MAUSAM, **69**, 1 (January 2018), 19-28

551.577.3 (540)

A Bayesian zero-inflated exponential distribution model for the analysis of weekly rainfall of the eastern plateau region of India

ARNAB HAZRA, SOURABH BHATTACHARYA and PABITRA BANIK*

Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, India

**Agricultural and Ecological Research Unit, Indian Statistical Institute, Kolkata, India*

(*Received 3 March 2017, Accepted 8 August 2017*)

***e mail : banikpabitra@gmail.com**

सार – लम्बे समय से जलवायु विज्ञानी और कृषि मौसम विज्ञानी का मुख्य अनुसंधान का क्षेत्र वर्षा के आँकडो की सॉख्यकीय मॉडलिंग करता रहा है। विशेषकर शीत सप्ताह के दौरान अक्तूबर शून्यमान की उच्च प्रतिशतता को शामिल करते हुए लघु अवधि वर्षा का मापन किया जाता है। शून्य से बढ़े हुए मॉडलो का मॉडलिग में ऐसे आँकड़ा सेटों का प्रायः प्रयोग किया गया है। इस शोध में हमने शून्य से धातांकीय रूप में बढ़े हुए वितरण का प्रयोग करके साप्ताहिक वर्षा आँकड़ा मॉडल का तैयार करने का प्रयास किया है। हालांकि मौसम विज्ञानियों द्धारा अधिकतम संभावित अनुमान को प्रायः प्राथमिकता दी जाती है। इस शोध पत्र में कुछ कमियों के बारे में चर्चा की गई है। इसलिये हमने बायेसियन मॉडल पर विचार किया है। अर्थात हमने माना कि नियत मात्राओ के बजाय मॉडल के अनियमित प्राचल होगे जैसे कि अधिकतर प्रयास में होता है। जैसे कुछ बायेसियन विशलेषण के वास्तविक भागों में है। हमारी मुख्य चर्चा की प्राथमिकता प्राचलों के उत्तरोतर वितरणों पर आधारित व्यतिकरण और उत्तरोन्तर संभावित वितरणों पर आधारित संभावित वर्षा मात्राओं की विभिन्न प्रतिशत की गणना को दी है। भारत में गिरडीह में वर्ष 1969 से 2009 तक के लिये गये वर्षा आँकडों के सेटों का हमने सप्ताहिक विशलेषण किया है पूर्वी पठार क्षेत्र में सिचांई सुविधाओं में कमी के कारण केवल वर्षा के परिमाण कृषि निर्भर होती है। हमने 10%, 30%, 50%, 70%, 90%, संभावित वर्षा मात्राओं को उपलब्ध कराया है जो कृषि के कामों पर विशेष कदम के लिये सटीक सप्ताह को तय करने में मदद करेगा।

ABSTRACT. Statistical modelling of rainfall data has been a major research area of the climatologists and agro-meteorologists for quite a long time. Short-period rainfall measurements often include a high percentage of zero values, particularly during the winter weeks. Zero-inflated models are often used in modelling such datasets. In this paper, we attempt to model weekly rainfall data using zero-inflated exponential distribution. Though a frequentist approach (mainly maximum likelihood estimation) is often preferred by meteorologists, it has a few shortcomings discussed in this paper. Hence, we consider a Bayesian model, *i.e*., we assume the model parameters to be random instead of fixed quantities as in a frequentist approach. As some obvious parts of a Bayesian analysis, we discuss the prior choices, inference based on the posterior distributions of the parameters and calculations of the different percentage probability rainfall amounts based on the posterior predictive distributions. We analyze weekly rainfall dataset for the years 1969-2009 collected at Giridih, India. In the eastern plateau region, agricultural operations depend solely on the rainfall quantities because of the lack of irrigation facilities. We provide 10%, 30%, 50%, 70%, 90% probability rainfall amounts which would help in deciding the accurate week for a particular step of an agricultural procedure.

Key words – Zero-inflated exponential distribution, Bayesian paradigm, Prior distribution, Posterior distribution, Posterior predictive distribution, Quantiles.

1. Introduction

The percentage of rural population in India is veryhigh (68.84% in 2011; source: http://censusindia. gov.in) and the main source of income and employment in the rural areas is agriculture. Forecasting of different meteorological parameters, mainly rainfall, is very important, particularly in the rain-fed agricultural ecosystem. Lack of proper irrigation system often makes the agricultural practices solely dependent on rainfall. An estimate of the future rainfall amount becomes a necessary tool before implementing any step of an agricultural procedure. Thus, modelling short-period rainfall, e.g., weekly data is very important from this perspective.

In the literature, several studies are available for rainfall analysis and the commonly used probability models are- exponential (Burgueño *et al*., 2005 & 2011; Taewichit *et al*., 2013), gamma (Husak *et al*., 2007; Liang *et al*., 2012; Krishnamoorthya and León-Novelo, 2014), log-normal (Kwaku and Duke, 2007; Sharma and Singh, 2010), Weibull (Burgueño *et al*., 2005; Lana *et al*., 2017),

Pearson Type-III/V/VI (Hanson and Vogel, 2008; Khudri and Sadia, 2013; Mayooran and Laheetharan, 2014), log-logistic (Fitzgerald, 2005; Sharda and Das, 2005). In the case of short-term rainfall such as weekly rainfall, for most part of the globe, there is a high chance of the observation being zero, considering not only the wet weeks but also the dry weeks. Thus, an obvious choice for modelling short-period rainfall is some zero-inflated positive continuous distribution recently studied by Muralidharan & Pratima (2017) under practical scenarios including rainfall modelling. The zero-inflated exponential distribution has been used under different practical scenarios by several authors, e.g., Kale and Muralidharan (1999), Muralidharan (1999) and Velarde *et al*. (2004). Muralidharan and Kale (2002), Singh *et al*. (2009) and Kumar *et al*. (2015) have analyzed the rainfall data based on zero-inflated gamma distribution while Muralidharan and Lathika (2005) have used zero-inflated Weibull distribution for modelling rainfall data collected at Jalgaon and Coimbatore divisions in India during a 10 year period from 1961 to 1970. Hazra *et al*. (2014) have considered different types of the zero-inflated positively continuous distributions for modelling Nakshatra-wise rainfall data of the eastern plateau region of India.

