Krishnamurty et al. (2009). Rather than use a fixed threshold for the entirety of India, they utilized data derived percentile thresholds which changed depending on spatial location. However, their research was focused on trend analysis. Since we were able to include spatial covariates, we only consider fixed thresholds for the entire country found in Attri and Tyagi (2010). This report defined three categories of rainfall: light rainfall ($0 < x < 64.4\,\text{mm day}^{-1}$), moderate rainfall ($64.4 \leq x < 124.4\,\text{mm day}^{-1}$), and extreme rainfall ($\geq 124.4\,\text{mm day}^{-1}$).

[1] http://www.ncdc.noaa.gov/

[2] http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.derived.html

[3] http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/detrend.nino34.ascii.txt

[4] http://www.jamstec.go.jp/e/

All stations were were marked each day with indicators of these categories to be used in the modeling. Only observations considered to be within monsoon season were used. This conservatively included the time period from 1 May to 31 October (184 days) for each year. We fit models for each year (excluding 1975) from 1973–2013. To account for spatial variability, we fit a random intercept by weather station (WS) in the following logit-normal model:

$$\text{Level 1:} Y_{\text{WS,day}} | \boldsymbol{u} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\theta_{\text{WS, day}}\right), \quad (32)$$

$$\text{logit}\left(\theta_{\text{WS,day}}\right) = \boldsymbol{x}^T_{\text{WS,day}}\boldsymbol{\beta} + u_{\text{WS}}, \quad (33)$$

$$\text{Level 2:} U_{\text{WS}} \overset{\text{ind.}}{\sim} \mathcal{N}\left(0, \sigma^2_{\text{WS}}\right). \quad (34)$$

5.2 Results of GLMMs

To aid interpretation and provide a basis for comparison among models, we performed tests of significance for both fixed and random parameter estimates. We also give results from a goodness-of-fit test for each year's model.

In order to provide tests of significance for fixed effect coefficients, we propose the following procedure:

1. Run a generalized linear model (GLM) with all eligible fixed covariates.

2. Run a GLM with all eligible fixed covariates except the one we are testing.

3. Do a likelihood ratio test (LRT) to compare these models and get a p value from the asymptotic χ^2_1 distribution.

We recognize this method does not include the variance component and is thus, not the same model that we are proposing. The likelihood ratio test for GLM described above provides an idea about the relative important of various fixed effects covariates that may be influential for light, moderate or heavy precipitation. The above procedure may be supplemented by a multiple testing correction procedure, if needed. The details of this analysis is available from the authors. Also note that in this part of the analysis we did not include random effects, owing to a lack of viable and theoretically justifiable testing procedure when a random effect is present. Inclusion of random effects are likely to reduce variance attributed to noise, thus typically increasing significance levels.

We chose the GLM with all fixed covariates to provide a test of goodness-of-fit based on residual deviance being asymptotically χ^2. For details, refer to Faraway (2006). This compares the fitted model to the saturated model which contains one parameter for each observation. Failure to reject the null hypothesis of this test indicates a lack-of-fit.

Finally, a test of the variance component in the GLMM fit by `glmer` is done using a LRT with a nonstandard asymptotic distribution. Because our models have a single variance component, the asymptotic distribution for the LRT corresponds half of the p value obtained from the χ^2_1 distribution as noted by Zhang and Lin (2008). We only do this test for `glmer` since the other two methods do not use maximum likelihood, although it should be noted the likelihood produced by `glmer` is only approximate. Overall, we take a cautious view on the interpretation of these tests.

Table 6. This table indicates the percentage of significant p values at $\alpha = 0.05$ level for each of the 1973–2013 models. p values for fixed coefficients and goodness-of-fit test are from LRTs on GLM fits. p values for the variance components are from LRTs that compare GLM and `glmer` fits.

