Modelling the entire range of daily precipitation using mixture distributions

Hyunwoo Rho

1. Introduction

Measuring and predicting the precipitation are considered to be of great importance in many applications, including agriculture, forest management and hydrology. Common models of the precipitation are stochastic, in that they describe the phenomenon of rainfalls through some static or dynamic statistical models with specific distributions, with a suitable time interval as a base period of modelling. The choice of the time interval can be arbitrary in theory, but taking a daily basis is most natural and popular in the literature; see, for example, Richardson (1981).

Many of previous studies tried to model daily rainfall data through parametric approaches, for example, Ison *et al.* (1971); Mielke Jr and Johnson (1973); Richardson (1981); Stern and Coe (1984). Non-parameteric approaches have also been proposed by, e.g., Sharma and Lall (1999); Harrold *et al.* (2003). Though each approach has pros and cons, parametric models are generally advantageous in that they can model rainfall datasets with only a few parameters and can be easily extrapolated towards extreme area where rainfall observations are relatively rare in frequency. Simple parametric models found in the literature include exponential, gamma and some transformed normal distributions. However, these simple models often fail to capture important characteristics of the rainfall datasets which are typically skewed, over-dispersed, relatively heavy-tailed, and sometimes bell-shaped. As an alternative approach to improve the calibration, mixtures of two parametric distributions have been suggested in the literature.

Being natural extensions of those simple stochastic models, the class of mixture distributions is more flexible with more parameters, and thus can address unique shape characteristics of the rainfall datasets. Successful applications of the mixture distributions towards the precipitation datasets can be abundantly found in the literature. For example, references here.

Related to this, along some studies in hydrology community, the evidence of heavy-tailed phenomenon on the distribution of high precipitation amount has recently gained strong support from researchers; see, e.g., Koutsoyiannis (2004). The underlying idea of the heavy tail

essentially says that the extreme precipitation amounts over some high threshold value tend to have different distributional behaviours and, accordingly, can be modelled using a separate statistical distribution. In particular, the distribution for the tail is modelled by the generalized Pareto distiribution (GPD) justified by the standard result of extreme value theory (EVT). To this extent various methods were investigated and explored in the literature, focusing on the tail behavior of daily precipitation described by some power (or polynomial) survival function; see, for instance, Furrer and Katz (2008); Li *et al.* (2012); Papalexiou and Koutsoyiannis (2012); Papalexiou *et al.* (2013). Most notably, Li *et al.* (2012) presented a hybrid distribution to model the full spectrum of daily precipitation under the EVT framework where the body and tail parts of the dataset are modelled by an exponential and the GPD, respectively. However, there seems no universal agreement on how heavy the precipitation amounts are. For example, Wilson and Toumi (2005) pointed out that extreme daily precipitations of some areas can be adequately modelled by an exponential tails, which is shorter than a power tail. This implicitly suggests that there is no need of separate modelling for the tail as long as the candidate distributions have exponentially decaying tails. more similar references needed. Thus it remains controversial that whether the precipitation extremes follow a power-tail or an exponential-tail, and it is reasonable to say that, depending on the area under consideration, an exponential-type distribution or its mixture may serve as a good candidate to describe the rainfall datasets' body as well as tail.

In this paper, we are interested in modelling the entire range of typical precipitation datasets. For this, we propose the phase-type distribution (PHD) as an alternative mixture distribution to model the precipitation. The PHD is a class of mixture distributions that contains many well-known distributions, such as an exponential and Erlang mixture, as members. Though this distribution class has been used in reliability and insurance contexts, its applications towards the precipitation seems not found in the current literature. For the purpose of the present article, we closely follow the work of Li *et al.* (2012) where Texas rainfall datasets are analyzed using various single and hybrid models, both with and without the EVT framework. In particular, we use the identical datasets and model them with the PHD and compare with other existing parametric models. Later we also use the global precipitation datasets used in, e.g., Papalexiou and Koutsoyiannis (2012); Papalexiou *et al.* (2013) for further analysis as the global datasets exhibit somewhat different characteristics. For these datasets we carry out extensive fitting exercises and standard model validations and show that the PHD is a flexible, competitive and well-rounded alternative stochastic model in describing the whole range of precipitation datasets of different shapes.

The present article is organized as follows. In Section 2. I will revise below later. we compare various existing parametric models in the literature and further introduce a particular type of mixture distribution called phase-type distribution. Through this model we tried to fit the whole range of continuous daily precipitation. Moreover, model evaluation was done with concentration on the tail part as well as the entire data.

2. Data Sets

Let us first describe the datasets to be used throughout the present article. The main data set analysed 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, we selected Texas which has 49 stations (correct?), following Li *et al.* (2012). Further, the same data selection criteria were adopted so that all non-zero precipitation values during 1940 and 2009 are included, without taking care of missings.

The second set of data, which will be analyzed separately later in this article, is a global precipitation dataset from Daily Global Historical Climatology Network (GHCN-DAILY), which has also been widely used in precipitation analyses in the literature. This dataset includes about 100,000 stations and contains USHCN as its subset. We filtered stations according to Papalexiou *et al.* (2013), that is, we only included stations that have: (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 this dataset, see Menne *et al.* (2012). After data cleansing, we obtained the records of 19,328 stations from GHCN, which are almost identical to those used in Papalexiou and Koutsoyiannis (2012) and Papalexiou *et al.* (2013). Although USHCN data is a near subset of GHCN data, we analyze them individually as they have distinct characteristics.

3. Existing models

3.1. Single component models

Fundamental characteristrics of daily precipitation data are described as non-negative, continuous except at the spike on zero, and right-skewed. Assuming we only focus on the continuous part of the data, common candidate statistical models include exponential distribution by, e.g., Todorovic and Woolhiser (1975), gamma distribution by Ison *et al.* (1971); Wilks (1999); Schoof *et al.* (2010), and kappa distribution by Mielke Jr and Johnson (1973), to name a few. Let X denote the non-zero daily precipitation amount, then each model's probabilty density function is given by:

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

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

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

Clearly, these models are ordered in that they become increasingly more flexible with more parameters. Other distributions can also be found, such as, skewed-normal (Wan *et al.*, 2005), truncated power of normal distribution (Bardossy and Plate, 1992), and genralized Pareto distribution (GPD) provided that only the tail of dataset is modelled. The list of possible single component models can surely be much longer, but we restrict ourselves to these three models to avoid crowded model comparison. However, as evidenced in previous studies, it is now much agreed that single component models are generally inadequate to describe the whole range of daily precipitaion datasets. To illustrate this, we present the fitted result of these three models for Texas Station ID 2 (ID 2, in short) data in Figure 2. From the figure, we see that the exponential and gamma models underestimate the tail, and the kappa model overestimates it. With other stations, this tendency is unchanged, and the fits are generally extremely poor.