As of the authors' knowledge, all the papers implementing zero-inflated positively continuous distribution models are based on the frequentist paradigm, *i.e*., the model parameters are assumed to be unknown fixed numbers and estimated in terms of maximum likelihood estimates (MLEs) in general. In comparison, the parameters in Bayesian models are assumed to be random and inference about the parameters are drawn based on the posterior distributions, *i.e*., the conditional distributions of the parameters given the observed data. Bayesian methods are more robust in general, widely used among the statistical community (Gelman *et al*., 2014) and often lead to more meaningful estimates as discussed in this paper. Here we model weekly rainfall data for the years 1969-2009 collected at Giridih, India, using zero-inflated exponential distribution in a Bayesian framework. In the eastern plateau region of India, agricultural operations depend solely on the rainfall quantities because of the lack of irrigation facilities. Hence, the contribution of this paper is two-fold, in terms of the methodology as well as the analysis. We provide 10%, 30%, 50%, 70%, 90% probability rainfall amounts which would help the agricultural practitioners in deciding the accurate week for a particular step of an agricultural procedure.

2. **Materials and method**

The Data : Weekly rainfall data of Giridih, India (24°18' N, 86°30' E) for 41 years (1969-2009) are used in this study. The daily rainfall data have been collected from the Damodar Valley Corporation (DVC) for the period 1969-1989 and by the Indian Statistical Institute, Giridih, for the period 1990-2009. Weekly rainfall totals are calculated based on the definition of Standard Meteorological Weeks (SMW) provided in Table 1 of Chand *et al*. (2011). Rainfall data are collected using a rain gauge as specified by the India Meteorological Department (IMD).

The model: Short-period rainfall is a random variable which is non-negative, continuous on the set $(0, \infty)$ and has a positive probability of having zero rainfall mainly because of considering short-period rainfall over pre-monsoon and post-monsoon months. Suppose Y = total rainfall on a specific week. We model Y as follows:

$$Y = \begin{cases} Xw.p. & p \\ 0 \quad w.p. & 1-p \end{cases}.$$

For the random variable X we consider the exponential distribution with rate λ. So, the cumulative distribution function (CDF) of Y is of the form :

$$G(x|p,\lambda) = 1 - p + p\left(1 - e^{-\lambda x}\right) \quad \forall\, x \,\epsilon\, [0,\infty)$$

The goodness-of-fit of this model can be judged based on Q-Q plots which would confirm whether the modelling assumptions holds quite well or not. Also, we can check the quantiles of the data as well as the model after censoring the zero values which is valid because the model considers the zero and non-zero parts separately (Hazra *et al*., 2014).

Significance of the model parameters : Here the parameter $p$denotes the probability of observing a wet week, *i.e*., $p = \Pr(Y > 0)$. Thus, a larger value of p indicates higher chance of observing a wet week. Also, for any amount of rainfall x mm, $\Pr(Y > x) = p\, e^{-\lambda x}$. Thus, the larger is the rate parameter λ, the probability $\Pr(Y > x)$ decreases, *i.e*., the chance of observing high rainfall amount decreases with increase in λ.

Choice of prior : In Bayesian paradigm, we assume that the parameters are also random. For the purpose of Bayesian inference, we put priors on parameters p and λ. To make computation easy, we choose conjugate priors (*i.e*., the prior distribution and the posterior distribution belong to the same class. For more details, please refer to Ghosh *et al*. (2006) for both the parameters, *i.e*., $p \sim Beta(\alpha_p, \beta_p)$ and $\lambda \sim Gamma(\alpha_\lambda, \beta_\lambda)$. The density functions are given by :

$$P[p] = \frac{\Gamma(\alpha_p + \beta_p)}{\Gamma(\alpha_p)\Gamma(\beta_p)} p^{\alpha_p - 1}(1-p)^{\beta_p - 1}$$

