Investigation of Precipitation Thresholds in the Indian Monsoon Using Logit-Normal Mixed Models

Lindsey R. Dietz and Snigdhansu Chatterjee

Abstract Previous literature showed the relevance of using logit-normal mixed models for understanding climate variable associations with Indian summer monsoon precipitation probabilities. We further this work by exploring fixed and station-based threshold definitions used to study monsoon precipitation intensity. Fixed thresholds are used to illuminate physical differences, such as the effect of temperature or tropospheric winds, as precipitation levels increase. Also, non-negligible station and year random effects indicate idiosyncrasies in probabilities of threshold exceedences by station and year. Station-based percentile thresholds are used to discuss predictions of threshold exceedances in particular stations where cyclical trends appear. Both types of thresholds provide meaningful information and expand the use of the logit-normal mixed model.

1 Logit-Normal Mixed Models in Indian Monsoon Precipitation

Generalized linear mixed models (GLMM) are commonly used in biostatistical and epidemiological settings, but are relatively new to climate data modeling. A proof-of-concept was done in Dietz and Chatterjee (2014) and indicated a logit-normal model was useful in understanding Indian summer monsoon precipitation. We extend this use of GLMM to examine other previously studied types of thresholds in precipitation data. Station-defined percentile thresholds were used in Krishnamurty et al (2009) and fixed level thresholds were used in Goswami et al (2006) to ex-

L. R. Dietz
University of Minnesota – Twin Cities, School of Statistics, 313 Ford Hall, 224 Church St SE, Minneapolis, MN 55455; e-mail: diet0146@umn.edu

S. Chatterjee
University of Minnesota – Twin Cities, School of Statistics, 313 Ford Hall, 224 Church St SE, Minneapolis, MN 55455; e-mail: chatterjee@stat.umn.edu

plore trends in monsoon rainfall intensity. Our study focuses on the inclusion of relevant covariates and uses both threshold definitions with distinct purposes. We use the fixed threshold model to elicit a physical interpretation across rainfall levels, and percentile-based thresholds for understanding local predicted probabilities of threshold exceedances and possible cycles in their occurrence.

Theory exposition for all models used within this study can be found in McCulloch and Searle (2010); information on estimation techniques available for GLMM can be found in Breslow and Clayton (1993), Jiang (1998), and Lele et al (2010).

Annual logit-normal models with a station random effect were used in Dietz and Chatterjee (2014). Rather than estimating different models for each year, we took a more robust approach and fit a single model for the entire time period, added additional relevant covariates, and kept the station random effect. We also tested a model with separate station and year random effects. The larger model is depicted in the following box.

Logit-Normal Mixed Model for Indian Monsoon Precipitation

Let station $i \in \{1,...,m\}$, day $j \in \{1,...,n_i\}$, and year $k \in \{1,...,K\}$. Given a threshold τ and precipitation event Z_{ijk}, let $Y_{ijk} = I(Z_{ijk} > \tau)$. Let $\mathbf{x}_{ijk}$ be a vector of covariates and $\mathbf{U}$ and $\mathbf{W}$ be vectors of random effects for station and year, respectively. Then,

$$\text{Level 1: } Y_{ijk}|\mathbf{U}=\mathbf{u},\mathbf{W}=\mathbf{w} \overset{\text{ind.}}{\sim} \text{Bernoulli}(\theta_{ijk}), \tag{1}$$

$$\text{logit}(\theta_{ijk}) = \mathbf{x}_{ijk}^{\mathrm{T}}\beta + u_i + w_k, \tag{2}$$

$$\text{Level 2: } U_i \overset{\text{ind.}}{\sim} \mathcal{N}(0,\sigma^2_{station}), W_k \overset{\text{ind.}}{\sim} \mathcal{N}(0,\sigma^2_{year}) \tag{3}$$

$$U_i \text{ independent of } W_k \text{ for all } (i,k). \tag{4}$$

To provide benchmark models to the GLMMs, we fit a generalized linear model (GLM) which does not take into account repeated measures by station or year and a generalized estimating equations (GEE) model with an auto-regressive lag 1 structure for repeated events within weather station. Model selection was not used within this study; instead, we selected scientifically relevant covariates to investigate based earlier literature.