3.2. Multi component models

Further improvement in fitting can be achieved by inserting additional distribution to single component models. In this regard, mixture type distributions were suggested in the literature, including a mixed exponential distribution (Wilks, 1999) and dynamic mixture of gamma and GPD (Vrac and Naveau, 2007). Alternatively, hybrid models could be considered, os which the density is created by stitching two densities at a particular threshold. A hybrid distribution is also known as a composite distribution in the statistical literature. In recent studies, Furrer and Katz (2008) suggested a hybrid distribution of gamma and generalized Pareto distribution (GGP in short), and Li *et al.* (2012) proposed a hybrid of exponential and generalized Pareto distribution (EGP). Both of these models take generalized Pareto distribution (GPD) as the upper tail distribution to reflect heavy tail behaviuor of precipitation extremes (Koutsoyiannis, 2004). Just like single component models, one can create infinitely many different mixture or hybrid models, and it is not feasible to consider all possible candidates. An extensive comparative study of Li *et al.* (2012) shows that the performance of hybrid models defeat the one of mixture models in general Please revise this sentence. it is too vague to me. make it more detailed.. Hence, we mainly cover the hybrid models to be compared rather than other mixture distributions.

$$f(x;\alpha,\beta,\xi,\sigma,\theta) = f_{gam}(x;\alpha,\beta)I(x \le \theta) + [1 - F_{gam}(\theta;\alpha,\beta)]f_{GP}(x;\xi,\sigma,\theta)I(x > \theta)$$
$$x \ge 0, \quad \alpha,\beta,\xi,\theta > 0 \tag{4}$$

Where f_{gam} and F_{gam} are probabilty density function and cumulative distribution fucntion of gamma respectively. To make it continuous at the threshold θ, , 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,

$$\sigma = \frac{1 - F_{gam}(\theta;\alpha,\beta)}{f_{gam}(\theta;\alpha,\beta)} \tag{5}$$

As a result, its parameter set reduces to $\{\alpha,\beta,\xi,\theta\}$. Even though the model complexity have been descreased, still the threshold selection problem remains, which doesn't have any

guaranteed method and entails cumbersome trial-and-error task. Hybrid of exponenetial and generalized Pareto distribution Li *et al.* (2012) has an advantage on such huddle by passing away it via analytical derivation of threshold.

$$f(x;\lambda,\xi,\sigma,\theta) = \frac{1}{1+F_{exp}(\theta;\lambda)}[f_{exp}(x;\lambda)I(x\le\theta) + f_{GP}(x;\xi,\sigma,\theta)I(x>\theta)]$$
$$x \ge 0, \quad \lambda,\xi,\sigma,\theta > 0 \tag{6}$$

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

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

Then the parameter set of EGP model becomes $\{\lambda,\xi,\sigma\}$. By sacrificing model flexibity with taking exponential rather than gamma on its body part, it obtains a primary advantage circumventing the threshold selection problem.

4. Phase-type distribution class

5. Analysis

Model parameters are estimated via maximum likelihood method. Especially for fitting Phase-type distribution, a kind of EM algorithm proposed by Asmussen *et al.* (1996).

For GGP model, estimation procedure of by Furrer and Katz (2008) was applied. Also 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. In Li *et al.* (2012), this model indicated its best performace with threshold $\theta = 3.99$, that is around .Thus we set threshold value used in GGP model moderately as 55% quantile of each sample and avoid cumbersome optimal threshold selection procedure.

5.1. Model comparison

As shown in Figure 2, single component models are not enough to sufficiently depict the entire range of daily precipitation amount. Hence we exclude those models at in depth comparison analysis among models and mainly cover the multi component models in this section; GGP, EGP, and PH model.

Figure 6 presents fitting result of three models on three different stations of USHCN Texas data set. All the models show good performances in general especially in the body part of the data. Differences stand out at the tail part. Throughout three of stations, EGP model estimates its tail thickness heavier than sample data points, because points at tail part of QQ plot are

marked on the lower side of 45 °line. For ID44, EGP gets poor result by estimating its tail thickness too large. On the other hand, GGP and PH model results exhibit similar pattern, even performance of GGP highly depends on choosing the threshold value.

Visual model evaluation methods like QQ plot have advantage on concise and dramatical comparison bewteen distinct models. With Figure 6, however, accurate comparison is problematic. Therefore numerical measurement is needed for precise research. Akaike information criteria(AIC) Akaike (1974) is well known and most widely used model selection method. It has an advantage of reflecting model complexity by giving penalty on the number of parameters. AIC surpassed visual comparison methods especially in this case, since three models above have different number of parameters, which are 4 of GGP, 3 of EGP, 5 of PH distribution.

AIC values for each stations and each models are presented in Table 2. With AIC values, the smallest is the best. Especially for ID2, ID11 and ID44 from Figure 6, PH distribution is chosen as the best model according to AIC values. Even with all 49 stations of USHCN Texas, PH distribution is selected as the best model of 41 stations among them, but GGP distribution couldn't at least once. Despite we can enhance the fitting performance of GGP distribution by finding the optimal threshold value, still such cumbersome procedure is not desirable.

Even we can try to compare more limited version of PH distribution and EGP distribution. As we restrict the number of phases to two, also with Coxian assumption, the number of parameters of PH model becomes three, same with EGP model.

5.2. Observation on extreme values

As mentioned earlier, there is a controversy on the tail behavior of daily precipitation data. There are two well known approaches and corresponding statistical methods to investigate the extreme values; block maxima(BM) with generalized extreme value distribution(GEV) and peaks over threshold(POT) with generalized Pareto distribution(GPD). Both of distributions have shape parameter ξ, which determines tail behavior of data. GHCN data set, which contains about 20000 stations around the world, is used to figure out thet general tail characteristics of daily precipitation data. Also, to make consistency with foregoing models; GGP and EGP, we take POT-GPD approach on examining tail data.

Then it becomes contentious issue that how to define the extreme values. Some commonly used criteria are above 95% or 98%. But we adopt a conventional way in hydrology Cunnane (1973), that is to make the number of values above the threshold equal to the number of years of observation record.

In the extreme value theory(EVT), the tail thickness of data is determined by the shape parameter ξ. If it has negative ξ, it has upper bound, or if ξ is zero, it has a exponential tail, or if it has positive ξ, it has a polynomial tail which is much heavier than exponential. Figure 6 is the empirical distribution of $\hat{\xi}$ and it seems symmetric bell-shape distribution like normal distribution. Fitted line is a normal density curve having sample mean and sample standard deviation as its parameter, which are 0.0896 and 0.1814 respectively. About 29% of estimates are obtained in negative value. One can say that this estimating results support that the shape parameter ξ has postive value in general. However, in the point of view of hypothesis testing, it is hard to say that ξ is not zero.