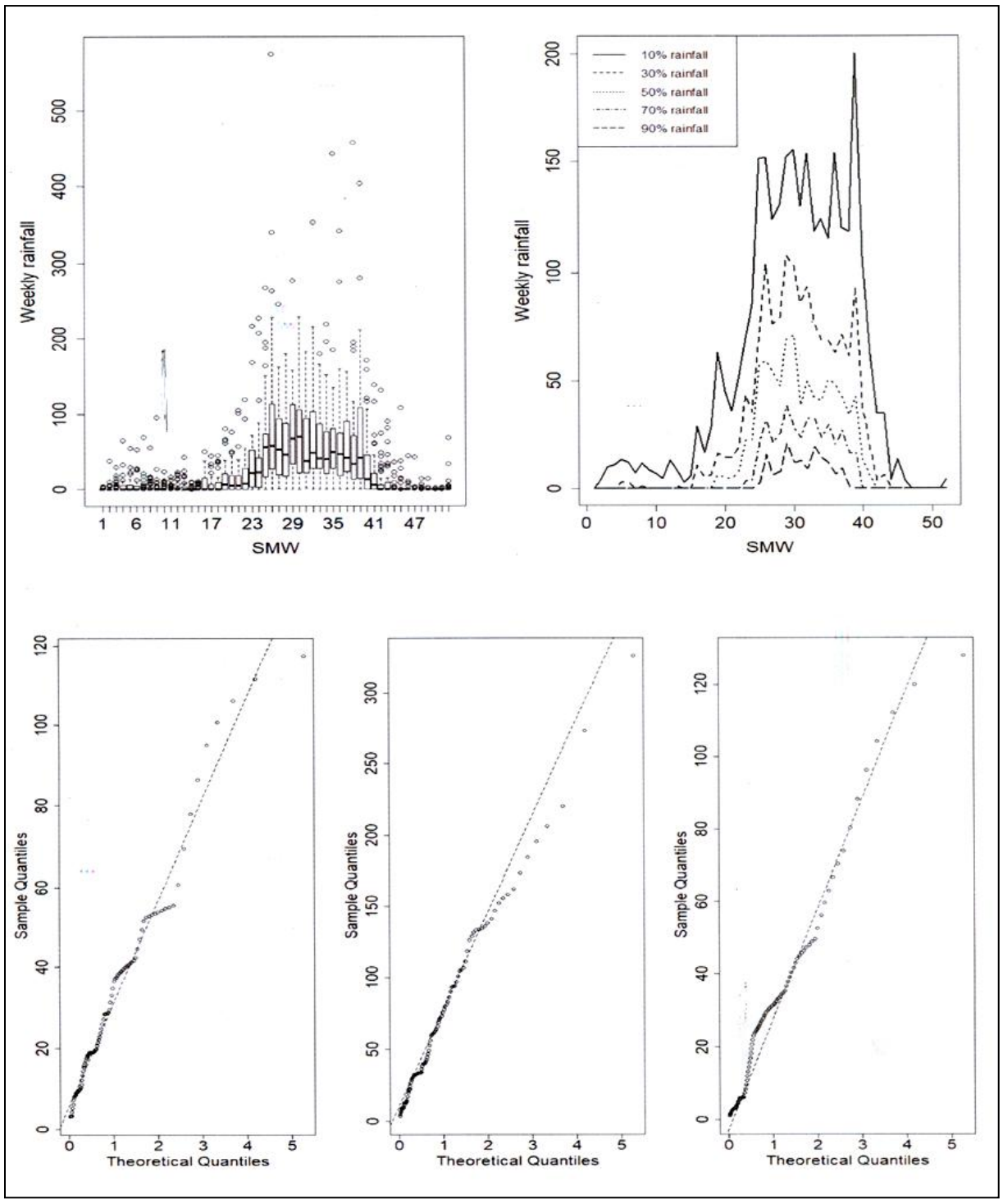

Fig. 1. Box plots and the corresponding nonparametric estimates of weekly rainfall total at different probability levels for 52 weeks (top panel). Q-Q plots for SMW22, SMW32 and SMW42 respectively (bottom panel)

$$P[\lambda] = \frac{{\beta_\lambda}^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)} \lambda^{\alpha_\lambda - 1} e^{-\beta_\lambda \lambda}$$

We also assume that $p$and λ are independently distributed. The hyper-parameters, *i.e.*, $\alpha_p, \beta_p, \alpha_\lambda, \beta_\lambda$ are

assumed to be known. We consider them to be $\alpha_p = 1$, $\beta_p = 1, \alpha_\lambda = 0.01, \beta_\lambda = 0.01$, *i.e.*, $p \sim Beta\,(1,1) = Uniform\,(0,1)$ and $\lambda \sim Gamma\,(0.01, 0.01)$ so that E(λ) = 1 and Var(λ) = 100, which is large enough. The idea for choosing such hyper-parameter values comes from the fact that in case we haven't observed the data and as p can take any value in [0, 1], we assume that it has a homogeneous distribution over [0, 1]. Small mean and large variance of λ flattens its distribution enough and hence these choices are non-informative (Syversveen, 1998).

Posterior distribution of parameters : For a particular week, suppose $Y_1, Y_2, \ldots, Y_n$ denote the data for *n* years. With the priors we consider, the posterior distribution of p is, $P[p|Y_1, Y_2, \ldots, Y_n] = Beta(n - \sum_{i=1}^{n} I(Y_i = 0) + 1, \sum_{i=1}^{n} I(Y_i = 0) + 1)$ and the posterior distribution of λ is $P[\lambda|Y_1, Y_2, \ldots, Y_n] = Gamma\ (n - \sum_{i=1}^{n} I(Y_i = 0) + 0.01, \sum_{i=1}^{n} Y_i + 0.01)$. Here I denotes an indicator; $I(A) = 1$ if the event A happens and 0 otherwise. The derivations and more mathematical details are moved to Appendix. We consider posterior means as the estimates of p and λ and call them "*Bayesian estimates*" henceforth in this paper. Thus, the estimates are :

$$\hat{p} = \frac{n - \sum_{i=1}^{n} I(Y_i = 0) + 1}{n + 2}$$

$$\hat{\lambda} = \frac{n - \sum_{i=1}^{n} I(Y_i = 0) + 0.01}{\sum_{i=1}^{n} Y_i + 0.01}$$

In comparison, the MLEs are given by $\hat{p}_{MLE} = \frac{n - \sum_{i=1}^{n} I(Y_i = 0)}{n}$ and $\hat{\lambda}_{MLE} = \frac{n - \sum_{i=1}^{n} I(Y_i = 0)}{\sum_{i=1}^{n} Y_i}$. Thus, the estimates of both the approaches are close but not exactly the same. Along with the *Bayesian estimates*, we provide 95% credible regions (*i.e.*, the upper and lower 2.5% quantiles of the posterior distributions) which indicates that the parameters lie within the corresponding intervals 95% of the time (The inference is different from that of 95% confidence intervals in a frequentist context where a 95% confidence interval means the intervals contain the true value of the parameter 95% of the times).

