Response to Referee 2 Comments

Lindsey Dietz and Snigdhansu Chatterjee

June 19, 2014

We'd thank the referee for their useful and detailed comments. It is clear that to address the reviewer's concerns, we need to re-compute several parts of our data analysis as well as rewrite considerable parts of our manuscript. We plan to do for our revision. We believe working to smooth out the reviewer's concerns will improve the paper, and the additional and more thorough data analysis we carried out in response to the comments by both reviewers has brought out important features we plan to discuss in the revised manuscript.

The following includes the comments of the reviewer and our response. We are in the process of revising the manuscript to include the results of these comments. The reviewer has made two general comments followed by several specific comments where the general comments overlap with the specific ones. Our response is itemized following the same pattern.

General Comments 1: *First, there appears to be too little focus on the issue being addressed, extremes of rainfall for the Indian monsoon. In general, identification of extreme events at the station level and of their drivers at a larger-than-station level is of some interest to the researchers studying extremes of rainfall. However, insufficient attention is paid to these important points in both the analysis as well as at the level of motivation of the research (including an incomplete and somewhat incoherent review of the existing results) and choice of the data set (see below for specific details). The overall conclusions of the study are, correspondingly, very vague e.g. ..promise of modeling some of the unobservable physical features within the complicated network of interactions the underly extreme climate patters while maintaining statistical integrity. (pp 206). It is not explained in what sense the method(s) illustrated here are better than (or add to) those used in the prior literature, nor is it indicated how larger scale patterns (e.g. regional distinctions in effect of covariates) in rainfall extremes are to be thought of, in the context of the model presented. Finally, there is no discussion of (or motivation for) the possible differences in effects of potential large-scale drivers (e.g. ENSO and the IOD) between normal and extreme rainfall.*

General Comment 1 Response:

- We understand the reviewer's concerns and appreciate these comments. We will rewrite different parts of the paper to address the issues raised here. See further details relating to motivation in our response to specific comment on *Extremes for the Indian sub-continent Comment 1.*
- Our choices of data sets were made based on what was widely and freely available in order to ensure reproducibility of research, and further methodological development free of data protection concerns. Future work may hone in on more specific data including the IMD data set.
- We are carefully revising overall conclusions to be more clear and pointed. We are especially working to include how use of GLMM adds the literature for study of precipitation extremes along with identifying regional distinctions in the effect of covariates.

- We have fit new models that include both an ENSO and IOD covariate. We will include discussion of differences in effects on "light", "moderate", and "extreme" rainfall.

General Comment 2: *Second, given that the focus of the analysis almost exclusively on the statistical aspects of the GLMM model, and in general statistical analysis, the model details themselves are sketchy and/or incomplete to the point of incoherence. Basic details are sometimes not provided (e.g. the MSIM model in §3.2 is hurriedly and incompletely sketched and in the Bayesian model, in §3.3, virtually no details or equations are provided for the posterior distributions and method of simulation, a key point in Bayesian analysis) while even the basic GLMM model set up, in § 2, suffers from many problems, including unlabeled and unexplained objects ($b(\eta_i)$ and $c(y_i)$ are used (inappropriately) interchangeably and the first equation in the paper, pp 196, is clearly wrong). These issues can leave the reader unfamiliar with GLMM models confused, defeating the very purpose of the detailed exposition of GLMM models provided in the study. This is rather surprising, given the relatively standard nature of these tools in the statistical literature. Finally, §5, a key section of the study, suffers from excessive brevity in both methodology and discussion of results. Too many of the coefficients are not significant, and further, the simulation results (§4.2) indicate that many models are severely biased and do not at all perform well. Given this, the authors do not explain how condent the applied researcher can be with his results, upon using this methodology. Further, model goodness-of-fit measures (e.g. LR tests) are neither reported nor even discussed, rendering the robustness of these models (and, by extension, their applicability) somewhat suspect. Finally, the writing is in general confused, with far too many clumsy constructions, meaningless (e.g. pp 195, lines 9-11, first sentence, pp 196, sentence beginning on line 14, pp200) or even incomplete (line 13, pp 198, beginning with and the approximate...) sentences. I suggest the authors carry out a very careful editing of the entire paper in order to eliminate these (and other) errors.*

General Comment 2 Response:

- Model details will be added to clarify exposition of the models as well as each of the estimation methodologies in the manuscript.
- Unlike standard fixed but unknown parametric statistical models that are commonly used (including in the literature on Indian summer precipitation), models with random effects have additional technical complications when it comes to assessing the level of significance of a variable, or goodness-of-fit, and several other things. We have implemented the steps that are currently available for these, in the current problem under study, despite the well-known limitations and caveats for these. In the revision, we plan to report our results on significance of model coefficients and model goodness-of-fit measures, while including the caveats and references. Further details on this matter are discussed below in the responses to *Methodology Comments 6, 7, and 8.*
- Deficiencies in §5 are addressed below in the response to *Methodology Comment 8.*
- Simulation results in §4 indicated glmmPQL was biased. Also, since it contributes minimally, we plan to remove the bootstrapping dclone method from the revised paper to increase clarity and focus. The other 3 methods accurately estimated the truth, albeit at varying speeds. Thus, these 3 are used in the application to Indian rainfall.