Within the rest of the chapter, we provide discussion on the fixed and percentile-based threshold models. Section 2 provides an overview of the data and software methodology. Section 3 focuses on the interpretation of fixed threshold models in understanding covariates and variability at different threshold levels. Section 4 discusses the use of percentile-based threshold models to provide predicted probabilities on a local scale. Final commentary and future directions for this work are highlighted in Section 5.

2 Data Processing and Software

Daily data for station level covariates of minimum temperature, maximum temperature, elevation, latitude and longitude were collected from the National Climatic Data Center (NCDC)[1] in the National Oceanic and Atmospheric Administration (NOAA).

Data were collected from 1973 to 2013. Only observations considered to be within the summer monsoon season (1 June to 30 September) were used. Station-level data had a large amount of missing observations, therefore, only stations with at least 40% of days were included in the analysis. Two years in particular, 1975-1976, were also excluded from the analysis due to the high level of missingness. The processed data included a total of 36 weather stations.

Along with the NCDC data, reanalysis data (Kalnay et al (1996)) were collected. These data include tropospheric temperatures from 200 and 600 mb levels, u-winds from 200 and 850 mb levels, and v-winds from 200 and 850 mb levels[2]. Since these data[3] are gridded, they were aligned with the station closest in Euclidean distance by latitude and longitude. The wind variables were kept in their original form while the two temperatures were averaged to create tropospheric temperature difference (ΔTT) as suggested by Xavier et al (2007). All of these tropospheric variables affect the monsoon circulation, and are of physical importance for inclusion in the model.

A final covariate of interest was the Niño-3.4 anomaly series collected from the National Centers for Environmental Prediction (NCEP) Climate Prediction Center (CPC)[4]. This index is a measure of the sea surface temperature which is known to be an important global climate driver. Again, since these data are gridded, they were assigned to stations in the same method as the previous gridded covariates.

Analysis in this article was done using SAS/STAT® 9.3 for the Windows® operating system. Several approximate likelihood estimation methods were tested and produced similar results, thus, we used output from PROC GLIMMIX estimated by the residual subject-specific pseudo-likelihood (RSPL) method. GLM and GEE models were estimated using PROC GENMOD. Uncertainty estimates within this study correspond to the default methods in these procedures. GLMM approximate standard errors for fixed effects are obtained by use of the delta method on the predicted population averaged probability estimates; variance component standard errors are based on asymptotic theory. GLM estimates use asymptotic normal standard errors while GEE provides empirically based standard errors. Detailed information on these procedures can be found in SAS Institute Inc. (2011).

---

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

[2] Positive u-winds move west to east (westerlies); positive v-winds move south to north (southerlies).