Posterior predictive distribution : For a future observation Y^*, the posterior predictive distribution is given by the conditional density of Y^*given the data, *i.e.*, $P[Y^*|Y_1, Y_2, \ldots, Y_n]$. Here the posterior predictive distribution is *zero-inflated Lomax* distribution with shape parameter $\sum_{i=1}^{n} I(Y_i = 0) + 0.01$ and scale parameter $\sum_{i=1}^{n} Y_i + 0.01$ with zero-inflation parameter $\frac{\sum_{i=1}^{n} I(Y_i = 0) + 1}{n+2}$. The mathematical details are provided in Appendix. The amount of $100r\%$ probability rainfall on each week is given by the $(1-r)^{th}$quantile of the posterior predictive distribution.

3. Results

The box plot of the data for 52 weeks are provided in the top-left panel of Fig. 1. For SMW 1-15 and 44-52, the boxes are concentrated at zero with only a few positive outliers. For SMW 23-39, the boxes do not include the zero value and exhibit a few very high positive valued outliers. For other weeks, the boxes do not include zero value and display similar patterns of positive outliers. Thus, in an overall sense, the distribution is positively skewed with inflation at the value zero for each week. The 10%, 30%, 50%, 70%, 90% probability rainfall amounts based on the data (before fitting a parametric model) are provided in the top-right panel of Fig. 1. Even for the 90% probability level, we note that the rainfall totals are non-zero for the weeks 25-38 and thus, it is highly likely to observe a wet week during that period. For the 10% probability level, the rainfall totals are non-zero for the weeks 1, 47, 49-51. Thus, it is highly unlikely to observe a wet week during these weeks. For 50% probability level, the weeks 19-41 show non-zero amount of rainfall.

For three weeks SMW 22 (pre-monsoon), 32 (monsoon), 42 (post-monsoon), we provide the Q-Q plots in the bottom panel of Fig. 1 which shows that an exponential distribution gives a justified fit for the non-zero observations and hence, we fit the Bayesian zero-inflated exponential distribution model separately for each week. The Bayesian estimates of p and λ for each SMW are provided in the top panel of Fig. 2. The 95% credible regions are also provided along with the posterior means. The estimates of p remain above 0.5 for the weeks 19-41, same as the wet weeks at 50% probability level and remain above 0.9 for the weeks 25-37 while the wet weeks at 90% probability rainfall are SMW 25-38. Also, the weeks with the estimates of p less than 0.1 conforms to the dry weeks at 10% probability rainfall. We also notice that the estimates of p drop soon after the monsoon period and again increase for the weeks 51-52. The estimates of λ are minimum for SMW 26 with value 0.0107 mm^{-1} while maximum for SMW 50 with value 1.5827 mm^{-1} and remains less than 0.02 mm^{-1} for the weeks 25-40. For the weeks 27 and 31, the Bayesian estimates of p are 0.9767, while the corresponding frequentist estimates (MLEs, *i.e.*, the proportions of wet weeks) are 1. The standard errors of the MLEs are zero while the standard deviation of the posterior distribution is non-zero (2.27×10^{-2}). For the parameter λ, the Bayesian estimates and the MLEs conform after rounding the results till three decimal places.