	Variable	Light	Moderate	Extreme
Fixed	DMI	84 %	30 %	14 %
	Niño34	68 %	20 %	13 %
	ΔTT	98 %	95 %	70 %
	elevation	95 %	98 %	95 %
	max. temp.	100 %	98 %	100 %
	min. temp.	75%	100%	40 %
	latitude	90 %	30 %	8 %
	longitude	90 %	33 %	55 %
Random	station	93 %	53 %	28 %
	lack-of-fit?	38 %	0 %	0 %

5.2.1 Discussion of rainfall models

Results of significance testing for each of the models can be found in Table 6. All covariates were significant in the light rainfall model in the majority of the years. However, 38 % of the years showed lack-of-fit based on the deviance test. The moderate and extreme rainfall models showed no lack of fit, but had far fewer significant covariates over the years.

Clearly, maximum daily temperature was important in all three levels of rainfall aligning with the Clausius–Clapeyron equation regarding water vapor capacity of the atmosphere. Minimum temperature was significant in most years for light and moderate rainfall, but was only significant in a minority of the extreme rainfall models.

Elevation was also significant in many years for all rainfall levels. This aligns with the physical explanations of warm moist air cooling at higher altitudes to produce precipitation.

Latitude and longitude were both significant in most light rainfall years. Moderate and extreme rainfall did not indicate latitude as significant in most years. Longitude was significant in just over half of the extreme models. Coefficient estimates for latitude indicated the probabilities of rain increasing going south to north. Longitude estimates were mostly negative indicating a decreasing probability of rainfall going west to east.

DMI was significant for the majority of light rainfall models; however, it was significant in very few of the moderate

