Modelling the entire range of daily precipitation using mixture distributions

Hyunwoo Rho

1. Introduction

To measure and predict the precipitation are the most oldest concerns of mankind. It is an essential factor in agriculture, forest management, hydrology, and preparing for disasters like flood and drought. For generating stochastic precipitaion model, there are some feasible time intervals as a base period of modelling, but taking a daily basis is considerably natural Richardson (1981). Constructing daily stochastic precipitaion model can be extended to precipitation analytics like monthly or yearly basis approach and spatial analysis.

Many of previous studies tried to model daily rainfall data through parametric approaches e.g. Ison *et al.* (1971); Mielke Jr and Johnson (1973); Richardson (1981); Stern and Coe (1984) and also non-parameteric approaches e.g. Sharma and Lall (1999); Harrold *et al.* (2003)Along some studies in hydrology community, the evidence of heavy-tailed phenomenon on the distribution of high precipitation amount Koutsoyiannis (2004). In this perspective, various methods were introduced more focused on tail behavior of daily precipitation Furrer and Katz (2008); Li *et al.* (2012); Papalexiou and Koutsoyiannis (2012); Papalexiou *et al.* (2013).

In this paper, we compare various type of parametric models and introdcue a particular type of mixture distribution called phase-type distribution. Through this model we tried to fit the whole range of continuous daily precipitation.

2. Data Sets

The main data set used to analyse is United States Historical Climatology Network(USHCN) Daily data set. Raw data set contains daily record of precipitation, snowfall, snow depth, maximum temperature, minumum temperature, and information about flag. Among 48 files of each sontiguous states, According as Li *et al.* (2012), Texas was selected. For following their work, the same data selection criteria were adopted which are all nonzero precipation of 1940 to 2009 without taking care of missings.

Additionally, Daily Global Historical Climatology Network(GHCN-DAILY) was used to do extra analyses Papalexiou and Koutsoyiannis (2012); Papalexiou *et al.* (2013) which has identical data format with USHCN. This set contains about 100 thousand stations and also encompasses USHCN as its subset. By filtering stations along Papalexiou *et al.* (2013), (a) record length of over 50 years, (b) percentage of missing values less than 20%, data assigned with suspicious

"quality flags" less than 0.1%. The screen values of quality flags are two, one with "G"(failed gap check), and another with "X"(failed bound check). For more information about data set, see Menne *et al.* (2012).

Finally we handled 49 stations in USHCN-Texas data, and 19328 stations in GHCN data after filtering.

3. Existing models

3.1. One component models

Fundamental characteristrics of daily precipitation data are described as follows; non-negative, continuous except at the spike on zero, right-skewed. Assuming we only focus on the continuous part of the data, such distributions include exponential Todorovic and Woolhiser (1975), gamma Ison *et al.* (1971); Wilks (1999); Schoof *et al.* (2010), and kappa Mielke Jr and Johnson (1973), and so on. Let X denote the nonzero daily precipitation amount, then each models can be presented as their probabilty density functions as below.

$$f(x;\lambda) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x \geq 0, \quad \lambda > 0, \tag{1}$$

$$f(x;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x \geq 0, \quad \theta > 0, \tag{2}$$

$$f(x;\alpha,\beta,\theta) = \frac{\alpha\theta}{\beta}(\frac{x}{\beta})^{\theta-1}[\alpha + (\frac{x}{\beta})^{\alpha\theta}]^{-\frac{\alpha+1}{\alpha}}, \quad x > 0, \quad \alpha,\beta,\theta > 0, \tag{3}$$

Other distributions like skewed-normal Wan *et al.* (2005), truncated power of normal distribution Bardossy and Plate (1992), and only for the tail of data, genralized Pareto distribution) were introduced. To avoid crowded model comparison and to aim our goal fitting the whole range of the data, we are going to consider only three models of one component model above. Additionally, each of distributions represent one, two, and three parameter model respectively.