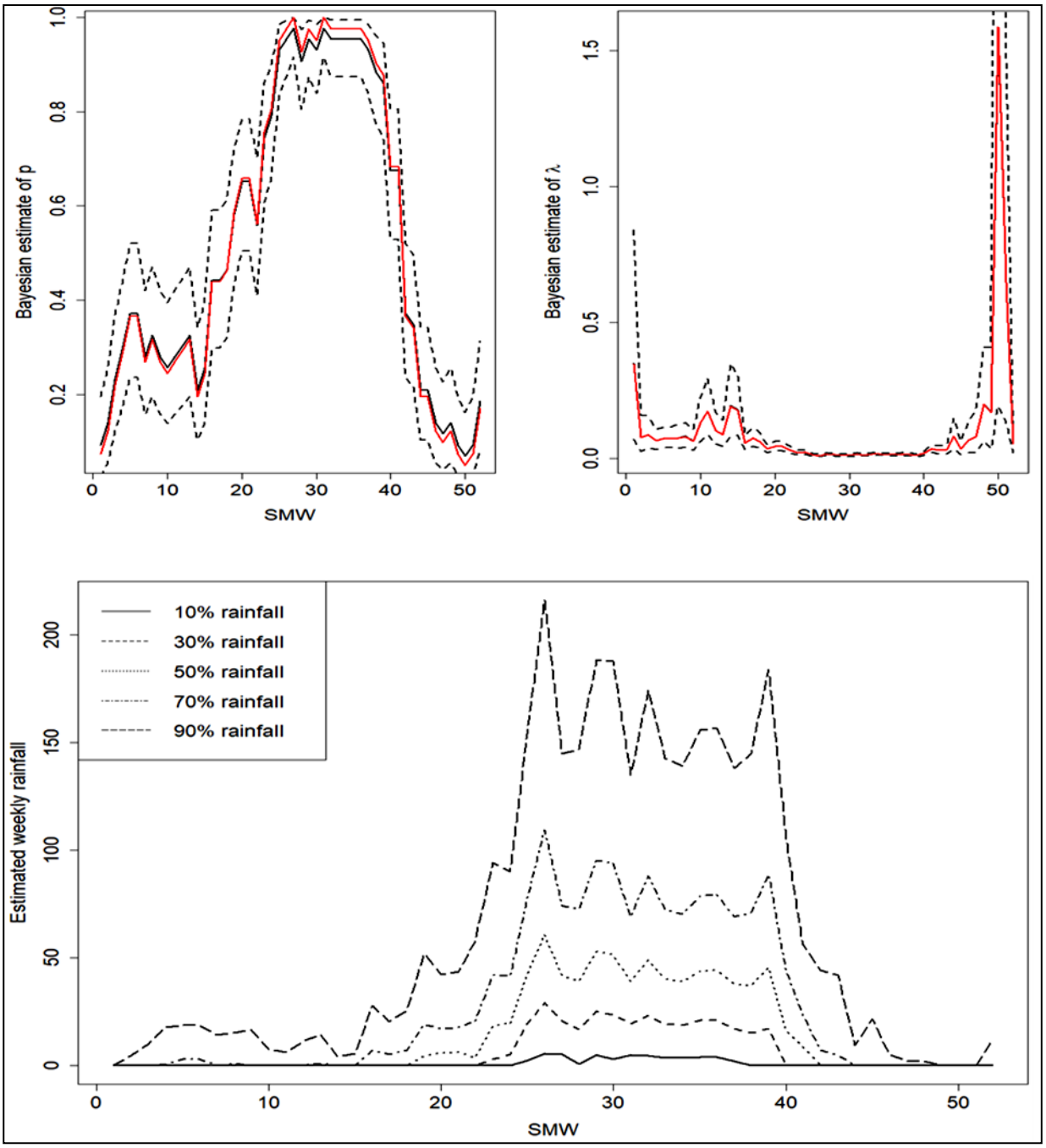

Fig. 2. Bayesian estimates of and for 52 weeks. The dashed lines show the 95% credible regions. The red lines show the corresponding maximum likelihood estimates (top panel). The model based estimates of weekly rainfall total at different probability levels (bottom panel)

The amount of estimated weekly rainfall totals based on our model at different probability levels (10%, 30%, 50%, 70%, 90%) based on the posterior predictive distribution are provided in the bottom panel of Fig. 2. At the 30% probability level, SMW 22-41 receive more than 200 mm rainfall with maximum for SMW 26 (109.37 mm). Starting from SMW 23, aggregate of SMW 23-26 shows a total of around 200 mm rainfall. At the 50 % probability level, SMW 25-39 receive more than 200 mm rainfall with maximum for SMW 26 (60.66 mm) again.

Starting from SMW 25, the aggregate of SMW 25-28 shows a total of around 175 mm rainfall.

4. Discussions

The amounts of weekly rainfall amounts at different probability levels in Figs. 1 and 2 match quite accurately indicating that the model fits the data well. The standard errors of the MLEs are zero for the weeks 27 and 31, which is never desirable for a statistical model. On the other hand, the standard deviation of the posterior distribution is non-zero (2.27×10^{-2}). Thus, the Bayesian estimates are more reliable and also more meaningful. The frequentist and Bayesian estimates of λ almost coincide due to our flat (high variance) prior choice for λ. In case some prior information about the parameters are available (through the elicitation of prior; refer to Ghosh *et al*., 2006), it can be incorporated and the estimates would be significantly different in that case. The estimates of λ remain less than 0.02 mm^{-1} for the weeks 25-40 indicating a high chance of a large amount of rainfall total during these weeks given that a positive amount of rainfall occurs and can be considered to be peak-monsoon period necessary for agricultural purposes though the chance of positive rainfall decreases sharply at the last weeks. A frequentist approach indicates certain positive amount of rainfall for the weeks 21 and 37 while the chance of rainfall is high but not certain in case of Bayesian approach. These weeks are mainly during the peak-monsoon period and for water requirement for kharif crops, there should be a minimal availability of irrigation as otherwise it may lead to crop failure.

