Modelling the Entire Range of Daily Precipitation Using Mixture Distribution

Hyunwoo Rho

Dept. Applied Statitstics, Yonsei University

2016. 04. 12

Introduction

- Measuring and predicting the precipitation are critical issues in many fields; e.g. agriculture, hydrology, and forestry.
- To model the precipitation amount, the most common unit interval is daily basis.
- There are many studies on daily precipitation both parametric and non-parametric, but parametric approaches especially have advantage on describing extreme part which has rare observation.
- Main goal of this study is to find precipitation model such that
    1. fits the entire range of the daily precipitation data
    2. performs well at extreme part of data
    3. generally accepted with any precipitation characteristics
- Some features of daily precipitation data
    - Non-negative
    - Two distict part: dry days and rainy days
    - Right skewness on continuous part

U.S. Historical Climatology Network(USHCN), Texas

- Used by [Li et al., 2012].
- 49 stations across the Texas.

	ID	Label	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	410120	ID1	0.254	1.524	5.334	11.27	14.22	737.9
2	410144	ID2	0.254	1.524	5.08	12.02	14.73	308.4
3	410174	ID3	0.254	0.762	2.794	6.544	8.128	99.57
4	410493	ID4	0.254	1.524	4.826	10.4	13.46	179.1
5	410498	ID5	0.254	1.524	3.81	7.539	9.652	104.9
6	410639	ID6	0.254	1.27	4.064	10.74	12.7	269.5
7	410832	ID7	0.254	1.524	4.826	11.39	14.48	443.7
8	410902	ID8	0.254	1.27	4.064	10.89	12.95	226.8
9	411000	ID9	0.254	1.524	4.572	9.268	12.19	114.3
10	411048	ID10	0.254	1.524	4.572	11.66	14.73	263.7
11	411138	ID11	0.254	1.778	5.842	11.87	15.49	167.6
12			...	...	...	...	...	...

Table: Summary of USHCN Texas data set

Global Historical Climatology Network(GHCN)

- Used by [Papalexiou and Koutsoyiannis, 2012, Papalexiou et al., 2013].
- There exist around 100 million stations around the world, but we used only 19328 stations after data cleansing with following criteria [Papalexiou et al., 2013].[1]
    - Stations having record over 50 years
    - Missing days less than 20
    - Suspicious quaily of flags less than 0.1

[1] For more infomation on data, see [Menne et al., 2012]

Station Sample: Alice, TX(ID2)

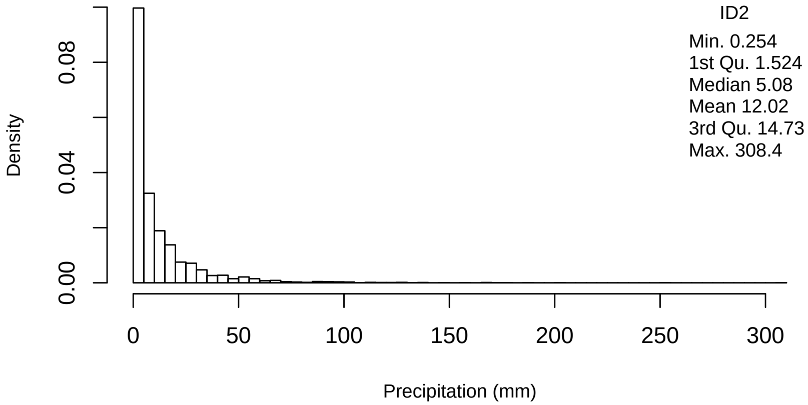


Figure: Sample histrogram of USHCN Texas data set

Existing Models : Single-Component Models

- Exponential [Todorovic and Woolhiser, 1975]