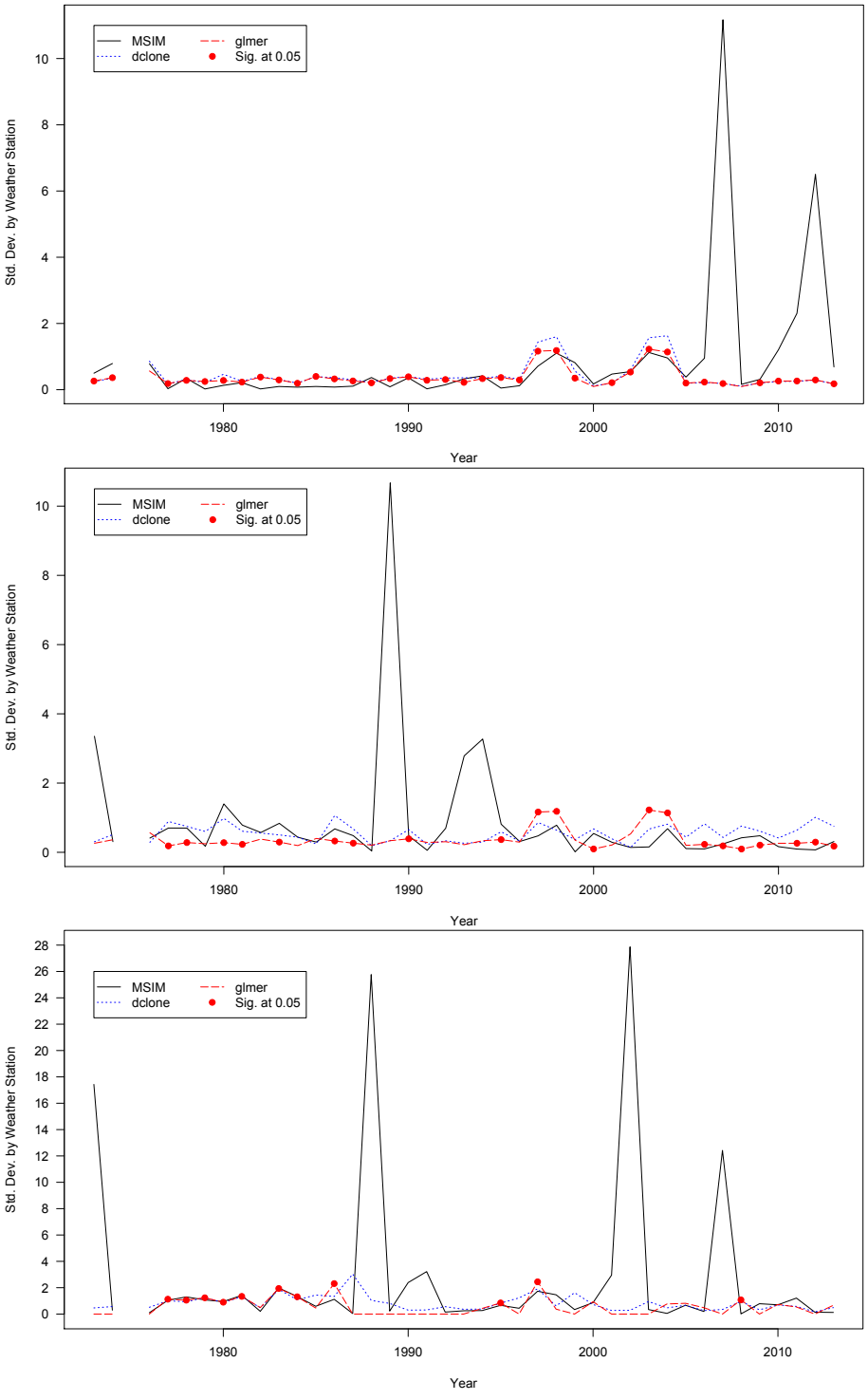

Figure 2. Top panel: weather station standard deviation estimates for logit-normal models with light Indian rainfall ($0 < x \leq 64.4\,\mathrm{mm\,day^{-1}}$) as the response from 1973–2013. Estimates over time indicate variability near 0, however, most of the `glmer` estimates are significant at the 0.05 level. Middle panel: weather station standard deviation estimates for logit-normal models with moderate Indian rainfall ($64.4 \leq x < 124.4\,\mathrm{mm\,day^{-1}}$) as the response from 1973–2013. Approximately half of the `glmer` estimates are significant at the 0.05 level. Bottom panel: weather station standard deviation estimates for logit-normal models with extreme Indian rainfall ($\geq 124.4\,\mathrm{mm\,day^{-1}}$) as the response from 1973–2013. Approximately one-quarter of the `glmer` estimates are significant at the 0.05 level