A few studies about rainfall water management in the eastern plateau region are available. For example, Singh *et al*. (2009) have studied the frequencies of drought in nearby Ranchi region based on the departure of aridity index, the percentage ratio of the total annual water deficit to the total annual water need. They have categorized the study years 1970-2004 into moderate, large, severe and disastrous drought years and also concluded the weeks 26-39 to be the water surplus weeks which conforms with our findings as well. Singh *et al*. (2010) have performed a water balance study for the Ranchi region with concluding about high coefficient of variation of monthly rainfall. The average, maximum and minimum lengths of growing season are found to be 18, 12 and 28 weeks respectively and the authors have suggested cultivating short duration paddy variety Birsa Gora-101, maize, *i.e*., Devki, Ganga-11, Suran and kharif pulses. Availability of 12 weeks during the peak-monsoon period conforms to our results and hence short duration crops are preferable for Giridih as well. Kumari *et al*. (2014) have opted a preliminary data analysis of weekly (along with monthly, seasonal and annual basis) rainfall in the nearby Palamau region of Jharkhand for the years 1956-2011 without considering any statistical modelling and have concluded about most frequently observed droughts during weeks 23-26 and 37-40. They have suggested short duration, low water requiring but high value crops like maize, pulses, oilseeds to be opted in order to minimize the production risk. Our analysis shows that the amount of weekly rainfall has steep upward and downward trends for the weeks 23-26 and 37-40 respectively which conforms with the findings of Kumari *et al*. (2014) as well.The cultivated area of Jharkhand is about 1.8 million ha and only 9.3% of these areas have irrigation facilities (Source: www.icar.org.in). Some major concerns are- drought in uplands, low soil fertility and low-coverage of high yielding varieties. As the water management situation in Giridih conforms to other studies at the nearby regions, we also suggest drought-tolerant short duration varieties of rice like Vandana, Anjali, Birsa dhan 109; 110, maize, pulses etc.for cultivation.

5. Conclusions

We propose a mixture probability model of two distributions, *i.e*., degenerate at zero and a one-parameter exponential distribution for each week which takes care of the dry spells. Rather than the frequentist approach of finding MLEs, we recommend the Bayesian paradigm for its many-fold advantages - we believe the paradigm to be promising enough for modern statistical scientists, particularly for statistical climatologists (Clark and Gelfand, 2006).

Banik *et al*. (2002) analyzed weekly rainfall data of Giridih using a two-stage Markov chain and the drought index was calculated to be 0.16 indicating severe drought-proneness. Hence, a proper statistical modelling is very important for a proper crop planning. The MLEs of the zero-inflation parameters are just the proportions of wet weeks while the Bayesian estimates are not and always have positive standard deviations. The estimation of the zero-inflation parameter using MLE leads to zero standard deviation for the weeks which are wet for all the years considered and hence, to a less meaningful interpretation and possibly erroneous crop planning. In this situation, a frequentist approach indicates certain positive amount of rainfall while in case of Bayesian approach, the chance of rainfall is high but not certain, a more realistic situation. As these weeks are mainly during the peak-monsoon period and water requirement for kharif crops is necessary, there should be a minimal availability of irrigation facilities even on these weeks according to our analysis while a MLE approach says the irrigation requirements are not necessary for these weeks. Our study indicates that a complete ignorance about the irrigation facilities even during the peak-monsoon may lead to crop

failure. The reservoirs of the DVC can be considered as possible sources of water during these weeks in case there is no rainfall.

From the application point of view, we have provided the estimates of rainfall at different probability levels and have drawn conclusion regarding the cropping times and the choice of varieties. Aggregate rainfall amount of SMW 23-26 is sufficient before sowing or transplanting on SMW 27 and it is fine for long/medium-duration rice cropping. The aggregate rainfall of SMW 25-28 is fine before sowing or transplanting on SMW 28 or 29 and it is fine for short-duration rice cropping. Thus, opting a short-duration rice cropping has less chance of crop failure due to scarcity of rain water. The aggregate of weekly rainfall totals for the weeks in late September and early October is sufficient for the second crop (winter crop). It can be grown in the eastern plateau area, but proper irrigation is required as the rainfall amount drops sharply after that period. Hence, the reservoirs of the DVC can be considered as sources of water during the second week of October in case there is no rainfall.

For avoiding computational complexities, here we have confined ourselves to the zero-inflated exponential model only. It is possible that some other zero-inflated model (e.g. zero-inflated gamma / log-normal / Weibull) would provide better fit than our model for a set of weeks. The corresponding Bayesian inference can be carried out via Markov Chain Monte Carlo techniques (for more details, refer to Ghosh *et al*., 2006). We reserve these as part of our future endeavour.

Acknowledgement

The authors would like to thank an anonymous reviewer whose suggestions have improved the paper in several ways. The contents and views expressed in this research paper/article are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Banik, P., Mandal, A. and Rahman, M. S., 2002, "Markov chain analysis of weekly rainfall data in determining drought-proneness", *Discrete Dynamics in Nature and Society*, **7**, 231-239.

Burgueño, A., Martínez, M. D., Lana, X. and Serra, C., 2005, "Statistical distributions if the daily rainfall regime in Catalonia (Northeastern Spain) for the years 1950-2000", *Int. J. Climatol*., **25**, 1381-1403.

Burgueño, A., Martínez, M. D., Serra, C. and Lana, X., 2011, "Statistical distributions of daily rainfall regime in Europe for the period 1951-2000", *Theor. Appl. Climatol*., **102**, 213-226.

Chand, R., Singh, U. P., Singh, Y. P., Siddique, L. A. and Kore, P. A., 2011, "Analysis of weekly rainfall of different period during rainy season over Safdarjung airport of Delhi for 20th century - A study on trend, decile and decadal analysis", *Mausam*, **62**, 2, 197-204.

Clark, J. S. and Gelfand, A. E., 2006, "A future for models and data in environmental science", *TRENDS in Ecology and Evolution*, **21**, 7, 375-380.

Fitzgerald, D. L., 2005, "Analysis of extreme rainfall using the log logistic distribution", *Stoch. Environ. Res. Risk Assess*., **19**, 249-257.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B., 2014, "Bayesian data analysis", Volume 2, Chapman and Hall, CRC Boca Raton, Florida, USA.

Ghosh, J. K., Delampady, M. and Samanta, T., 2006, "An Introduction to Bayesian Analysis: Theory and Methods", Springer.