3.2. Two-componenet models

Most general way to combine two distributions is to using mixture distribution. Mixed exponenetial could be a typical case.

$$f(x;\omega,\lambda_1,\lambda_2) = \frac{\omega}{\lambda_1} e^{-x/\lambda_1} + \frac{1-\omega}{\lambda_2} e^{-x/\lambda_2}, \quad x > 0, \quad \lambda_1,\lambda_2 > 0, \quad \omega \in [0,1], \tag{4}$$

Alternatively, hybrid model could be considered, which is a distribution combining two different distributions at particular threshold. These hybrid approaches have more attention on

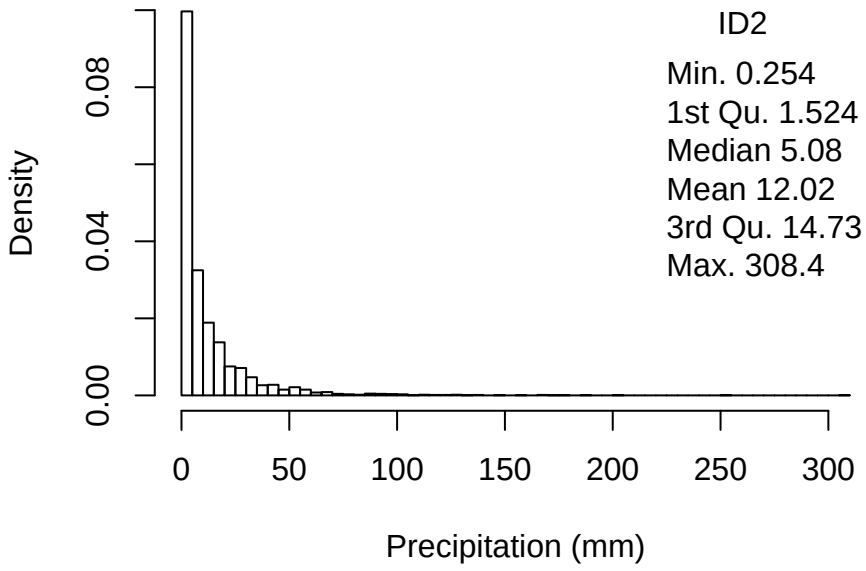

Figure 1: sample histrogram

the extreme value of daily rainfall by adopting heavy-tailed distribution in the tail part, like generalized Pareto, because some researches find out it has heavy tail behavior Koutsoyiannis (2004). From the precedent studies, the performance of hybrid models defeat the one of mixture models in genereal Li *et al.* (2012); Furrer and Katz (2008). Hence, we mainly cover the hybrid models to be compared rather than other mixture distributions.

$$f(x;\alpha,\beta,\xi,\sigma,\theta) = f_{gamma}(x;\alpha,\beta)I(x \le \theta) + [1 - F_{gamma}(\theta;\alpha,\beta)]f_{GP}(x;\xi,\sigma,\theta)I(x > \theta) \quad (5)$$

At the threshold θ, to make it continuous, a constraint $f(\theta-) = f(\theta+)$ is needed, which yields that the scale parameter σ of generalized Pareto distribution is expressed as the reciprocal of the gamma hazard funciton,

$$\sigma = \frac{1 - F_{gamma}(\theta;\alpha,\beta)}{f_{gamma}(\theta;\alpha,\beta)} \quad (6)$$

Thus the number of paramters in this model reduces from five to four. But still the selection problem of the threshold *theta* remains without any guaranteed selection method. To pass away such huddle, Li *et al.* (2012) sugested another type of hybrid model, combination of exponential and generalized Pareto distributions.

$$f(x;\lambda,\xi,\sigma,\theta) = \frac{1}{1+F_{exp}(\theta;\lambda)}[f_{exp}(x;\lambda)I(x \le \theta) + f_{GP}(x;\xi,\sigma,\theta)I(x > \theta)] \tag{7}$$

With same constraint above, $f(\theta-) = f(\theta+)$, the number of parameters can be reduced again.

$$\theta = -\lambda \ln \frac{\lambda}{\sigma} \tag{8}$$

Especially this model outstands by deriving the threshold θ analytically.

4. Phase-type distribution class

5. USHCN Texas daily precipitation data set

From Li *et al.* (2012), various kind of models were used to fit each stations of USHCN Texas daily precipitation data set. We chose five models among them to compare with PH model; Exponential, Gamma, kappa, hybrid of gamma and generalized Pareto(GGP), hybrid of exponential and generalized Pareto(EGP).

For GGP model, its performance is affected by the choice of threshold value. Too large threshold would estimate its tail short by taking large emphasis on gamma distribution. Otherwise too small threshold set large emphasis on genralized Pareto and estimate its tail heavy. Thus we set threshold value used in GGP model moderately as 60% quantile of each sample.

Overall, single component models; exponential, gamma, kappa, show inferior performances than other multi component models. With particular sample station(ID2), Exponential and gamma model underestimate tail thickness, besides kappa model overestimates it. Even with any other stations, gamma tends to exhibit short tail and kappa tends to exhibit heavy tail.