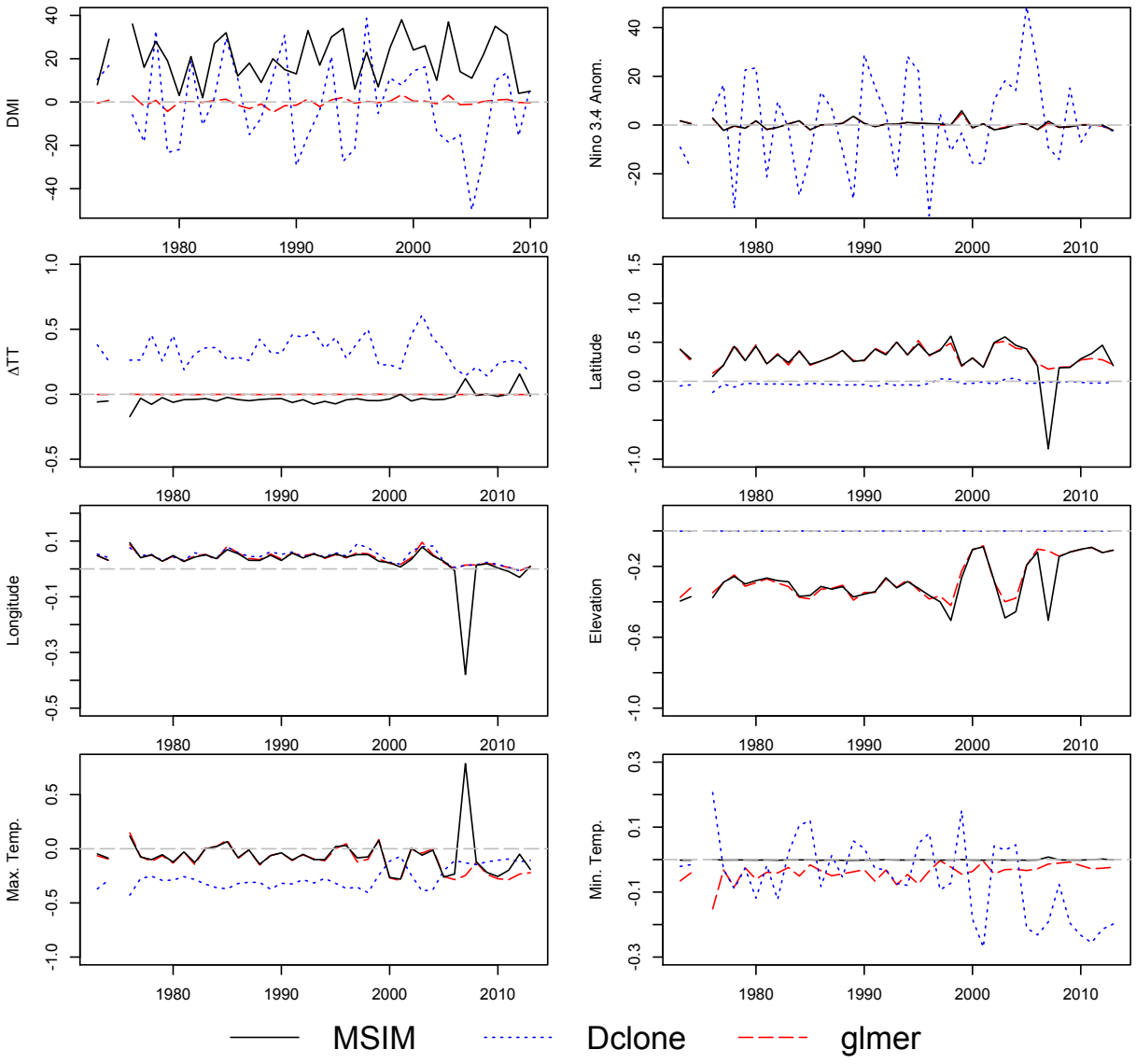

Figure 3. Fixed coefficient estimates for logit-normal models with light Indian rainfall ($0 < x < 64.4\,\text{mm day}^{-1}$) response from 1973–2013.

and extreme models. This corresponds to the DMI influence in non-ENSO drought years as hypothesized in previous literature.

ΔTT was significant in most years for all three rainfall levels corroborating the hypothesis that it is instrumental in monsoon circulation.

The Niño 3.4 anomaly index was significant in the majority of light rainfall models, but in less than 20 % of both moderate and extreme models. This could be related to the possible weakening of the relationship between ENSO and the Indian monsoon as noted in Chang et al. (2001) but may also be a function of the other covariates included in the modeling.

The station variance component was significant in nearly all of the light rainfall models. One can note from Fig. 2 (top panel), there is less variability in general in light rainfall even though most years are significant. The variance was significant in about half of the moderate and a quarter of the extreme rainfall models. As seen in Fig. 2, these models tended to have higher variability than light rainfall even though fewer years were significant. The variance component does provide additional explanation for the rainfall variability and thus, vets the methodology use in this application. This verifies the thesis of this paper: that a significant portion of the variability in any precipitation category is a random component that is distinguishable from random noise variability.

5.2.2 Estimation method performance

The coefficient estimates over the time period for all fixed effects at each level of rainfall are depicted in Figs. 3–5.