Phase-type distribution is a mixture of several gamma distributions and one exponential distribution. Thus its tail behavior follows exponential tail theoretically. The fundamental characteristic of extreme value of daily precipitation data is still arguable, like comparable fitting results of Figure 6. More precise research on extreme values is needed to figure out that which kind of tail assumption is more appropriate with precipitaion data.

Some of methods for testing goodess of fit like Kolmogorov-Smirnov(KS), Cramer von Mises(CvM), and Anderson-Darling(AD) test include a difference between empirical distribution function F_n and model distribution fuction $\hat{F}$. Similarly, we take quantile function based approach that calculates difference between sample quantile and theoretical quantile. We adopt simple statistic $Q(\eta)$ as a quantile based model evaluation method.

$$Q(\eta) = \sum_{i:x_i \geq \eta}^{n} (q_i - \hat{q}_i)^2, \quad \eta \in [min(x_i), max(x_i)] \tag{8}$$

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

In the case of setting $\eta = min(x_i)$, $Q(\eta)$ represents model performance for the entire range of data. By taking η as 90%, 95%, or 98% sample quantile, we can concentrate on modelling extreme values.

5.3. Distinct instance

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. As a result two distinct shapes were observed called J-shaped and bell-shape distriution. A typical type of parametric model having J-shape distribution is exponential model. The level of J-shaped feature could be judged visually by high peak and steepness around zero; a steeper, a more J-shaped. This kind of empirical distribution is usual case of precipitation data. But bell-shape distributions is well known with noraml distribution. Though, in this case we have non-negative and right-skewed random variable, thus we say it has bell-shaped distribution if it has such shape on the body part of the distribution, like gamma distribtion with its shape parameter greater than one.

L-moments ratio diagram made with GHCN data set, Figure 6, shows rough positive linear relationship between L-variation and L-skewness. The larger value of L-variation and L-skewness, the more J-shaped. Square red dots plotted on Figure 6 represent 49 stations of USCHN-Texas data set. Those sample stations are more likely to be J-shaped rather than bell-shaped according to their position on L-moments ratio diagram. Both GGP and EGP model are only simulated only with such J-shaped sample data, and did not consider other extreme cases of bell-shaped data.

Especially for EGP model, it has fundamental defect on modelling bell-shaped data since it takes exponential distribution as its body part model. Accordingly, estimates couldn't be obtained with original parameter space in 6, but estimation succeeded with negative value of

ξ. Negative value of shape parameter is still embarrassing result since it infers bounded tail as reported by extreme value theory.

In Figure 7, unlike USHCN Texas samples, body parts of GGP and EGP QQ plot are under the 45 °line. PH model, however, still performs well with bell-shaped sample on both body and tail part. Such strength comes from flexible mixing procedure of PH distribution. This result means PH model can be the generally accepted model for daily precipitation data including both J-shaped and bell-shaped distribution.

6. Conclusion

References

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

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

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

Cunnane, C. (1973). A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. *Journal of hydrology*, **18**(3), 257–271.

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).

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).

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.

Vrac, M. and Naveau, P. (2007). Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water resources research*, **43**(7).

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.

Wilson, P. and Toumi, R. (2005). A fundamental probability distribution for heavy rainfall. *Geophysical Research Letters*, **32**(14).

