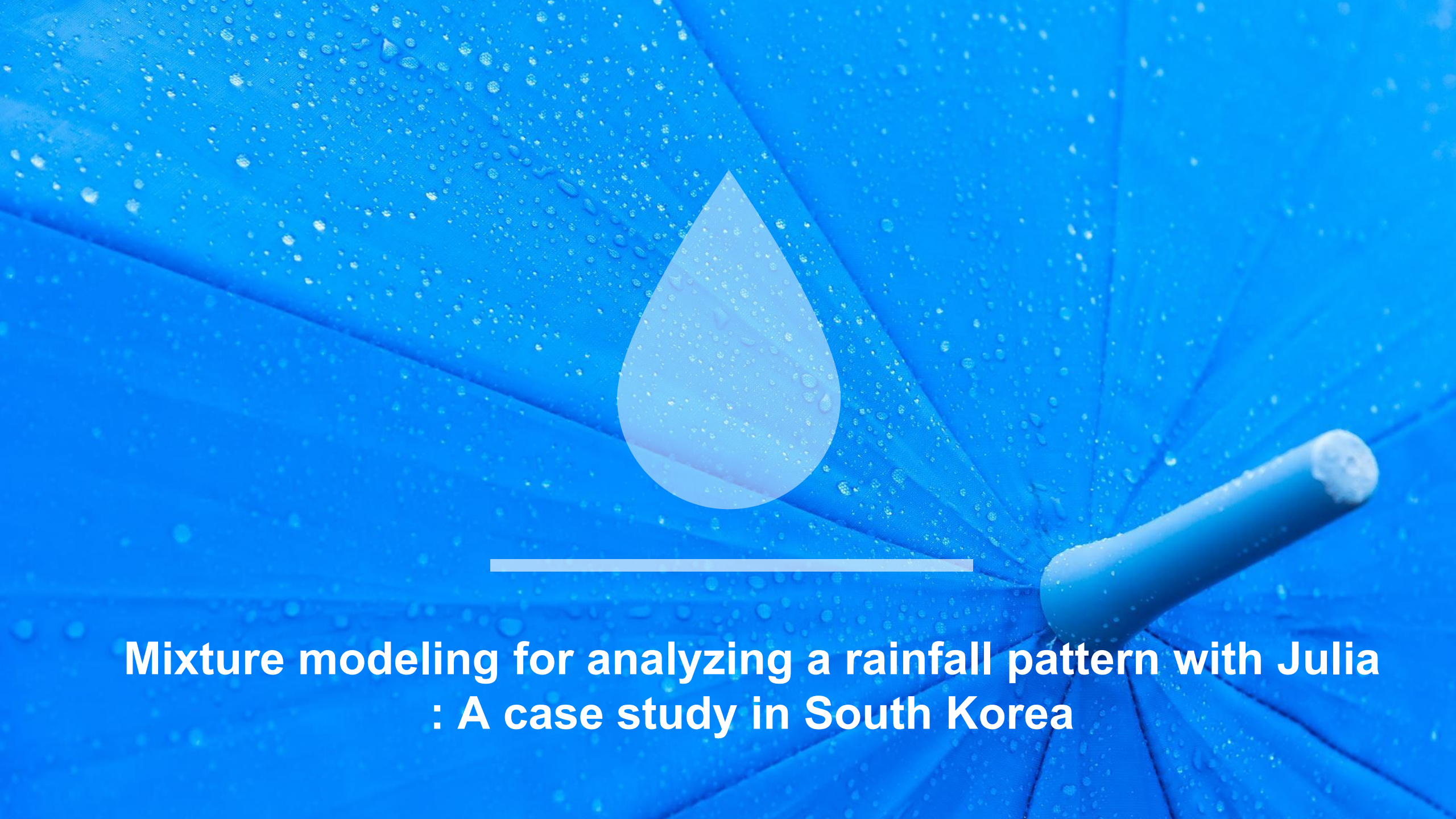


Mixture modeling for analyzing a rainfall pattern with Julia : A case study in South Korea

Contents

Contents

01. Introduction – Mixture model

A mixture model provides a principled approach to modeling such as complex data that might be multimodal – containing multiple regions with high probability mass.

Generally, the probability density function of finite mixture model is defined as

$$f_X(x) = \sum_{i=1}^{K} \pi_i f_i(x|Z_i)$$

K : number of components

f_i : the component probability density functions of the mixture

$Z_i \in \{1, \ldots, K\}$: component variables

$P(Z_i = k) = \pi_i$: component proportions

01. Introduction – Mixture model

The EM algorithm turns out to be a general way of maximizing the likelihood when some variables are unobserved, and hence useful for other things besides mixture models.

The log-likelihood function of mixture distribution:

$$l(\theta) = \sum_{j=1}^{N} \log \sum_{i=1}^{K} \pi_i f(x_j|\theta_i)$$

θ_i : parameter vector of i-th component density function

01. Introduction – Mixture model

An expectation-maximization(EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables.