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

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

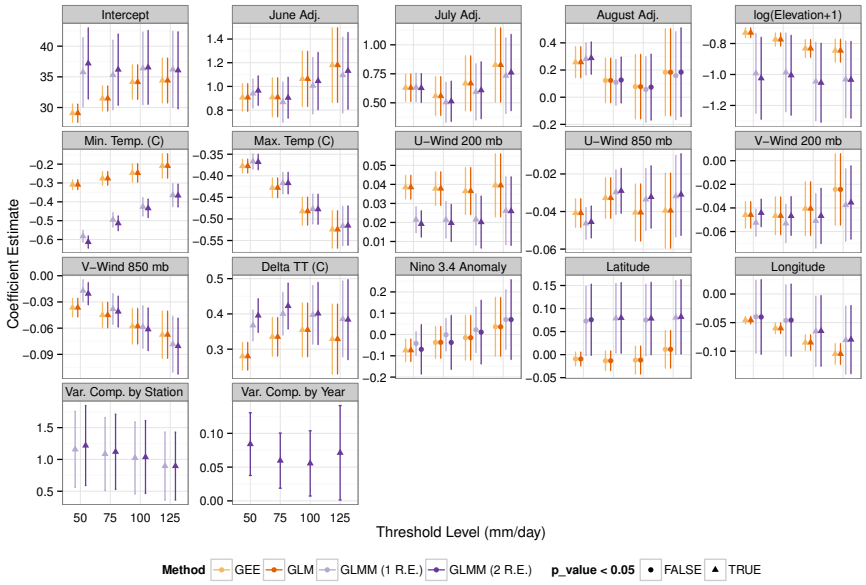

Fig. 1: Fixed threshold fixed coefficient estimates. Statistical significance at $\alpha = 0.05$ level is represented by marker shape. The reference level for month is September, i.e., statistical significance indicates significant difference from September. Bars represent 2 standard errors.

3 Fixed Threshold Logit-Normal Models

We selected 50, 75, 100, and 125 mm/day as fixed thresholds. 50 and 75 mm/day are light to moderate thresholds. 100 mm/day was the high setting used within Goswami et al (2006). 124.4 mm/day was the high setting in Dietz and Chatterjee (2014) based on Attri and Tyagi (2010), thus, 125 mm/day is used to approximate this.

3.1 Fixed Threshold Fixed Effect Analysis

Coefficients for fixed thresholds are seen in Figure 1. Covariates are not scaled within the models to facilitate comparisons across different model types (GLM, GEE, GLMM). The Niño 3.4 anomaly, latitude, and longitude generally display nonsignificant estimates, although longitude is significant at higher thresholds.

Intercepts are higher in the GLMMs compared to GEE or GLM. Thus, we'd expect higher probability of rainfall in the GLMM models based on the fixed effects only. The intercept is constant over thresholds in the GLMMs, while GEE and GLM coefficients increase with threshold.

Monthly adjustments for June and July indicate a significant positive effect compared to September. August is not significantly different from September. June and July show a slight increasing trend as the threshold increases inducing a higher probability of more extensive rainfall in June and July in comparison to September. This insight is consistent with earlier summer months typically containing larger amounts of rainfall events than September.

Western low elevation coastal areas and northeastern low lands receiving a large amount of rainfall may contribute to the significantly negative coefficient for log(Elevation+1). This estimate is relatively constant over threshold levels indicating a consistent effect. Both minimum and maximum temperatures coefficients are significantly negative. However, as the threshold increases, the magnitude of the minimum temperature coefficient decreases while the magnitude of the maximum temperature coefficient increases.

All monsoon circulation variables are significant in the models. The u-wind coefficients are positive at 200 mb and negative at 850 mb. Both are relatively constant as the threshold increased. The v-wind coefficients are negative at both pressure levels. The 850 mb coefficient decreased as threshold increased while the 200 mb is essentially constant as the threshold increased. The coefficient for ΔTT is significantly positive indicating higher probability of threshold exceedance as ΔTT increases.

3.2 Fixed Threshold Random Effect Analysis

Testing for the variance components[5] indicates both the intercept by station and intercept by year are significant over all threshold levels. However, the annual component makes up a much smaller proportion of the estimated variability. The station component decreases slightly as threshold increases.

In Figure 2, estimated random effects of the 125 mm/day exceedance GLMM with both random effects are shown for two different years. Positive (negative) random effects correspond to a higher (lower) probability of rainfall than that estimated by the fixed effects alone. Stations tend to consistently indicate either positive or negative (of varying magnitudes by year) random effects.