List of Figures

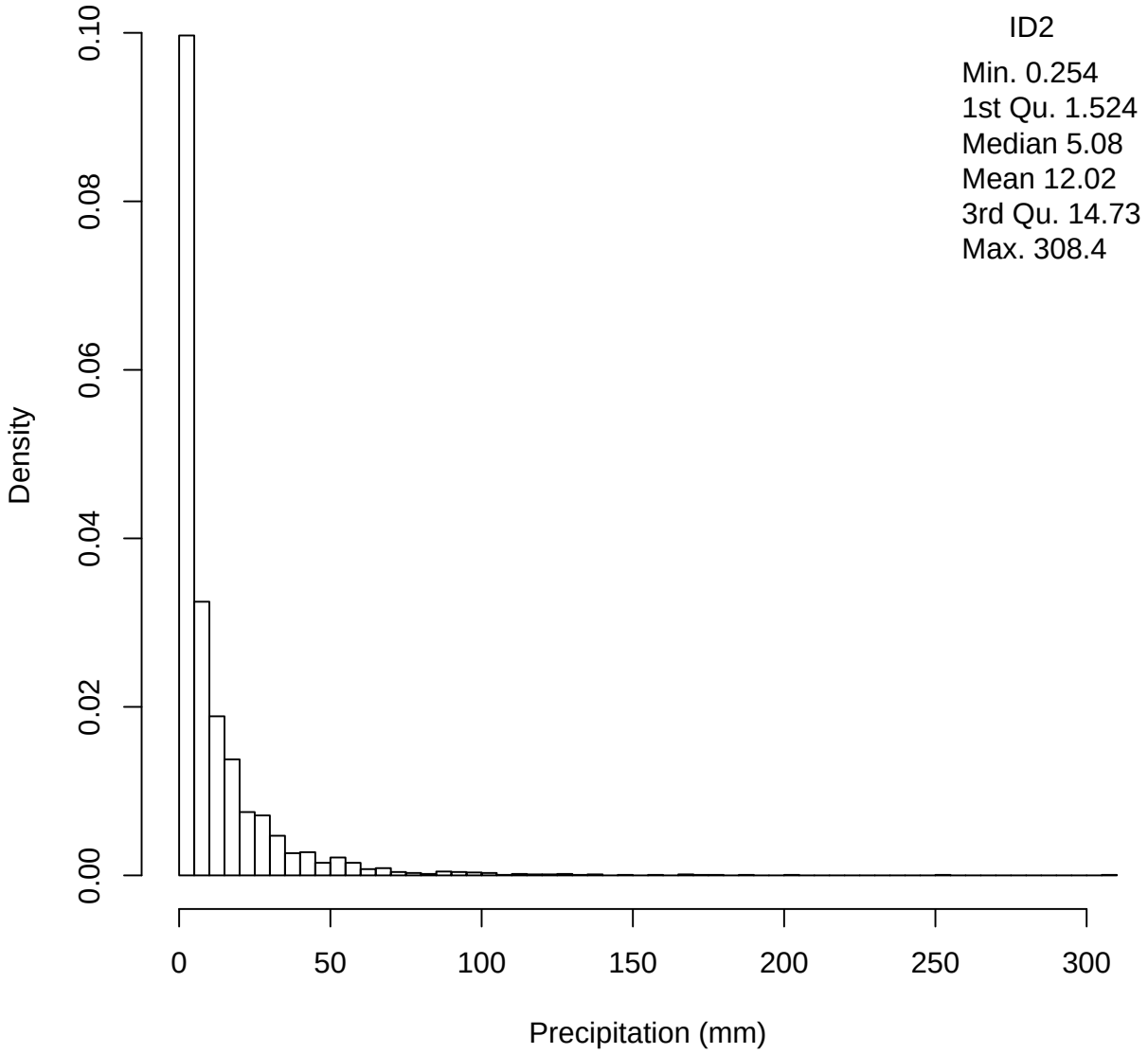

Figure 1: Sample histrogram : ID2

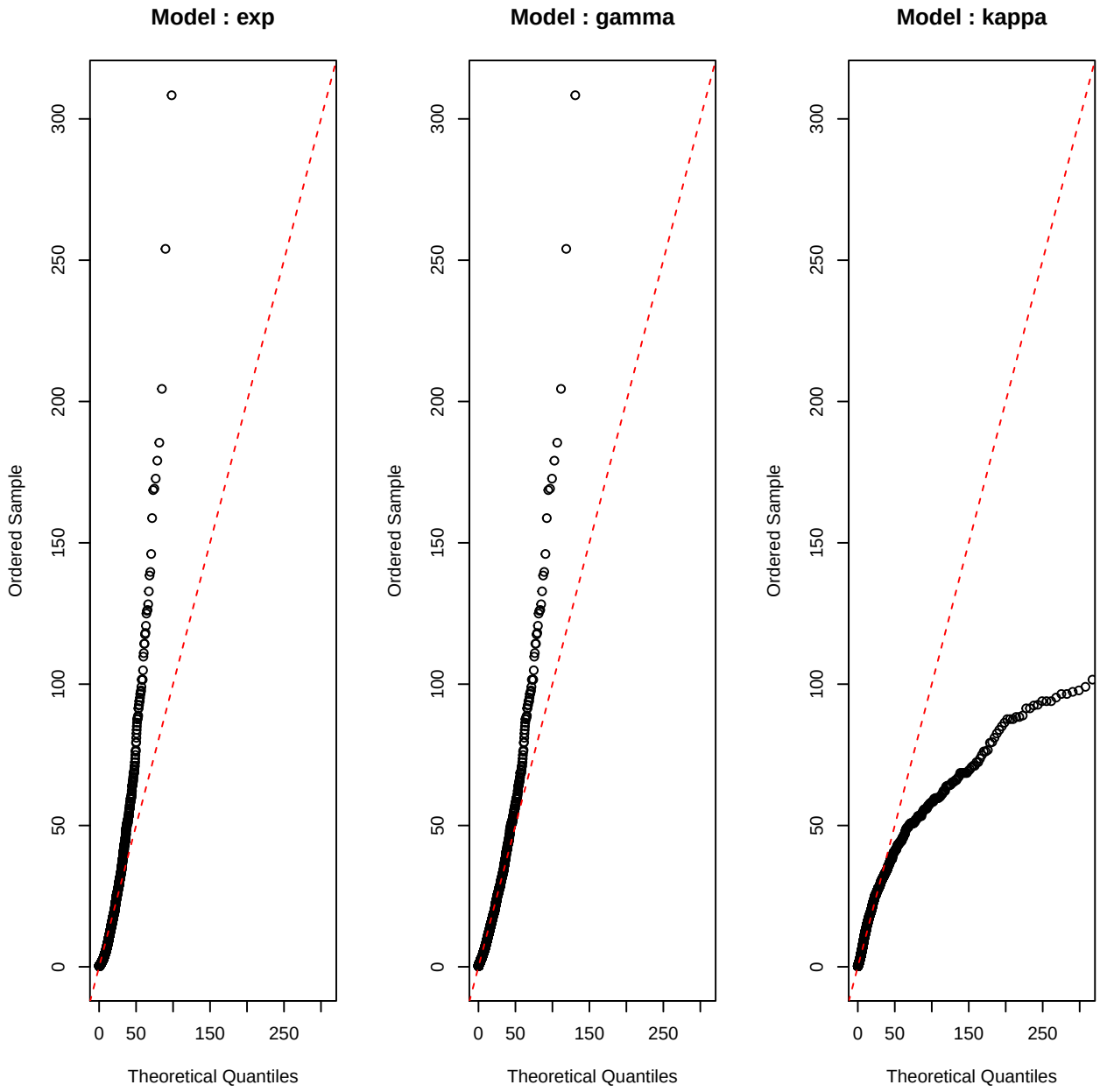

Figure 2: QQ plots of ID2 modelled by single component models; Exponential, gamma, and kappa