With this sample, GGP, EGP, and PH models show visually similar performances. Since the ultimate goal is to find general method generating stochastic daily precipitation model, it it essential to compare each methods with differenct cases. Furthermore, not only with qq plot, AIC Akaike (1974) and BIC Schwarz *et al.* (1978) were used to do numerical comparison.

6. World data

In the research of Papalexiou and Koutsoyiannis (2012), wordwide GHCN data set was analyzed to figure out its shape of empirical distribution and tail behavior. To capture the shape characteristics of each sample, L-moments ratio was used. A typical L-moments ratio diagram is made with L-skewness versus L-kurtosis, but with precipitation data, but we prefer L-varaiation versus L-skewness as Papalexiou and Koutsoyiannis (2012) done before.

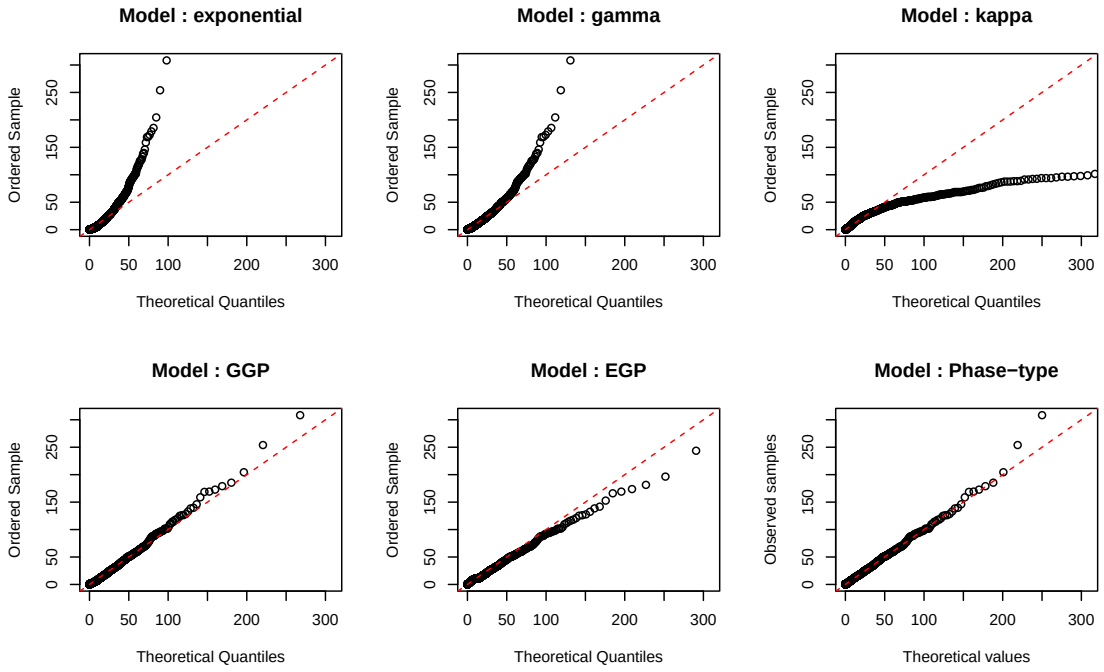

Figure 2: QQ plots of ID2 modelled by each distributions