In 1987, negative random effects were larger and mostly fell within the center of India. In 2007, the positive random effects were more pronounced especially along the west coast and northern areas of the subcontinent. The two years examined were compared with Indian Meteorological Society rainfall data[6]. This annual summer monsoon season data provides percentage deviations from average rainfall amounts for four geographic demarcations in India- northwest, central, northeast, and south peninsula. In 1987, all but northeast India indicated drought which agrees with the stronger negative random effects produced by the model. 2007 had higher than average rainfall in all but northwest India; again, this agrees with the stronger positive

[5] Note that this is the standard deviation (σ) of the random effects distribution

[6] `http://www.imd.gov.in/section/nhac/dynamic/data.htm`

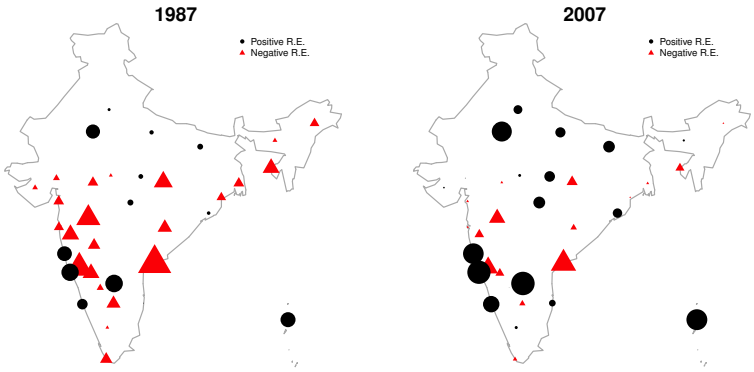

Fig. 2: Estimated random effects for $>$125 mm/day. The magnitude is depicted by the relative size of the marker. Triangles (circles) indicate negative (positive) estimated random effects.

random effects and higher chances of a large precipitation even. The correspondence is not one-to-one because the model is fitting probabilities of exceedances rather than actual rainfall amount, but provides some intuition for the random effects.

4 Percentile Threshold Logit-Normal Models

In Krishnamurty et al (2009), the median of the yearly 90^{th} and 99^{th} percentiles were used as thresholds for examining station-level percentile exceedances. Because of missing data, thresholds were defined using the direct 90^{th}, 95^{th}, and 99^{th} percentiles of the data. Models for the 99^{th} percentile failed to converge and are excluded.

4.1 Percentile Threshold Predictions for Selected Stations

Threshold exceedance predictions for 4 representative stations are displayed in Figure 3. Box plots indicate the expected pattern of decreasing probability as the threshold moves from the 90^{th} to the 95^{th} percentile. West coast stations, represented by Bombay, have markedly higher probabilities of exceeding their station thresholds. Bombay has station thresholds of 59.9 mm/day (90^{th}) and 92.9 mm/day (95^{th}). In comparison, more moderate exceedance probabilities were seen by Calcutta and New Delhi. Calcutta has thresholds of 39.9 mm/day (90^{th}) and 56.9 mm/day (95^{th}) and New Delhi thresholds are 34.0 mm/day (90^{th}) and 52.1 mm/day (95^{th}). Thiruvanthapuram, in the southmost region of India, indicated low predicted probabilities of exceeding it's extreme thresholds of 34 mm/day (90^{th}) and 49 mm/day (95^{th}).

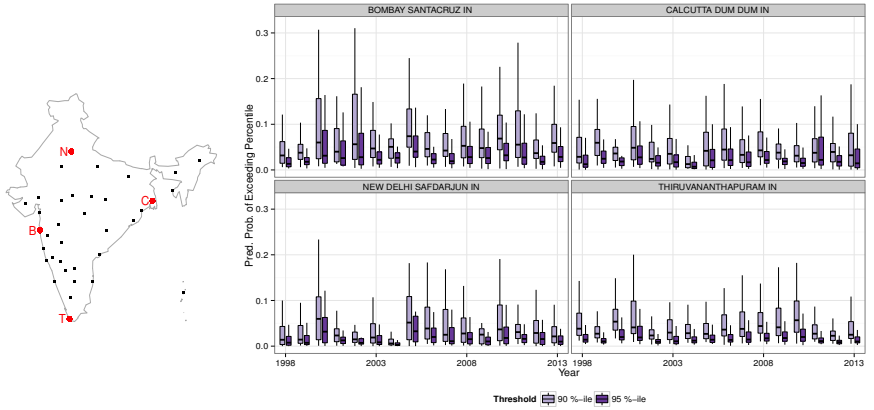

Fig. 3: Percentile Threshold Predictions. Box plots show the distribution of daily predictions by year. Outliers are not shown for clarity of the graphics and consisted of <5% of yearly predictions.