Hanson, L. S. and Vogel, R., 2008, "The probability distribution of daily rainfall in the United States", Conference proceeding paper, World Environmental and Water Resources Congress.

Hazra, A., Bhattacharya, S. and Banik, P. 2014, "Modelling Nakshatra-wise rainfall data of the eastern plateau region of India". *Mausam*, **65**, 2, 264-270.

Husak, G. J., Michaelsen, J. and Funk, C., 2007, "Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications", *Int. J. Climatol.*, **27**, 935-944.

Kale, B. K. and Muralidharan, K., 1999, "Optimal estimating equations in mixture distributions accommodating instantaneous or early failures", *J. Indian Statistical Association*, **38**, 317-329.

Khudri, M. M. and Sadia, F., 2013, "Determination of the Best Fit Probability Distribution for Annual Extreme Precipitation in Bangladesh", *European J. Scien. Res*., **103**, 391-404.

Krishnamoorthya, K. and León-Novelo, L., 2014, "Small sample inference for gamma parameters: one-sample and two-sample problems", *Environmetrics*, **25**, 107-126.

Kumar, N., Patel, S. S., Chalodia, A. L., Vadaviya, O. U., Pandya, H. R., Pisal, R. R., Dakhore, K. K. and Patel, M. L., 2015, "Markov chain and incomplete Gamma distribution analysis of weekly rainfall over Navsari region of south Gujarat", *Mausam*, **66**, 4, 751-760.

Kumari, P., Ojha, R. K., Wadood, A. and Kumar, R., 2014, "Rainfall and drought characteristics for crop planning in Palamau region of Jharkhand", *Mausam*, **65**, 1, 67-72.

Kwaku, X. S. and Duke, O., 2007, "Characterization and frequency analysis of one day annual maximum and two consecutive days maximum rainfall of Accra, Ghana", *ARPN Journal of Engineering and Applied Sciences*, **2**, 5, 27-31.

Lana, X., Serra, C., Casas-Castillo, M. C., Rodríguez-Solà, R., Redaño, A. and Burgueño, A., 2017, "Rainfall intensity patterns derived from the urban network of Barcelona (NE Spain)", *Theor. Appl. Climatol*., published online.

Liang, L., Zhao, L., Gong, Y., Tian, F. and Wang, Z., 2012, "Probability distribution of summer daily precipitation in the Huaihe basin of China based on Gamma distribution", *Acta Meteorol. Sin*., **26**,72-84.

Mayooran, T. and Laheetharan, A., 2014, "The Statistical Distribution of Annual Maximum Rainfall in Colombo District", *Sri Lankan Journal of Applied Statistics*, **15**, 107-130.

Muralidharan, K., 1999, "Tests for the mixing proportion in the mixture of a degenerate and exponential distribution", *J. Indian Stat. Assn.*, **37**, 2, 105-119.

Muralidharan, K. and Kale, B. K., 2002, "Modified gamma distribution with singularity at zero", *Comm. Statist.-Simulation and Computations*, **31**, 1, 143-158.

Muralidharan, K. and Lathika, P., 2005, "Statistical modelling of rainfall data using modified Weibull distribution", *Mausam*, **56**, 4, 765-770.

Muralidharan, K. and Pratima, B., 2017, "Analysis of lifetime model with discrete mass at zero and one", *J. Stat. Theory Pract.*, published online.

Sharda, V. N. and Das, P. K., 2005, "Modelling weekly rainfall data for crop planning in a sub-humid climate of India", *Agricultural Water Management*, **76**, 120-138.

Sharma, M. A. and Singh, J. B., 2010, "Use of probability distribution in rainfall analysis", *New York Science Journal*, **3**, 9, 40-49.

Singh, P. K., Rathore, L. S., Singh, K. K., Baxla, A. K. and Athiyamani, B., 2009, "Incomplete Gamma distribution of rainfall for sustainable crop production strategies at Palampur, Himachal Pradesh", *Mausam*, **60**, 1, 73-80.

Singh, P. K., Rathore, L. S., Athiyamani, B., Singh, K. K., Baxla, A. K., Kumar, A. and Bhargava, A. K., 2009, "Frequencies of drought at Ranchi regions, Jharkhand", *Mausam*, **60**, 4, 455-460.

Singh, P. K., Bhargava, A. K., Mitra, V., Prasad, A. and Jayapalan, M., 2010, "Water balance studies for the crop planning in Ranchi, Jharkhand", *Mausam*, **61**, 2, 233-238.

Syversveen, A. R., 1998, "Non-informative Bayesian priors-interpretation and problems with construction and applications", *Technical report*, Department of Mathematical Sciences, NTNU, Trondheim.

Taewichit, C., Soni, P., Salokheb, V. M. and Jayasuriya, H. P. W., 2013,"Optimal stochastic multi-states first-order Markov chain parameters for synthesizing daily rainfall data using multi-objective differential evolution in Thailand", *Meteorol. Appl.*, **20**, 20-31.

Velarde, L. G. C., Migon, H. S. and Pereira, B. B., 2004, "Space-time modelling of rainfall data", *Environmetrics*, **15**, 561-576.