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

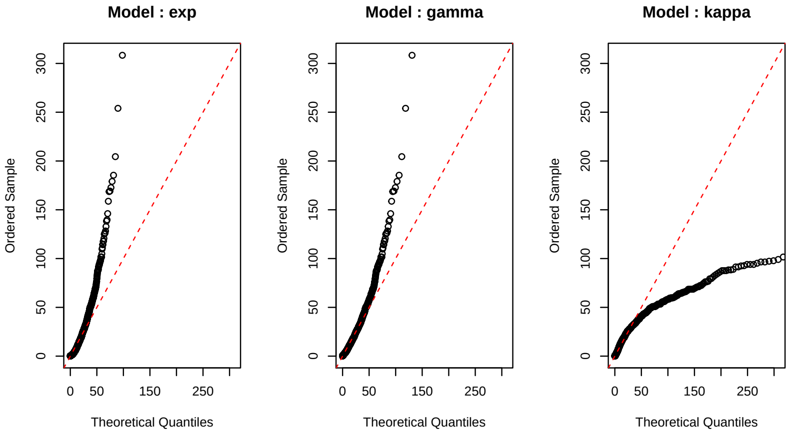
- Gamma [Ison et al., 1971, Wilks, 1999, Schoof et al., 2010]

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

- Kappa [Mielke Jr and Johnson, 1973]

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

Existing Models : Single-Component Models



Existing Models : Multiple-Component Models

- Hybrid models were introduced to reflect heavily distributed tail of daily precipitation data
- Hybrid of Gamma and Generalized Pareto(GGP) [Furrer and Katz, 2008]

$$
\begin{aligned}
f(x;\alpha,\beta,\xi,\sigma,\theta) = f_{gam}(x;\alpha,\beta)I(x \leq \theta) + \\
[1 - F_{gam}(\theta;\alpha,\beta)]f_{GP}(x;\xi,\sigma,\theta)I(x > \theta)
\end{aligned}
\tag{4}
$$

- Hybrid of Exponential and Generalized Pareto(EGP) [Li et al., 2012]

$$
f(x;\lambda,\xi,\sigma,\theta) = \frac{1}{1 + F_{exp}(\theta;\lambda)}[f_{exp}(x;\lambda)I(x \leq \theta) + f_{GP}(x;\xi,\sigma,\theta)I(x > \theta)] \tag{5}
$$

EGP model has an advantage on avoiding threshold selection problem.

Existing Models : Multiple-Component Models

- Number of parameters in hybrid models are 5 and 4 respecitvely, but those reduce to 4 and 3 due to continuity constraint on threshold θ.
- for GGP model, $f(\theta-) = f(\theta+)$ derives

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

- for EGP model, $f(\theta-) = f(\theta+)$ derives

$$\theta = -\lambda \ln \frac{\lambda}{\sigma} \tag{7}$$

Phase-type Distribution(PHD) [2]

- The phase-type distribution is defined as the distribution of the time until absorption in a continuous Markov chain with absorbing state, denoted by $PH(\alpha, T)$.
- Consider Markov chain with a state space $\{1, \cdots, p, p+1\}$ with initial vector $(\alpha, 0)$, $\alpha\mathbf{1} = 1$, and also with transition rate matrix,

$$Q = \begin{pmatrix} T & t \\ \mathbf{0} & 0 \end{pmatrix}, \quad t = -T\mathbf{1}, \tag{8}$$

- Where T is a p x p square matrix and $p+1$ state indicates aborbing state.
- Its distribution function and density function is given as,

$$F(x) = 1 - \alpha e^{Tx}\mathbf{1}, \quad x > 0, \tag{9}$$

$$f(x) = \alpha e^{Tx} t, \quad x > 0 \tag{10}$$

- Its quantile function cannot be obtained in closed form.

[2] introduced by [Neuts, 1974]

Phase-type Distribution(PHD)

- General phase-type distribution with m phases includes $m^2 + m$ parameters. (m^2 from T, m from α)
- Phase-type distribution includes several types of distributions with parameters constraints and we could control model complexity by adopting particular limited distribution.
  1. Exponenetial and exponential mixture
  2. Erlang and Erlang mixture
  3. Coxian and generalized Coxian
  4. ..., etc.
- EM algorithm for phase-type distribution called EMphat introduced by [Asmussen et al., 1996] is used for parameter estimation.[3]

[3] C script for EMpht is available at
http://home.math.au.dk/asmus/pspapers.html

Special Cases of PHD

- Exponential
  The simplest case of phase-type distribution is the exponential distibution with PHD parameters $T = -\lambda, \alpha = 1$.
- Erlang (m parameters)

$$T = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ 0 & -\lambda & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & & -\lambda \end{bmatrix}, \quad \alpha = (1, 0, \ldots, 0),$$