Compared with the fixed thresholds analysis, the percentile-based analysis suggests the use of much lower thresholds for understanding local monsoon behavior.

We note the appearance of an irregular cycle in the probability predictions shown for each of the stations in the 1998-2013 period. The cycle is not consistent among all stations. This may be due to the random effects for each station in each year which captures some of the idiosyncratic features of a location.

5 Summary and Future Work

The analysis in this study serves as a starting point for climate scientists in exploring thresholds. These thresholds are useful in an explicit context of understanding risk to civil structures or in an implicit context of further modeling. Specifically, fixed threshold analysis statistically examines the relationships of climate covariates with rainfall probabilities in the context of increasing thresholds. This may be useful in a large scale analysis of the Indian monsoon. Percentile based thresholds are useful at a local scale for understanding risks of certain levels of rainfall.

Possible limitations of our approach include model fit and data issues. One measure of fit provided within SAS is a Generalized Chi-Square (GCS) statistic. We'd expect this statistic to be around 1 if the model fits well. Fixed threshold models GCS ranged from 1.06 to 2.06, and increased with the threshold, indicating a slight issue in fit at the higher thresholds. There were also outliers indicated by residual plots which indicate the need to employ a more robust fit in future. Missing data could be driving some of the results; several possibly important areas of India are not included in the data set based on availability. Unfortunately, the wet northeast as

well as the central and northwest regions of India are poorly covered. Aggregating data may provide a different perspective and a more stable fit.

However, in general, we believe the logit-normal mixed model in this context provides valuable physical insights, such as the increasing importance of maximum temperature as threshold increases, as well as understanding of local predictions and their cycles. In future work, model residuals may be used in a spatial correlation testing framework to establish high thresholds. We also plan to investigate model selection techniques in the context of GLMM to identify a "best" model.

Acknowledgements This research was partially supported by the National Science Foundation under grants # IIS-1029711 and # SES-0851705.

References

Attri SD, Tyagi A (2010) Climate profile of india. Tech. rep., Government of India, Ministry of Earth Sciences, India Meteorological Department

Breslow N, Clayton D (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88(421):9–25

Dietz L, Chatterjee S (2014) Logit-normal mixed model for indian monsoon precipitation. Nonlinear Processes in Geophysics 21:939–953

Goswami B, Venugopal V, Sengupta D, Madhusoodanan MS, Xavier PK (2006) Increasing trend of extreme rain events over india in a warming environment. Science 314:1442–1445

Jiang J (1998) Consistent estimators in generalized linear mixed models. Journal of the American Statistical Association 93(442):720–729

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D (1996) The ncep/ncar 40-year reanalysis project. Bulletin of the American Meteorological Society 77:437–470

Krishnamurty CKB, Lall U, Kwon HH (2009) Changing frequency and intensity of rainfall extremes over india from 1951 to 2003. Journal of Climate 22(18):4737–4746

Lele SR, Nadeem K, Schumuland B (2010) Estimability and likelihood inference for generalized linear mixed models using data cloning. Journal of the American Statistical Association 105(492):1617–1625

McCulloch CE, Searle SR (2010) Generalized, Linear, and Mixed Models, 2nd edn. Wiley Series in Probability and Statistics, John Wiley and Sons, Inc., Hoboken, New Jersey

SAS Institute Inc. (2011) SAS/STAT 9.3 User's Guide. SAS Institute Inc., Cary, NC

Xavier PK, Marzina C, Goswami BN (2007) An objective definition of the indian summer monsoon season and a new perspective on the enso–monsoon relationship. Quarterly Journal of the Royal Meteorological Society 133:749–764