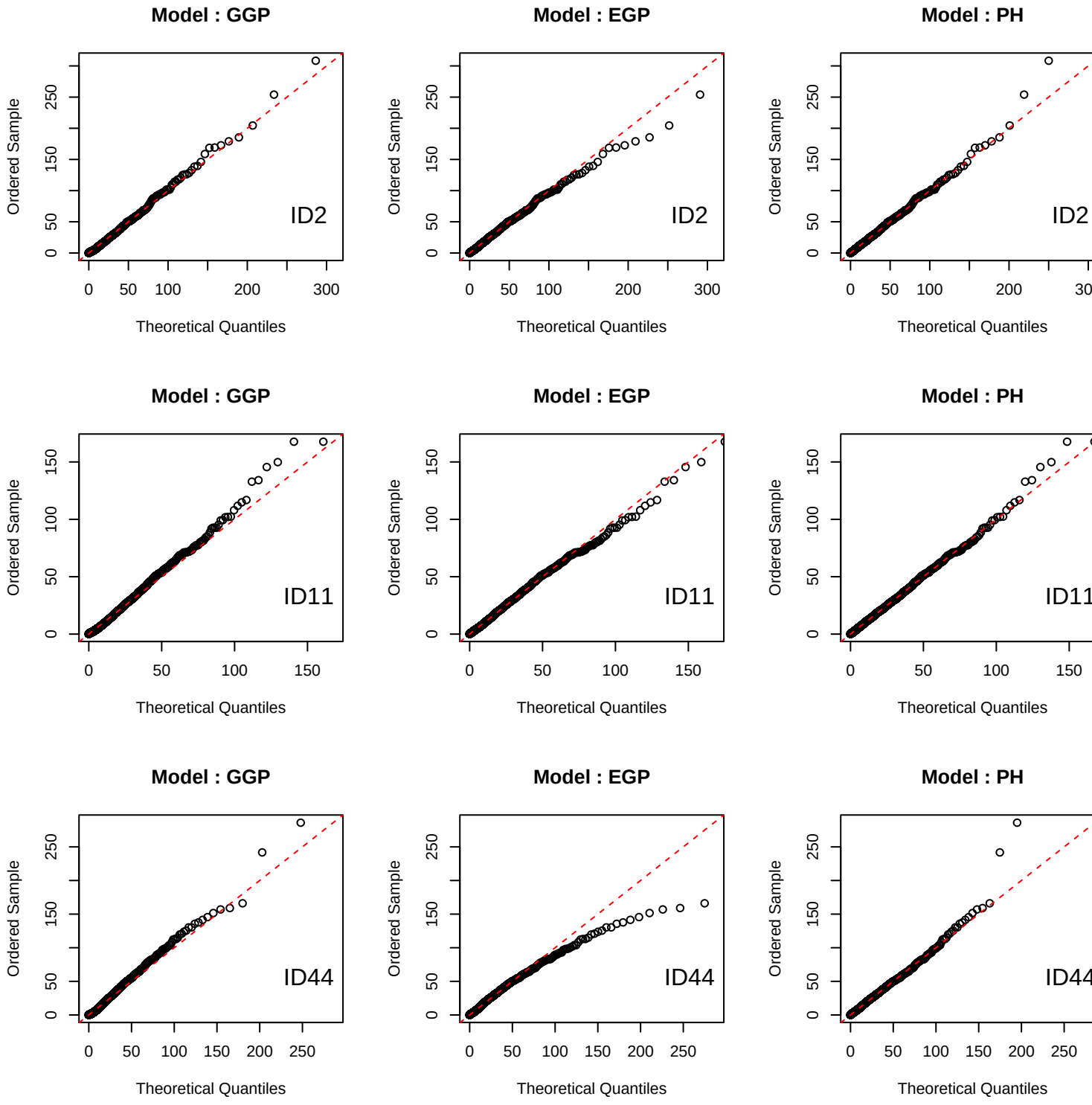

Figure 3: QQ plots of 3 different stations modelled by GGP distribution(left), EGP distribution(middle), and the PH distribution(right)

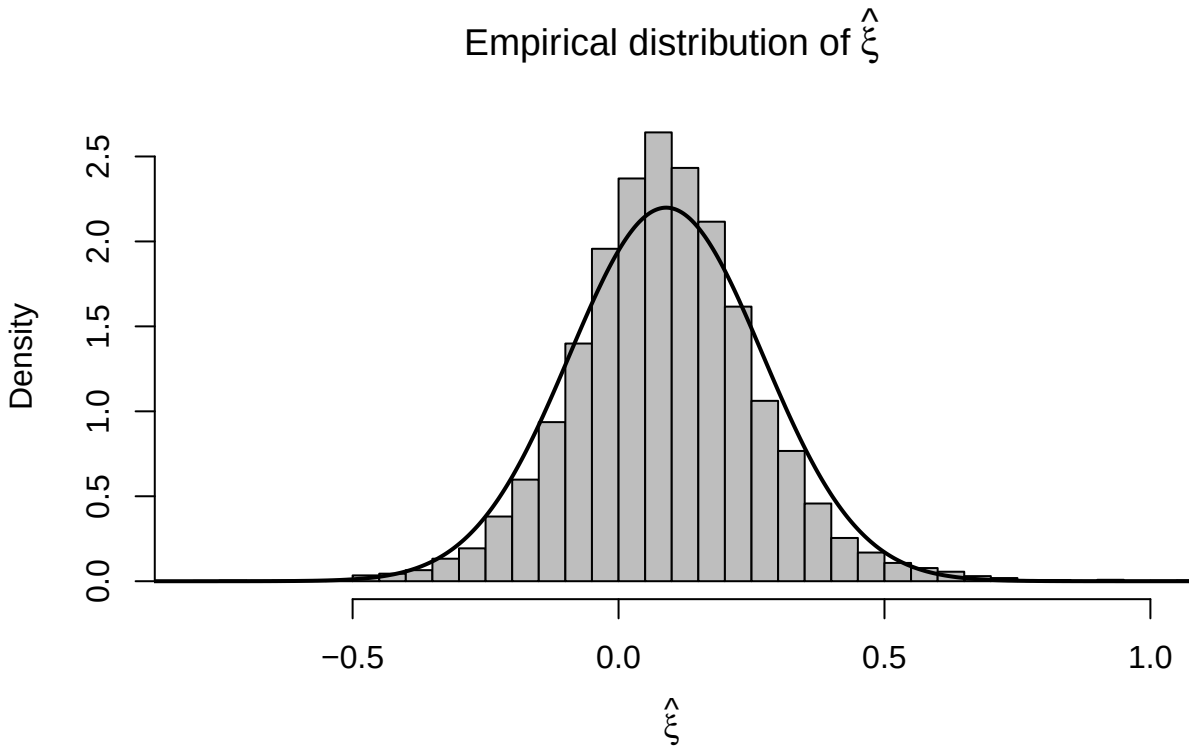

Figure 4: Empirical distribution of $\hat{\xi}$ and fitted normal curve. $\hat{\xi}$ is estimated from each of 19328 stations of GHCN data set.