- Coxian ($2m - 1$ parameters)

$$T = \begin{bmatrix} -\lambda_1 & p_1\lambda & 0 & \cdots & 0 \\ 0 & -\lambda_2 & p_2\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & & -\lambda_m \end{bmatrix}, \quad \alpha = (1, 0, \ldots, 0),$$

Model Comparison

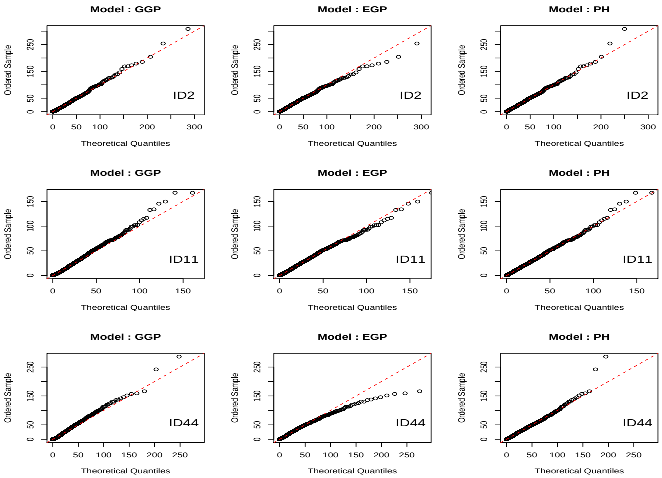


Figure: QQ plots of 3 different stations modelled by GGP, EGP, and PH model

Model Comparison

	ID	recordLen	AIC.GGP	AIC.EGP	AIC.PH
1	ID5	2911	17447.56	17333.05	17315.94
2	ID12	1516	10744.76	10646.41	10646.93
3	ID13	5687	41165.68	40892.36	40899.23
4	ID16	4401	27430.98	27077.82	27043.17
5	ID29	5288	34285.82	33730.29	33727.67
6	ID33	6212	44489.86	44091.62	44086.12
7	ID48	4673	31643.74	31243.72	31236.74
8	ID49	5119	34782.47	34398.06	34403.24

Table: Sample AIC Table

Model Comparison

- For numerical comparison, AIC is used which can reflect the model complexity.
- Number of parameters of each models
  - GGP : 4
  - EGP : 3
  - PH (restricted) : 5

EGP	8
PH	41

Table: Number of selection as best model by AIC

Even with consideration of model complexity, PH model is the most chosen method at USHCN Texas data.

On Tail Behavior

- From previous section, hybrid models adopt GPD as their tail distribution to reflect heavy tail behavior of daily precipitation.
- Even though [Koutsoyiannis, 2004] showed heavy tail properties of daily precipitation, some researches like [Booij, 2002, Park et al., 2011] used exponentially decaying distributions.
- Phase-type distribution has exponentail tail mathematically, but it could explain much more extreme values rather than single component exponential tail distributions like exponential, gamma, or Weibull.
- To evalutate model performance on extreme part, mean of square quantile error, denoted by $Q(\eta)$, is used.

$$Q(\eta) = \frac{\sum_{i:q_i \geq \eta}^{n} (q_i - \hat{q}_i)^2}{n}, \quad \eta \in [0,1] \tag{11}$$

n : sample record length, q_i : i^{th} sample quantile, $\hat{q}_i$: i^{th} theoretical quantile