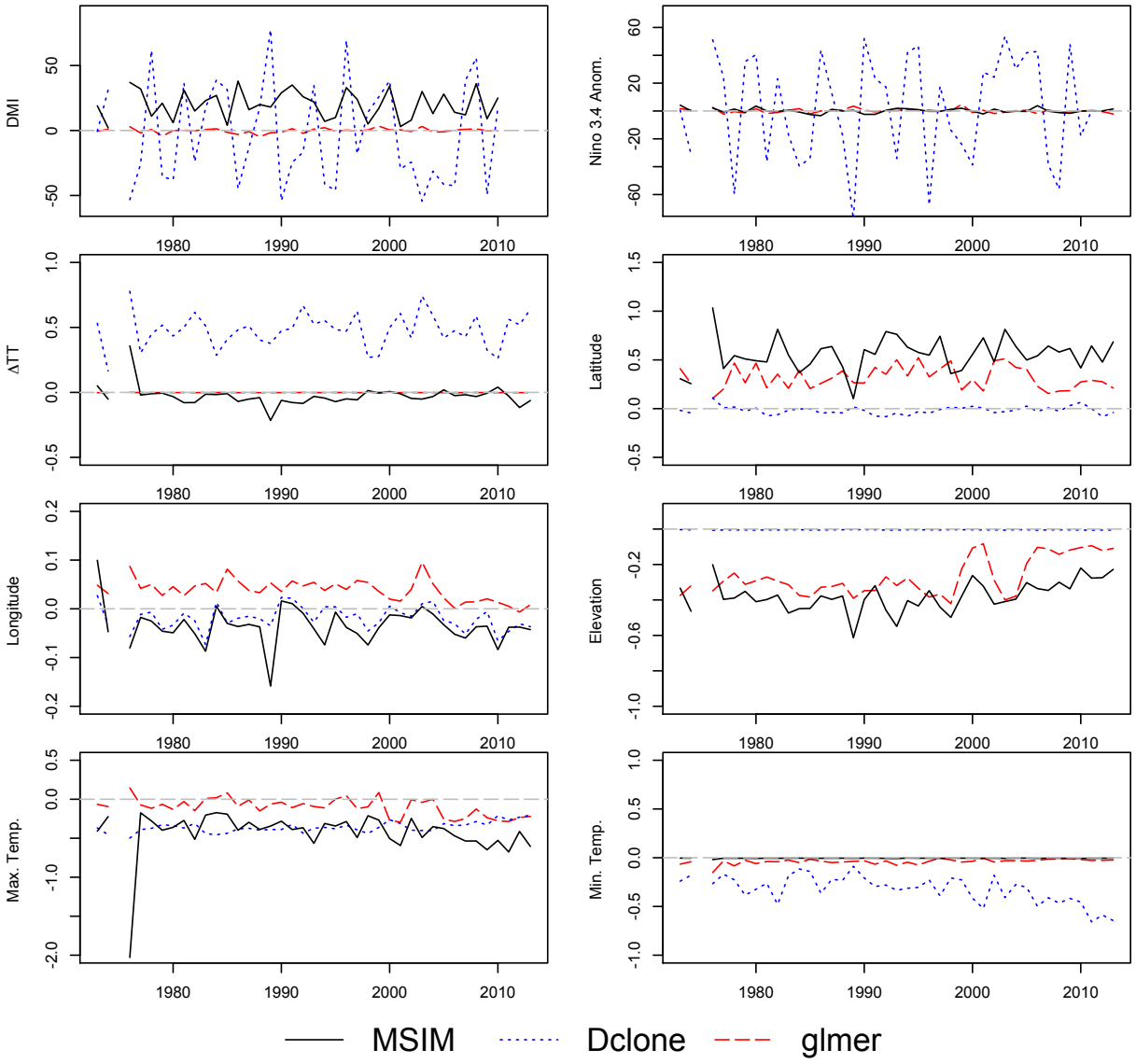

Figure 4. Fixed coefficient estimates for logit-normal models with moderate Indian rainfall ($64.4 \leq x < 124.4\,\mathrm{mm\,day^{-1}}$) response from 1973–2013.

The three estimation methods tended to produce different answers on at least a few of the coefficients in each of the models. The best agreement amongst all methods occurred in the estimates for maximum temperature (moderate, extreme), longitude (light, extreme), and the variance components (all). Estimates for `glmer` and MSIM tended to agree more often than either agreed with `dclone`. However, they were fairly different in magnitude for several covariates and did not always trend with each other. Light rainfall models displayed slightly more disagreement more than moderate and extreme rainfall.

Standard deviations from `dclone` estimates indicated that the algorithm had likely not converged for all parameters in the 10 000 samples taken from the posterior. As mentioned in Sect. 3.3, one of the issues with using this method is difficulty in assessing convergence. It would likely require a much larger sample to provide suitable answers in this application. Based on this, we would say the `dclone` results were mostly inconclusive in this application.

The outcome of this application indicated `glmer` and MSIM provided more reasonable estimates, however, a longer run of `dclone` may also be useful. The three methods are representative of distinct statistical paradigms of estimation including approximate likelihood, Bayesian, and method of moments. Each of the methods uses a different algorithm and different assumptions, thus, we recommend use of multiple methods when applying GLMM in an application.