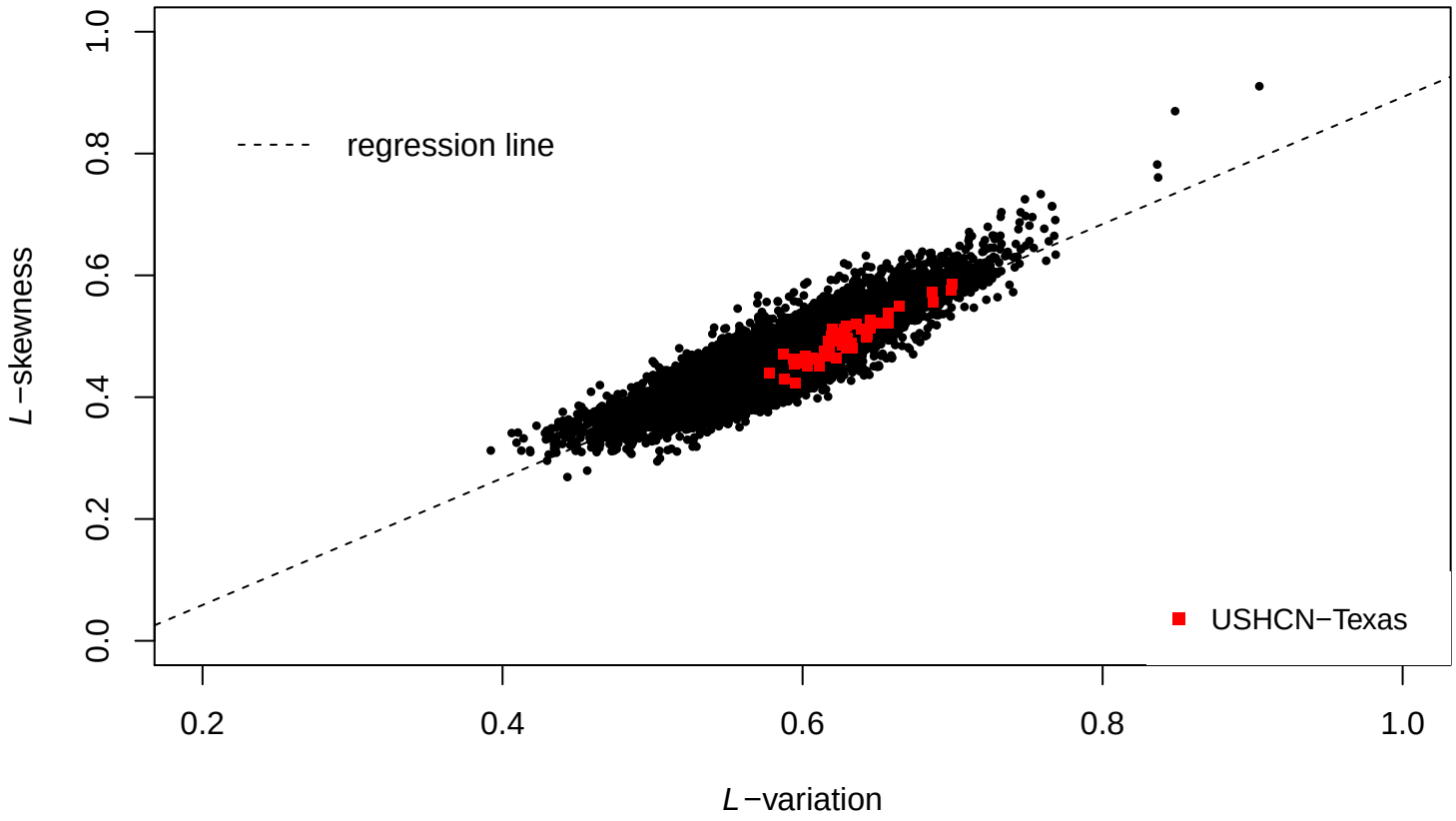

Figure 5: L-variation vs. L-skewness plot

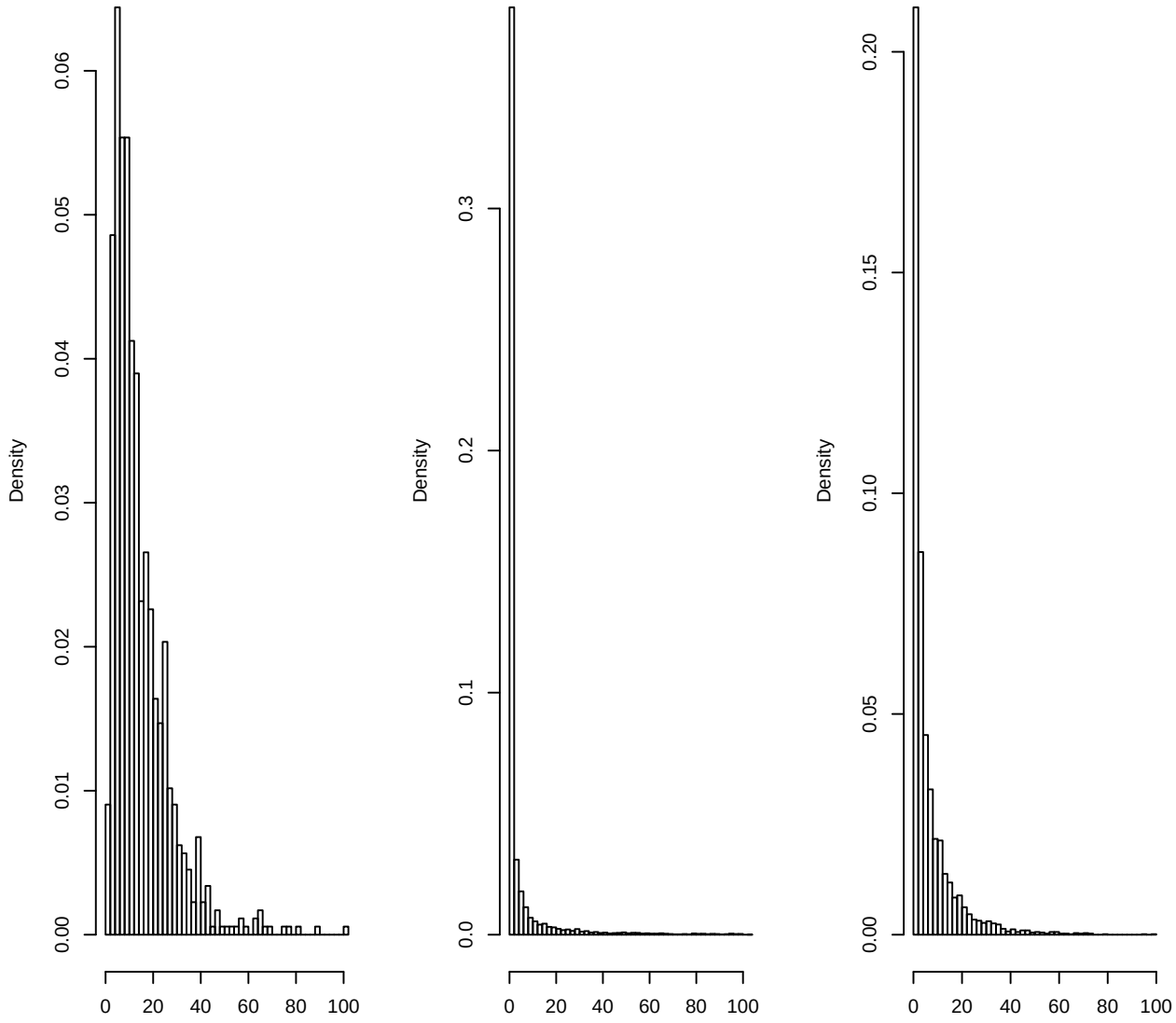

Figure 6: Histograms of bell-shape(right), J-shape(middle), and one of USHCN Texas(right) data