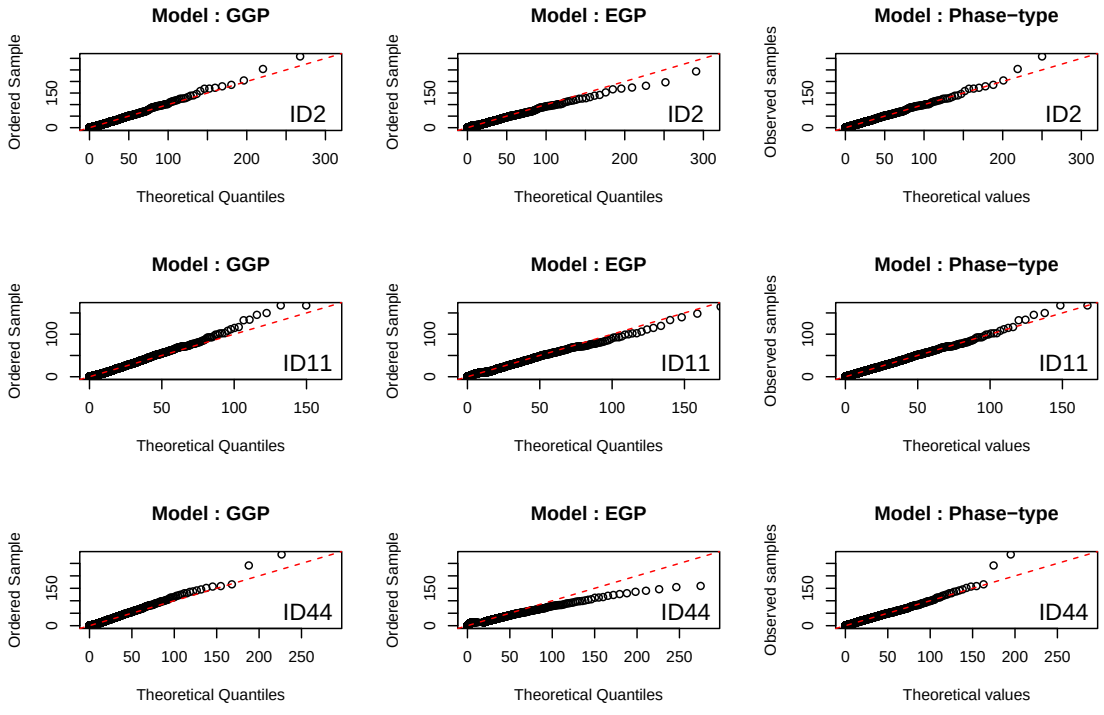

Figure 3: QQ plots for 3 different stations modelled by GGP distribution(lest), EGP distribution(middle), and the PH distribution(right)

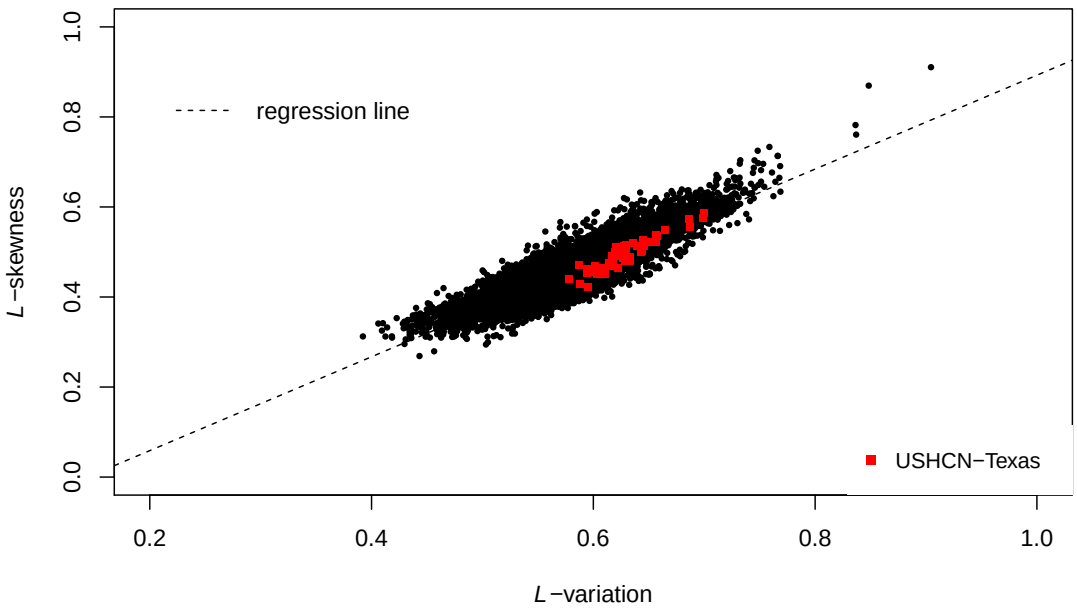

Figure 4: L-variation vs. L-skewness plot

7. Conclusion

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**(6), 716–723.

Bardossy, A. and Plate, E. J. (1992). Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, **28**(5), 1247–1259.

Furrer, E. M. and Katz, R. W. (2008). Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, **44**(12).

Harrold, T. I., Sharma, A., and Sheather, S. J. (2003). A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resources Research*, **39**(12).

Ison, N., Feyerherm, A., and Bark, L. D. (1971). Wet period precipitation and the gamma distribution. *Journal of Applied Meteorology*, **10**(4), 658–665.