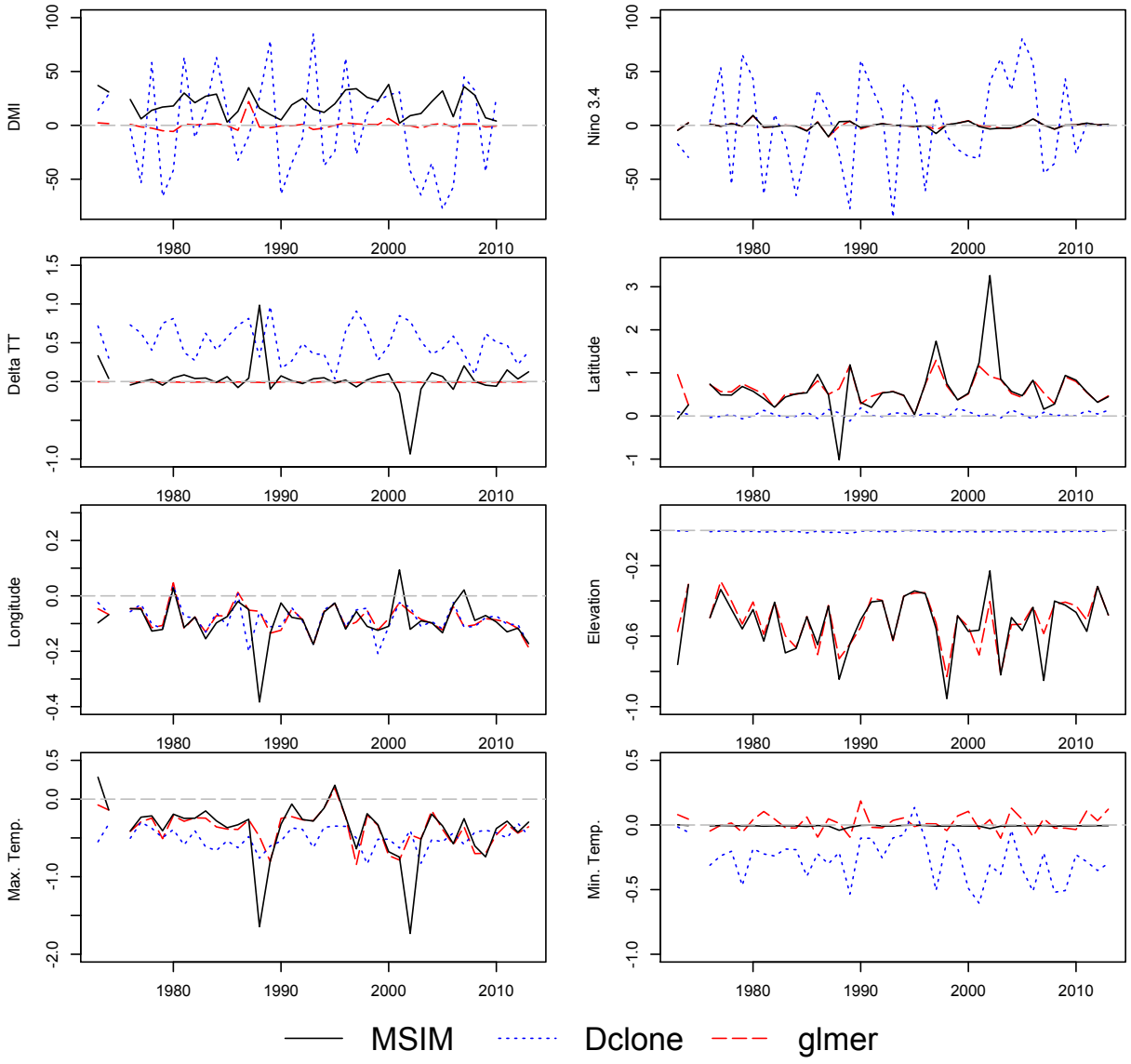

Figure 5. Fixed coefficient estimates for logit-normal models with extreme Indian rainfall ($\geq 124.4\,\mathrm{mm\,day^{-1}}$) response from 1973–2013.