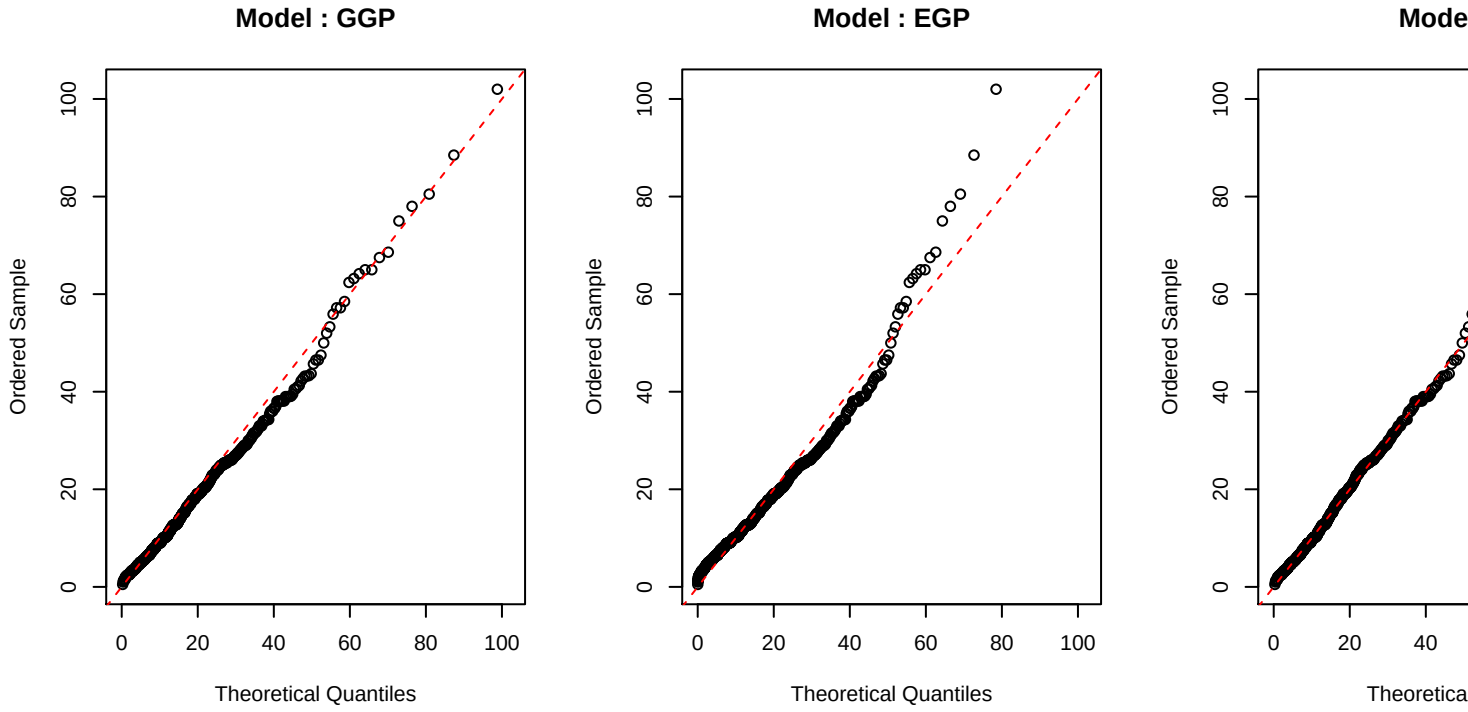

Figure 7: QQ plots for bell-shaped sample station modelled by GGP, EGP, and PH model

List of Tables

	ID	Label	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	410120	ID1	0.25	1.52	5.33	11.27	14.22	737.90
2	410144	ID2	0.25	1.52	5.08	12.02	14.73	308.40
3	410174	ID3	0.25	0.76	2.79	6.54	8.13	99.57
4	410493	ID4	0.25	1.52	4.83	10.40	13.46	179.10
5	410498	ID5	0.25	1.52	3.81	7.54	9.65	104.90
6	410639	ID6	0.25	1.27	4.06	10.74	12.70	269.50
7	410832	ID7	0.25	1.52	4.83	11.39	14.48	443.70
8	410902	ID8	0.25	1.27	4.06	10.89	12.95	226.80
9	411000	ID9	0.25	1.52	4.57	9.27	12.19	114.30
10	411048	ID10	0.25	1.52	4.57	11.66	14.73	263.70
11	411138	ID11	0.25	1.78	5.84	11.87	15.49	167.60
12	411528	ID12	0.25	2.29	6.10	13.43	16.51	254.00
13	411772	ID13	0.25	2.29	7.62	14.13	19.05	189.20
14	412015	ID14	0.25	0.76	3.05	10.17	11.43	250.40
15	412019	ID15	0.25	1.78	5.84	12.50	16.51	253.00
16	412121	ID16	0.25	1.27	3.81	8.77	10.92	146.00
17	412266	ID17	0.25	1.27	4.57	11.93	14.73	329.20
18	412598	ID18	0.25	1.52	5.08	11.33	14.22	221.20
19	412679	ID19	0.25	1.02	2.79	9.29	10.41	217.90
20	412797	ID20	0.25	0.76	2.03	4.53	5.40	72.14
21	412906	ID21	0.25	2.29	6.35	13.18	17.02	198.10
22	413063	ID22	0.25	1.27	3.81	10.09	11.43	254.00
23	413183	ID23	0.25	1.78	5.59	12.58	15.75	360.90
24	413280	ID24	0.25	1.02	3.05	7.14	8.89	132.60
25	413420	ID25	0.25	2.54	7.62	14.78	18.80	202.90
26	413734	ID26	0.25	1.78	6.35	13.09	17.78	189.20
27	413873	ID27	0.25	1.27	5.08	11.99	14.99	171.40
28	413992	ID28	0.25	1.27	4.32	9.56	11.68	363.00
29	415018	ID29	0.25	1.27	4.32	10.14	13.21	176.50
30	415196	ID30	0.25	2.54	7.11	15.04	19.05	469.90
31	415272	ID31	0.25	1.52	5.08	10.76	13.97	318.30
32	415429	ID32	0.25	1.27	4.19	10.87	13.21	267.50
33	415618	ID33	0.25	2.03	7.11	13.74	19.30	217.90
34	415707	ID34	0.25	1.27	3.56	8.13	9.91	231.90
35	415869	ID35	0.25	2.03	6.35	13.02	17.78	219.20
36	415875	ID36	0.25	1.52	4.83	9.34	12.19	141.70
37	416135	ID37	0.25	1.27	3.81	7.98	10.41	133.40
38	416276	ID38	0.25	1.27	4.83	11.47	14.73	466.10
39	416794	ID39	0.25	1.52	5.84	12.87	17.27	191.50
40	416892	ID40	0.25	1.02	3.05	6.84	8.13	111.30
41	417079	ID41	0.25	1.02	3.56	7.99	10.41	177.80
42	417336	ID42	0.25	1.27	4.57	10.09	12.95	204.00
43	417622	ID43	0.25	0.76	2.79	8.86	9.40	317.80
44	417945	ID44	0.25	0.76	3.05	9.45	10.92	286.00
45	418201	ID45	0.25	1.02	3.30	7.86	9.40	137.20
46	418433	ID46	0.25	1.52	5.08	10.43	13.97	148.80
47	418692	ID47	0.25	1.52	5.08	8.89	11.43	142.20
48	418910	ID48	0.25	1.27	5.08	11.67	15.24	244.30
49	419532	ID49	0.25	1.52	5.59	11.64	15.24	179.10