On Tail Behavior

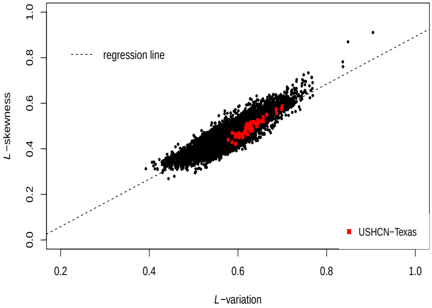
	$\eta = 0.9$			$\eta = 0.95$			$\eta = 0.98$		
ID	GGP	EGP	PH	GGP	EGP	PH	GGP	EGP	PH
ID2	0.78	3.87	1.55	0.70	3.86	1.53	0.68	3.77	1.50
ID5	2.45	4.87	0.22	2.35	4.86	0.22	2.32	4.79	0.18
ID9	1.30	1.22	0.25	1.01	1.13	0.23	0.70	1.10	0.22
ID11	1.83	0.91	0.32	1.66	0.90	0.31	1.13	0.86	0.29
ID15	2.65	0.76	0.65	2.46	0.71	0.65	2.11	0.70	0.63
ID24	1.18	4.21	0.17	1.16	4.20	0.16	0.92	4.19	0.15
ID31	8.86	4.16	2.23	8.55	4.08	2.22	8.18	4.05	2.21
ID44	2.98	13.92	2.68	2.68	13.87	2.67	1.94	13.83	2.65

Table: Mean of square qunatile difference with each value of η equals to 0.9, 0.95, 0.98

Distinct Instance: Bell-shaped Distribution

From [Papalexiou and Koutsoyiannis, 2012], shape of daily precipitation distribution around the world explained with L-moments ratios.

- Low L-variation & low L-skewness indicates bell-shaped distribution
- High L-varitation & high L-skewness indicates J-shaped distribution



- L-moments ratio diagram of 19238 stations in GHCN data set.
- USHCN-Texas samples are placed in relatively J-shape area.

Distinct Instance: Bell-shaped Distribution

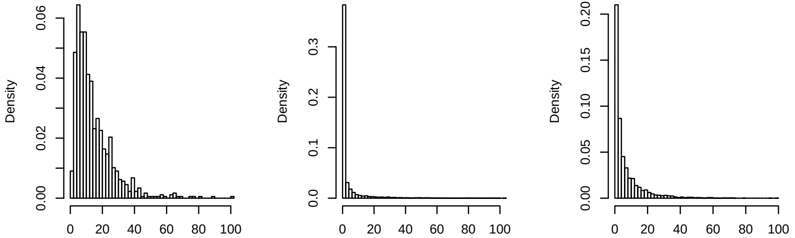


Figure: Histograms of bell-shape(right), J-shape(middle), and one of USHCN Texas(right) data

Distinct Instance: Bell-shaped Distribution

- Phase-type distribution has high flexibility on fitting body part of data
- GGP and EGP model have downside gap on body part
- Even EGP model can't obtain estimates with its original constraint $\xi > 0$

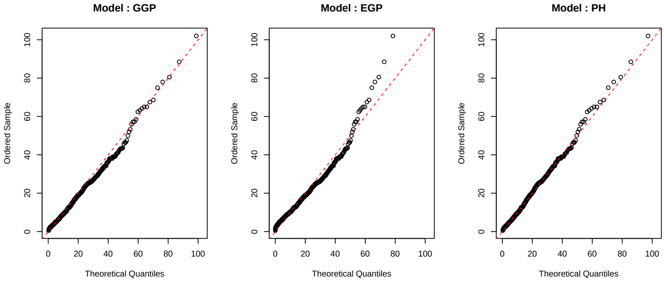


Figure: QQ plots for bell-shaped sample station modelled by GGP, EGP, and PH model

Conclusion

1. Among various approaches to model daily precipitation, PHD shows great performace in general.
2. Also, even though PHD has exponentially decaying tail, PHD appropriately caputures tail behavior of daily precipitation.
3. To be used as broadly accepted precipitation model, PHD surpasses other models with flexibility on fitting body part of distribution.

Asmussen, S., Nerman, O., and Olsson, M. (1996).
Fitting phase-type distributions via the em algorithm.
Scandinavian Journal of Statistics, pages 419–441.

Booij, M. (2002).
Extreme daily precipitation in western europe with climate change at appropriate spatial scales.
International Journal of Climatology, 22(1):69–85.

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

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.

Neuts, M. F. (1974).

Probability distributions of phase type.
Purdue University. Department of Statistics.

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.

Park, J.-S., Kang, H.-S., Lee, Y. S., and Kim, M.-K. (2011).
Changes in the extreme daily rainfall in south korea.
International Journal of Climatology, 31(15):2290–2299.

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).

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

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