6 Conclusions

Outcomes for Indian monsoon rainfall coincide with results in the Indian monsoon literature providing evidence of the usefulness of GLMM. Physical constructs are preserved by the models demonstrated by the importance of elevation, maximum temperature, and ΔTT in all levels of rainfall. Random effects are significant in several of the models indicating promise of modeling some of the unobservable and complicated interactions that underly climate patterns.

The GLMM methods explored in this article purposely included several styles of estimation including approximate likelihood, Bayesian, and method of moments type estimators. Each exhibited some drawbacks, thus, use of at least two out of three of the best methods, `glmer`, MSIM, and `dclone`, in the context of any application will help verify consistency of estimates. Use of multiple methods will provide researchers with higher confidence in results and will be more robust to limitations of any of the individual methods.

Since the relevance of GLMM in this context has been established, climate model output, such as that of CMIP5, will be explored to gain deeper intuition of the nature of this random effect. Further work on GLMMs may include studying other proposed drivers of Indian monsoons in their contributions to fixed or random effects. Additional random effects that include spatial correlation will be tested in future work. We also note that this model could be pursued in the future as a multinomial logit model.

It was suggested that Normalized Difference Vegetation Index (NDVI) may be a useful covariate for understanding

feedback of vegetation on precipitation. However, data coverage was limited, thus, it was not included in our models. It will be examined more closely in future modeling efforts.

Providing improvements to the GLMM estimation methods is another open research area. One limitation of GLMM, as presented here, is the reliance of modeling random effects as normal. Expanding the possible distributions of random effects to include extreme value distributions would be a breakthrough in mixed modeling. Providing more conclusive tests of significance for fixed and random effects is also important.

Appendix A: Additional specifications for simulations and applications

Computers used

- Assawa: 2010 Frontier i7 8-core Intel i7 940 (2.93 GHz) with 3 GB of RAM
- Geneva: 2011 Frontier i7 8-core Intel i7 950 (3.07 GHz) with 6 GB of RAM
- Nokomis: 2012 Optiplex 7010 8-core Intel i7-3770 (3.40 GHz) with 8 GB of RAM
- Tilde: 2013 Optiplex 7010 8-core Intel i7-3770 (3.40 GHz) with 8 GB of RAM

MSIM fast

- Number of Monte Carlo simulations: 100 000

MSIM slow

- Number of Monte Carlo simulations: 100
- Convergence criterion for Newton Raphson Method: Euclidean norm of change $\leq .01$

Dclone

- Clones: 5
- Prior for μ: $N(0, \frac{1}{0.0001})$
- Prior for $\frac{1}{\sigma^2}$: Gamma $(0.01, 0.01)$
- Adaptation length: 100
- Markov chain length after adaptation: 10 000

Acknowledgements. The comments of the two reviewers helped improve this paper considerably. We would also like to acknowledge Auroop Ganguly for input on data sets. Funding support of this research for L. R. Dietz is provided by means of the U. Minnesota Eva O. Miller Fellowship. This research is partially supported by the National Science Foundation under grant #IIS-1029711 and #SES-0851705, and by grants from the Institute on the Environment (IonE), and College of Liberal Arts, U. Minnesota.

Edited by: D. Wang
Reviewed by: S. Ghosh and C. K. B. Krishnamurthy

References

Ajayamohan, R., Merryfield, W., and Kharin, V.: Increasing trend of synoptic activity and its relationship with extreme rain events over central India, J. Climate, 23, 1004–1013, 2008.

Attri, S. D. and Tyagi, A.: Climate Profile of India, Tech. rep., Government of India, Ministry of Earth Sciences, India Meteorological Department, New Delhi, India, 2010.

Bates, D.: lme4: Mixed-effects modeling with R (Preprint), Springer, http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf (last access: 1 April 2014), 2010.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S.: Generalized linear mixed models: a practical guide for ecology and evolution, Trends Ecol. Evol., 24, 127–135, 2009.

Breslow, N. and Clayton, D.: Approximate Inference in Generalized Linear Mixed Models, J. Am. Stat. Assoc., 88, 9–25, 1993.