Table 1: Summary of USHCN Texas data set

	ID	recordLen	loglik.GGP	loglik.EGP	loglik.PH	AIC.GGP	AIC.EGP	AIC.PH	BIC.GGP	BIC.EGP	BIC.PH
1	ID1	4237	-14240.48	-14093.96	-14098.65	28488.96	28193.92	28207.29	28514.37	28212.97	28239.05
2	ID2	3485	-11852.44	-11693.06	-11681.29	23712.88	23392.12	23372.58	23737.51	23410.59	23403.36
3	ID3	4101	-11527.45	-11312.79	-11289.46	23062.90	22631.59	22588.91	23088.18	22650.54	22620.51
4	ID4	3819	-12575.33	-12440.55	-12433.67	25158.66	24887.10	24877.34	25183.66	24905.84	24908.58
5	ID5	2911	-8719.78	-8663.52	-8652.97	17447.56	17333.05	17315.94	17471.47	17350.98	17345.82
6	ID6	5118	-16762.39	-16482.29	-16458.08	33532.78	32970.58	32926.16	33558.94	32990.20	32958.86
7	ID7	5350	-18014.93	-17783.97	-17769.09	36037.86	35573.94	35548.18	36064.20	35593.70	35581.10
8	ID8	5615	-18477.57	-18158.59	-18120.59	36963.13	36323.18	36251.18	36989.67	36343.08	36284.35
9	ID9	1609	-5127.98	-5083.75	-5078.93	10263.96	10173.50	10167.85	10285.49	10189.65	10194.77
10	ID10	6457	-21797.86	-21431.11	-21402.14	43603.71	42868.23	42814.28	43630.80	42888.55	42848.14
11	ID11	4128	-14187.94	-14076.47	-14069.19	28383.87	28158.93	28148.38	28409.18	28177.91	28180.00
12	ID12	1516	-5368.38	-5320.20	-5318.47	10744.76	10646.41	10646.93	10766.05	10662.38	10673.55
13	ID13	5687	-20578.84	-20443.18	-20444.62	41165.68	40892.36	40899.23	41192.26	40912.30	40932.46
14	ID14	4698	-14704.92	-14304.84	-14287.76	29417.83	28615.69	28585.52	29443.65	28635.05	28617.79
15	ID15	5452	-18945.66	-18781.16	-18774.64	37899.31	37568.31	37559.28	37925.73	37588.12	37592.30
16	ID16	4401	-13711.49	-13535.91	-13516.58	27430.98	27077.82	27043.17	27456.54	27096.99	27075.12
17	ID17	6385	-21508.43	-21150.88	-21138.61	43024.86	42307.75	42287.21	43051.91	42328.04	42321.02
18	ID18	5149	-17317.21	-17097.58	-17092.30	34642.42	34201.16	34194.61	34668.60	34220.80	34227.34
19	ID19	3820	-11789.11	-11492.91	-11445.08	23586.22	22991.82	22900.16	23611.21	23010.56	22931.40
20	ID20	3052	-7489.22	-7352.53	-7338.42	14986.45	14711.06	14686.84	15010.54	14729.13	14716.96
21	ID21	2659	-9424.88	-9348.04	-9343.21	18857.76	18702.08	18696.42	18881.30	18719.74	18725.85
22	ID22	4174	-13373.11	-13101.81	-13092.87	26754.21	26209.62	26195.73	26779.56	26228.63	26227.42
23	ID23	5257	-18257.46	-18072.96	-18057.80	36522.92	36151.93	36125.59	36549.19	36171.63	36158.43
24	ID24	3021	-8792.73	-8666.98	-8651.50	17593.46	17339.97	17313.01	17617.52	17358.01	17343.08
25	ID25	1617	-5918.47	-5881.12	-5881.52	11844.94	11768.24	11773.05	11866.49	11784.40	11799.99
26	ID26	5538	-19481.08	-19293.92	-19286.09	38970.16	38593.83	38582.18	38996.64	38613.69	38615.28
27	ID27	5686	-19270.52	-19027.84	-19015.53	38549.05	38061.67	38041.06	38575.63	38081.61	38074.28
28	ID28	4562	-14510.96	-14305.84	-14307.34	29029.91	28617.68	28624.68	29055.61	28636.95	28656.80
29	ID29	5288	-17138.91	-16861.62	-16858.83	34285.82	33729.25	33727.67	34312.11	33748.97	33760.53
30	ID30	6429	-23512.71	-23291.95	-23294.56	47033.41	46589.90	46599.12	47060.49	46610.20	46632.96
31	ID31	4437	-14800.11	-14653.84	-14644.83	29608.23	29313.68	29299.67	29633.82	29332.87	29331.66
32	ID32	5712	-18786.99	-18470.13	-18452.05	37581.99	36946.26	36914.10	37608.59	36966.21	36947.36
33	ID33	6212	-22240.93	-22042.81	-22038.06	44489.86	44091.62	44086.12	44516.80	44111.82	44119.80
34	ID34	2935	-8910.75	-8797.32	-8776.03	17829.50	17600.64	17562.06	17853.44	17618.60	17591.98
35	ID35	4809	-16970.62	-16802.69	-16796.97	33949.24	33611.38	33603.94	33975.15	33630.82	33636.33
36	ID36	4037	-12916.61	-12810.85	-12806.33	25841.22	25627.69	25622.65	25866.44	25646.60	25654.17
37	ID37	3821	-11656.94	-11561.02	-11538.83	23321.87	23128.04	23087.66	23346.86	23146.78	23118.91
38	ID38	5062	-16955.29	-16686.02	-16685.86	33918.57	33378.04	33381.72	33944.69	33397.63	33414.37
39	ID39	5920	-20629.51	-20392.59	-20389.01	41267.02	40791.18	40788.01	41293.76	40811.24	40821.44
40	ID40	2334	-6706.41	-6621.67	-6610.74	13420.82	13249.35	13231.48	13443.84	13266.61	13260.26
41	ID41	4361	-13155.60	-12959.68	-12957.53	26319.21	25925.37	25925.07	26344.73	25944.51	25956.97
42	ID42	4230	-13682.85	-13504.08	-13499.74	27373.71	27014.17	27009.48	27399.11	27033.22	27041.23
43	ID43	4018	-12092.14	-11737.96	-11707.82	24192.27	23481.91	23425.64	24217.47	23500.81	23457.13
44	ID44	5182	-15975.17	-15558.27	-15541.86	31958.34	31122.53	31093.71	31984.55	31142.19	31126.48
45	ID45	3754	-11254.68	-11082.67	-11055.69	22517.36	22171.33	22121.38	22542.28	22190.03	22152.53
46	ID46	3429	-11317.88	-11219.34	-11214.89	22643.75	22444.69	22439.79	22668.31	22463.11	22470.49
47	ID47	3382	-10706.16	-10651.14	-10645.93	21420.32	21308.29	21301.85	21444.82	21326.67	21332.48
48	ID48	4673	-15817.87	-15618.86	-15613.37	31643.74	31243.72	31236.74	31669.54	31263.07	31268.98
49	ID49	5119	-17387.23	-17196.03	-17196.62	34782.47	34398.06	34403.24	34808.63	34417.68	34435.94

Table 2: Infomation Table

EGP	8
PH	41

Table 3: Number of selection as best model by AIC

EGP	8
PH	41

Table 4: Number of selection as best model by AIC