Koutsoyiannis, D. (2004). Statistics of extremes and estimation of extreme rainfall: I. theoretical investigation/statistiques de valeurs extrêmes et estimation de précipitations extrêmes: I. recherche théorique. *Hydrological sciences journal*, **49**(4).

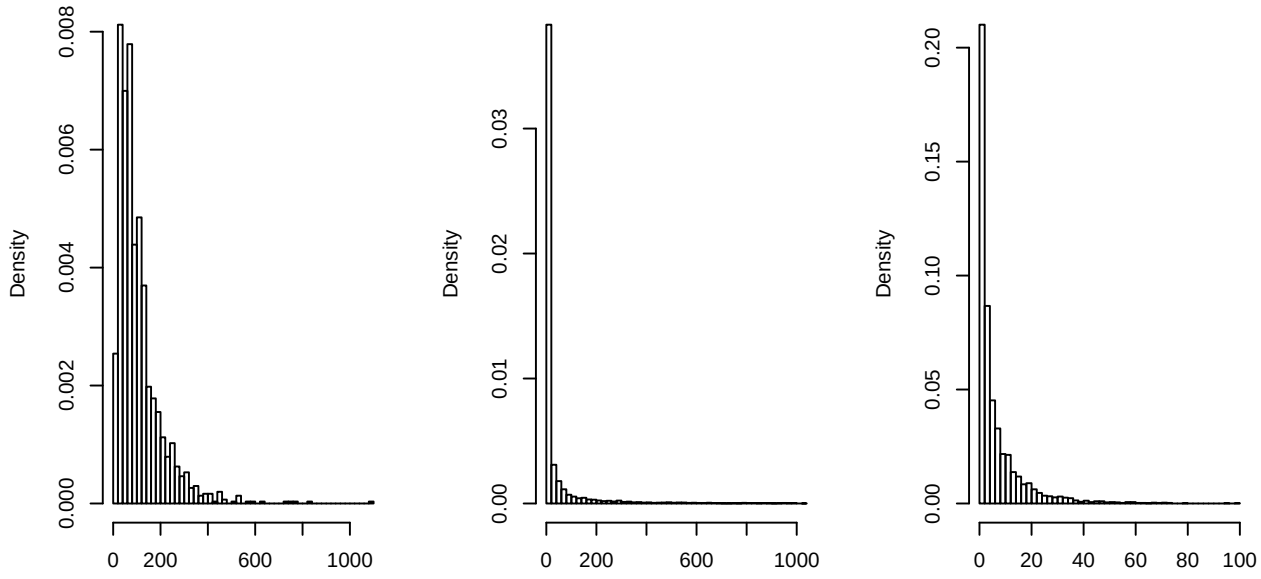

Figure 5: Histograms of bell-shape(right), J-shape(middle), and one of USHCN Texas(right) data

Li, C., Singh, V. P., and Mishra, A. K. (2012). Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water resources research*, **48**(3).

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, **29**(7), 897–910.

Mielke Jr, P. W. and Johnson, E. S. (1973). Three-parameter kappa distribution maximum likelihood estimates and likelihood ratio tests. *Monthly Weather Review*, **101**(9), 701–707.

Papalexiou, S., Koutsoyiannis, D., and Makropoulos, C. (2013). How extreme is extreme? an assessment of daily rainfall distribution tails. *Hydrology and Earth System Sciences*, **17**(2), 851–862.

Papalexiou, S. M. and Koutsoyiannis, D. (2012). Entropy based derivation of probability distributions: A case study to daily rainfall. *Advances in Water Resources*, **45**, 51–57.

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, **17**(1), 182–190.

Schoof, J., Pryor, S., and Surprenant, J. (2010). Development of daily precipitation projections for the united states based on probabilistic downscaling. *Journal of Geophysical Research: Atmospheres*, **115**(D13).

Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Sharma, A. and Lall, U. (1999). A nonparametric approach for daily rainfall simulation. *Mathematics and Computers in Simulation*, **48**(4), 361–371.

Stern, R. and Coe, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, pages 1–34.

Todorovic, P. and Woolhiser, D. A. (1975). A stochastic model of n-day precipitation. *Journal of Applied Meteorology*, **14**(1), 17–24.

Wan, H., Zhang, X., and Barrow, E. M. (2005). Stochastic modelling of daily precipitation for canada. *Atmosphere-Ocean*, **43**(1), 23–32.

Wilks, D. S. (1999). Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology*, **93**(3), 153–169.