- We are in the process of carefully editing of our final draft before submission to ensure the elimination of errors.

Extremes for the Indian sub-continent Comment 1: *A (very brief) discussion of existing literature beyond Goswami et al. (2006) would be helpful. In particular, the studies Krishnamurthy et al. (2009), Rajeevan et al. (2008), Ajayamohan and A. Rao (2008) and Ajayamohan et al. (2008) deal with different aspects of the issue and their inclusion, if only briefly, will particularly enrich the discussion and motivation.*

Extremes for the Indian sub-continent Comment 1 Response: We are currently looking at these studies and will add appropriate discussion to the introduction part of the document.

Extremes for the Indian sub-continent Comment 2: *The problems associated with the use of fixed rainfall levels for thresholds has been discussed, for the Indian context, in Krishnamurthy et al. (2009). While recognizing that the use of percentiles at the station level for the current study is probably too time consuming, I feel that at a minimum, a (very brief) discussion of possible problems with the approach used for identifying extreme rainfall in this study is likely very useful.*

Extremes for the Indian sub-continent Comment 2 Response:
Krishnamurthy et al. (2009) did utilize data derived percentile thresholds which changed depending on spatial location. However, their research was focused on trend analysis and did not utilize relevant covariates. Because our application of GLMM to this kind of data is nascent, and we are able to include spatial covariates, we only considered fixed thresholds for the entire country. We will further investigate the use of spatially varying thresholds in the future and will note this in the manuscript.

Methodology Comment 1: *Numbering the equations will make it easier for the reader (and the referee) to easily discuss them. The first three equation in §2 must be clearly explained, as it is not now. For instance, the first equation is clearly incorrect, and should read instead*

$$\mathbf{Y}_i|\mathbf{u} \overset{ind.}{\sim} f_{n_i}(Y_i|u) = \exp\left\{\frac{\mathbf{Y}_i\eta_i - b(\eta_i)}{\tau^2} - c(Y_i,\tau)\right\} \tag{1}$$

In addition, it is never explained to the reader that the expression above is actually the density of a random variable from the exponential family, with b(.) and c(.) having specific relationship to η_i and that τ is a scale parameter. Also, Y_i ,y_i and u, U are used interchangeably, confusingly and incorrectly, respectively. Finally, there is a typographical error in the first equation, where f_{n_i} reads f_{η_i}.

Methodology Comment 2: *Line 11, pp 197, introduces confusing notation for density function with parameters β, σ, as $f_{\beta,\sigma}(.)$. It might be less confusing (and cumbersome) to instead use $f(.|\beta,\sigma)$.*

Methodology Comments 1 and 2 Response: We meant for f_{η_i} to be written as we did with $f_{\beta,\sigma}$ in reference to the parameter of interest. However, we will clarify the notation in this section as suggested in order to eliminate confusion regarding the model.

Methodology Comment 3: *In §3.1, Penalized Quasi-likelihood methods, it is never clearly explained to the reader what is penalized. In the interest of the readers understanding, it might be worth inserting the equation corresponding to the Laplace approximation and mentioning exactly that the penalty term comes from the approximation process for the likelihood.*

Methodology Comment 3 Response: We will expand the discussion of PQL in the document to provide more detail on the methodology. This will include information on Laplace approximation and where the penalty term arises.

Methodology Comment 4: *In §3.2, the MSIM method is very confusingly described. It is not likely clear to the casual reader what is being simulated and where the unknown parameters enter into the second stage Newton-Raphson method. I recommend clarifying these aspects to the reader.*

Methodology Comment 4 Response: We are updating the section to state more precisely the set of estimating equations, pertinent simulation, and optimization technique.

Methodology Comment 5: *I also recommend adding, just before the beginning of §5, a brief summary of the rather confusing §4, to aid the reader in understanding the crux of the issues addressed in §4.*

Methodology Comment 5 Response: We are updating the summary provided in §4.4 to sufficiently describe the output of the simulations performed.