Chang, C.-P., Harr, P., and Ju, J.: Possible Roles of Atlantic Circulations on the Weakening Indian Monsoon Rainfall – ENSO Relationship, J. Atmos. Ocean. Tech., 14, 2376–2380, 2001.

Faraway, J. J.: Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, Texts in Statistical Science, Taylor and Francis, Boca Raton, Florida, USA, 2006.

Ghosh, S., Luniya, V., and Gupta, A.: Trend analysis of Indian summer monsoon rainfall at different spatial scales, Atmos. Sci. Lett., 10, 285–290, 2009.

Ghosh, S., Das, D., Kao, S.-C., and Ganguly, A. R.: Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes, Nat. Clim. Change, 2, 86–91, 2012.

Goswami, B., Venugopal, V., Sengupta, D., Madhusoodanan, M. S., and Xavier, P. K.: Increasing Trend of Extreme Rain Events Over India in a Warming Environment, Science, 314, 1442–1445, 2006.

Jiang, J.: Consistent Estimators in Generalized Linear Mixed Models, J. Am. Stat. Assoc., 93, 720–729, 1998.

Keener, R. W.: Theoretical Statistics: Topics for a Core Course, in: Springer Texts in Statistics, Springer, New York, USA, 2010.

Krishnamurty, C. K. B., Lall, U., and Kwon, H.-H.: Changing Frequency and Intensity of Rainfall Extremes over India from 1951 to 2003, J. Climate, 22, 4737–4746, 2009.

Kumar, K. K., Rajagopalan, B., and Cane, M. A.: On the Weakening Relationship Between the Indian Monsoon and ENSO, Science, 284, 2156–2159, 1999.

Lele, S. R., Nadeem, K., and Schumuland, B.: Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning, J. Am. Stat. Assoc., 105, 1617–1625, 2010.

Li, C. and Yanai, M.: The onset and interannual variability of the Asian summer monsoon in relation to land sea thermal contrast, J. Climate, 9, 358–375, 1996.

Lin, X. and Breslow, N. E.: Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion, J. Am. Stat. Assoc., 91, 1007–1016, 1996.

McCulloch, C. E. and Searle, S. R.: Generalized, Linear, and Mixed Models, in: Wiley Series in Probability and Statistics, 2nd Edn., John Wiley and Sons, Inc., Hoboken, New Jersey, 2010.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An Overview of the Global Historical Climatology Network-Daily Database, J. Atmos. Ocean. Tech., 29, 897–910, 2012.

Prell, W. L. and Kutzback, J. E.: Sensitivity of the indian monsoon to forcing parameters and implications for its evolution, Nature, 360, 647–652, 1992.

Rajeevan, M., Bhate, J., and Jaswal, A.: Analysis of variability and trends of extreme rainfall events over India using 104 years of gridded daily rainfall data, Geophys. Res. Lett., 35, 1–6, 2008.

Singh, D., Tsiang, M., Rajaratnam, B., and Diffenbaugh, N. S.: Observed changes in extreme wet and dry spells during the South Asian summer monsoon season, Nat. Clim. Change, 4, 456–461, 2014.

Solymos, P.: dclone: Data Cloning in R, R Journal, 2, 29–37, 2010.

Tierney, L. and Kadane, J. B.: Accurate Approximations for Posterior Moments and Marginal Densities, J. Am. Stat. Assoc., 81, 82–86, 1986.

Turner, A. G. and Annamalai, H.: Climate change and the South Asian summer monsoon, Nat. Clim. Change, 2, 587–595, 2012.

Wang, B.: The Asian Monsoon, Springer-Praxis Books in Environmental Sciences, Springer, Berlin, Germany, 2006.

Xavier, P. K., Marzina, C., and Goswami, B. N.: An objective definition of the Indian summer monsoon season and a new perspective on the ENSO–monsoon relationship, Q. J. Roy. Meteorol. Soc., 133, 749–764, 2007.

Zhang, D. and Lin, X.: Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and other Related Topics, in: Random Effect and Latent Variable Model Selection, vol. 192 of Lecture Notes in Statistics, edited by: Dunson, D., Springer, New York, 19–36, 2008.