APPENDIX

Posterior distribution of parameters : For a particular week, rainfall Y for n years can be considered as a sample $Y_1, Y_2, \ldots, Y_n \sim iidG(y|\, p, \lambda)$. The joint distribution of the observed variables $Y_1, Y_2, \ldots, Y_n$, is given by :

$$P[Y_1, Y_2, \ldots, Y_n | p, \lambda] = (1-p)^{\sum_{i=1}^{n} I(Y_i=0)} p^{n-\sum_{i=1}^{n} I(Y_i=0)} \prod_{i=1}^{n} \left(\lambda e^{-\lambda Y_i}\right)^{1-I(Y_i=0)}.$$

The prior density is given by :

$$P[p, \lambda] = P[p]P[\lambda] = \frac{\Gamma(\alpha_p + \beta_p)}{\Gamma(\alpha_p)\Gamma(\beta_p)} p^{\alpha_p - 1}(1-p)^{\beta_p - 1} \frac{{\beta_\lambda}^{\alpha_\lambda}}{\Gamma(\alpha_\lambda)} \lambda^{\alpha_\lambda - 1} e^{-\beta_\lambda \lambda}$$

Thus, the posterior distribution of p is

$$p|Y_1, Y_2, \ldots, Y_n \sim Beta(n - \sum_{i=1}^{n} I(Y_i = 0) + \alpha_p, \sum_{i=1}^{n} I(Y_i = 0) + \beta_p)$$

The posterior distribution of λ is

$$\boldsymbol{\lambda}|Y_1, Y_2, \ldots, Y_n \sim Gamma(n - \sum_{i=1}^{n} I(Y_i = 0) + \alpha_\lambda, \sum_{i=1}^{n} Y_i + \beta_\lambda)$$

For estimates of the parameters p and λ, we can consider the posterior mean or the posterior median. For a measure of dispersion, we can consider posterior variance or posterior standard deviation or some upper and lower quantiles, for example, 2.5% and 97.5% quantiles.

Posterior predictive distribution : For a future observation Y^*, the posterior predictive distribution is given by :

$$P[Y^*|Y_1, Y_2, \ldots, Y_n] = \int P[Y^*|p, \lambda] P[p|Y_1, Y_2, \ldots, Y_n] P[\lambda|Y_1, Y_2, \ldots, Y_n]\, dp\, d\lambda.$$

After integrating, we have Y^* distributed as follows

$$Y^* = \begin{cases} Zw.p. & q \\ 0 \quad w.p. & 1-q \end{cases}$$

where, $q = \frac{n-\sum_{i=1}^{n} I(Y_i=0)+\alpha_p}{n+\alpha_p+\beta_p}$. In comparison, we see that in case of frequentist approach, the maximum likelihood estimate turns out to be $\frac{n-\sum_{i=1}^{n} I(Y_i=0)}{n}$ [See Hazra et al. (2014)]. Now, the variable Zfollows the *Lomax* distribution (Pareto Type II distribution) with shape parameter $n - \sum_{i=1}^{n} I(Y_i = 0) + \alpha_\lambda$ and scale parameter$\sum_{i=1}^{n} Y_i + \beta_\lambda$.

Software (R code)

```
ZIE_Bayes<- function(Y, alpha_p = 1, beta_p = 1, alpha_lambda = 0.01, beta_lambda = 0.01){
Y <- Y[is.na(Y)!= 1]
nopar_est<- as.vector(quantile(Y, seq(0.9, 0.1, -0.2)))
  names(nopar_est) <- c("10% rainfall", "30% rainfall",
                        "50% rainfall", "70% rainfall", "90% rainfall")
mar.default<- c(5,4,4,2) + 0.1
par(mar = mar.default + c(0, 1, 0, 0))
  p <- ppoints(100); q <- quantile(Y[Y > 0], p = p);
  plot(qexp(p), q, main = "", xlab = "Theoretical Quantiles",
ylab = "Sample Quantiles", cex.lab = 2, cex.axis = 2)
qqline(q, distribution = qexp, lty = 2)
# prior: p ~Beta(alpha_p, beta_p), lambda ~ Gamma(alpha_lambda, beta_lambda)
zero_count<- sum(Y == 0); totals <- sum(Y);
posmean_p<- (length(Y) - zero_count + alpha_p) / (length(Y) + alpha_p + beta_p)
  names(posmean_p) <- "Bayesian estimate of p"
p_credible<-
qbeta(c(0.025, 0.975), length(Y) - zero_count+ alpha_p, zero_count + beta_p)
  names(p_credible) <- "95% credible region of p"
posmean_lambda<- (length(Y) - zero_count + alpha_lambda) / (totals + beta_lambda)
  names(posmean_lambda) <- "Bayesian estimate of lambda"
lambda_credible<- qgamma(c(0.025, 0.975),
          length(Y) - zero_count + alpha_lambda, totals + beta_lambda)
  names(lambda_credible) <- "95% credible region of lambda"
install.packages("VGAM");library("VGAM");
qziLomax<- function(q, p_hat, scale_par, shape_par){
if(q <= 1 - p_hat){
     out <- 0}else{out <- qlomax((p_hat + q - 1) / p_hat,
```

```
                scale = scale_par, shape3.q = shape_par)}
   out}
par_est<- sapply(seq(0.9, 0.1, -0.2),
        function(qq){qziLomax(qq, posmean_p, totals + beta_lambda,
                        length(Y) - zero_count + alpha_lambda)})
 names(par_est) <- c("10% rainfall", "30% rainfall",
                "50% rainfall", "70% rainfall", "90% rainfall")
 out <- list(nonparametric_estimate = nopar_est,
Bayesian_estimate_p = posmean_p, credible_p = p_credible,
Bayesian_estimate_lambda = posmean_lambda,
credible_lambda = lambda_credible,
model_based_estimate = par_est)
 out}
```