Methodology Comment 6: *In your discussion of the results (e.g. paragraph 3, pp 205) you compare coefficients across time (as in Fig. 6a). However, such a comparison of coefficients over time is carried out without any measure of individual confidence interval (i.e. it is not known how many of those individual fixed effects are significant), leading to difficulties of comparisons. In addition, in the absence of a joint CI for the whole period, the cyclic patterns detected can easily be a statistically meaningless artifact. It is important to at least flag these issues to the reader and in general, to not over interpret these results.*

Methodology Comments 7: *Regarding §5, my concern is also that there are really no results to speak of, since many of the effects are mentioned as being not significant (although there is no systematic enumerating of how many coefficients are not significant in each of the methods used). As mentioned already, this section is really too brief to allow the reader to judge the utility of the methods illustrated. A clearer, slightly expanded, narrative here will allay such concerns.*

Methodology Comments 6 and 7 Response: Testing in regard to the fixed effects of a GLMM is a fairly difficult problem and each of the estimation methods would require substantially different approaches. Nevertheless, we will include available tools for significance and goodness-of-fit tests and include the caveats on them as well. Also, in view of the comments and suggestions by the other reviewer, we will implement and report several standard tests in our revised data analysis.
The following is a list of issues with significance testing as-it-is in the framework used in the current paper:

- `glmer{lme4}` produces Wald z-statistics for each parameter in the model summary. We are wary of giving readers a false sense of security by reporting these p-values when even one of the lme4 package developers, Ben Bolker, has stated at `http://glmm.wikidot.com/faq` that the assumption of normal sampling distributions requires a "leap of faith" and is known to have problems in logit-normal contexts. He also states that an LRT is not reliable for fixed effects in finite-size cases. There are others who have expressed reservation about p-values elicited from standard GLMM implementation packages as well.

- `dclone{dclone}` provides sample mean and sample standard deviation from the MCMC chain for all parameters estimated. Based on the convergence to posterior distribution to a normal distribution, one could construct hypothesis tests. However, assessing convergence for MCMC can be very difficult, thus, this approximation could be quite poor and any p-values gleaned from it would not be meaningful. Bayesian methods of hypothesis testing, such as the use of the Bayes factor, don't provide p-values.

- Our current implementation of *MSIM* does not have a method of calculating standard errors. Jiang 1998 suggests bootstrapping to produce standard errors.

  Unfortunately, a typical bootstrap implementation (parametric or nonparametric) *does not* reproduce the null hypothesis scenario, as has been reported and discussed in literature. In general, bootstrap reproduces the data generation process up to a certain level of accuracy, which matches the null hypothesis world only if the null hypothesis is true and there are no additional complex (important or nuisance) parameters, a very simplistic scenario that is not valid for our model.

We should note that we have added several covariates as suggested by another referee which has in turn, changed our output in a meaningful way. 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.

The LRT for GLM is a much more reliable test than those available for GLMM estimation methods. However these tests are subject to multiple testing issues, which we will address with standard procedures for correcting for false discoveries. We can interpret this as a screening test for the covariates in our models.

Methodology Comment 8: *Finally, given that this study is primarily illustrating the applicability of a methodology to an applied audience, the lack of clear cut recommendations regarding the methods illustrated and an investigation of the robustness of some of the methods must be addressed by the authors. First, some guidelines regarding, or a discussion comparing, the different approaches to GLMM is likely to be very helpful to the applied researcher. For instance, it is never clearly explained which of the methods perform better with the data set investigated by the authors.*

Given this, can the applied researcher wishing to use the method in his own research, draw any broad conclusions? In addition, the issue of goodness-of-fit must be at least discussed. Can reliable *tests for the variance components be easily implemented for many of the methods investigated in the study? While a comprehensive investigation of these issues is likely outside the scope of the current study, a few simulation investigations and some discussion of these issues, possibly drawing on existing statistical literature, is surely pertinent here.*

Methodology Comment 8 Response: We will work to provide more clear cut recommendations regarding the estimation methods illustrated as well as the use of the model as a whole.

The robustness of the methods is somewhat addressed in the simulations where at least one of the standard methods (glmmPQL) did not produce accurate results in a controlled setting. Other statements regarding robustness of methodology will be added to §2 when discussing the estimation procedures.

We will add more specific guidelines regarding the different approaches to GLMM and further discussion as to which of the methods perform better with the data set investigated by the authors.

We have chosen to use the GLM to provide a test of goodness-of-fit based on the residual deviance. For details, refer to [1]. This compares the fitted model to the saturated model which contains one parameter for each observation.

Doing a parametric bootstrap is another goodness-of-fit possibility, however, this would not give us p-value, because we cannot generate conditions identical to the null hypothesis without knowing the parameter values.

Finally, we will test the variance component in the GLMM fit by `glmer` by using a LRT with a nonstandard asymptotic distribution. Because we are using models with a the single variance component, the asymptotic distribution for the LRT corresponds to 0.5 multiplied by the p-value obtained from the χ_1^2 distribution as noted by [2]. We can only do this test for `glmer` since the other two methods do not use maximum likelihood. As a final caveat, we note that the likelihood produced by `glmer` is only approximate due to the estimation procedure. Thus, we take a cautious view on the use of these tests. It is also difficult to apply this test in practice to any more complicated variance component structure.

References

[1] Julian J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Texts in Statistical Science. Taylor and Francis, 2006.

[2] Daowen Zhang and Xihong Lin. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In David Dunson, editor, *Random Effect and Latent Variable Model Selection*, volume 192 of *Lecture Notes in Statistics*, pages 19–36. Springer New York, 2008.