1. Start with guesses about the mixture components $\theta_1, \theta_2, \ldots, \theta_K$ and the mixing weights $\pi_1, \ldots, \pi_K$
2. Until noting changes very much:
   a. Using the current parameter guesses, calculate the weights w_{ij} (E-step)
   b. Using the current weights, maximize the weighted likelihood to get new parameter estimates (M-step)
3. Return the final parameter estimates (including mixing proportions) and cluster probabilities.

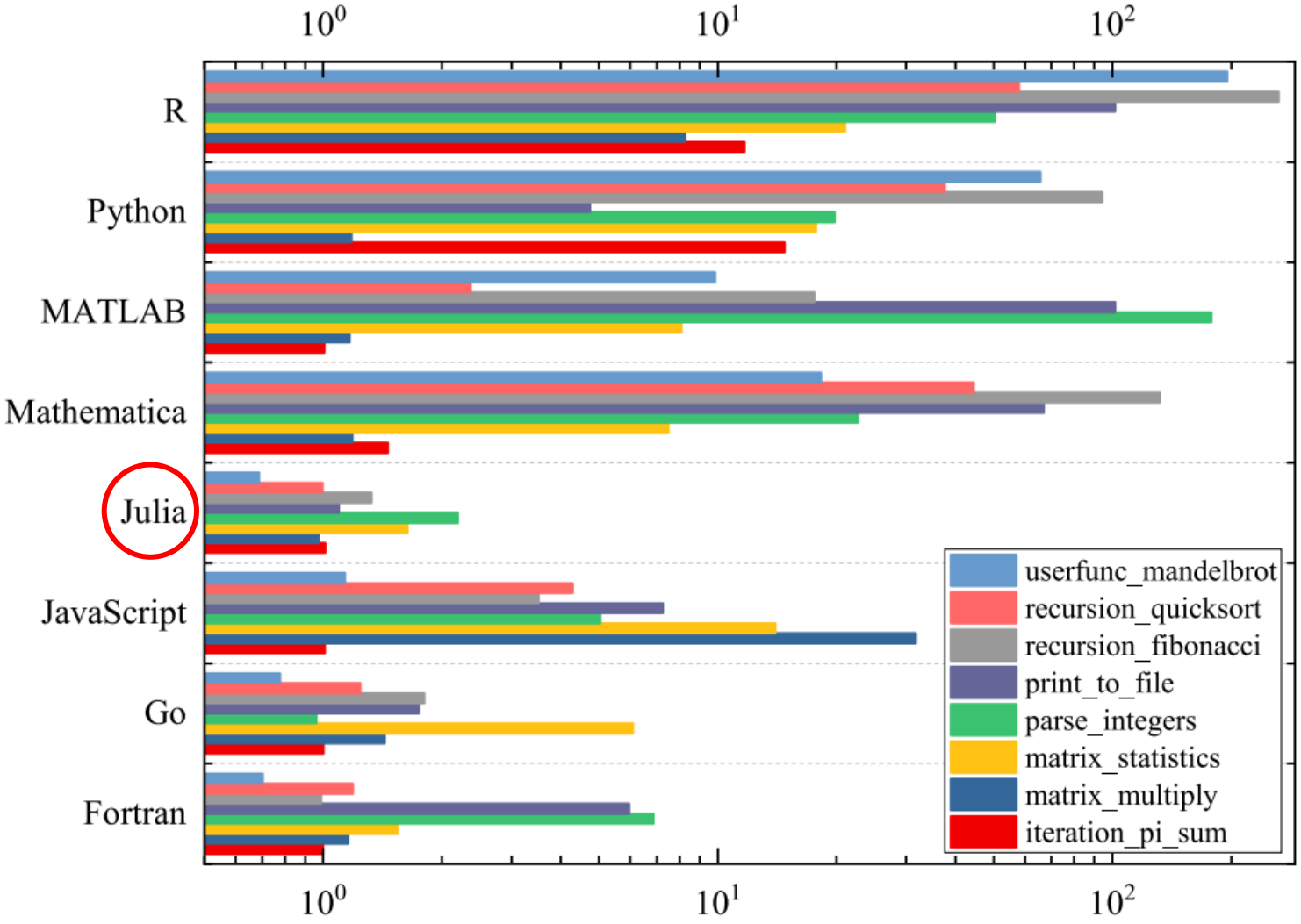
01. Introduction – Julia

Julia is a modern, expressive, and high-performance programming language for scientific computing and data processing. Its development started in 2009.

Julia's grammar is as readable as that of MATLAB or Python, and it can approach the C/C++ language in performance by compiling in real time. In addition, Julia is a free, open-source language that runs on all popular OS.

Julia successfully combines the high performance of a static programming language with the flexibility of a dynamic programming language.

01. Introduction – Julia



01. Introduction – Julia

Julia can be directly accessed from external libraries written in C or Fortran.
And also through the “PyCall” library, it is possible to access Python code or library and even share data between Python and Julia.

example)

```
py"""
import numpy as np

def sinpi(x):
    return np.sin(np.pi * x)

"""
py"sinpi"(1)
```

Julia is primarily aimed at people using scientific computing languages such as MATLAB, R.
Julia’s syntax for mathematical operations looks like mathematical formulas written outside the computing environment.

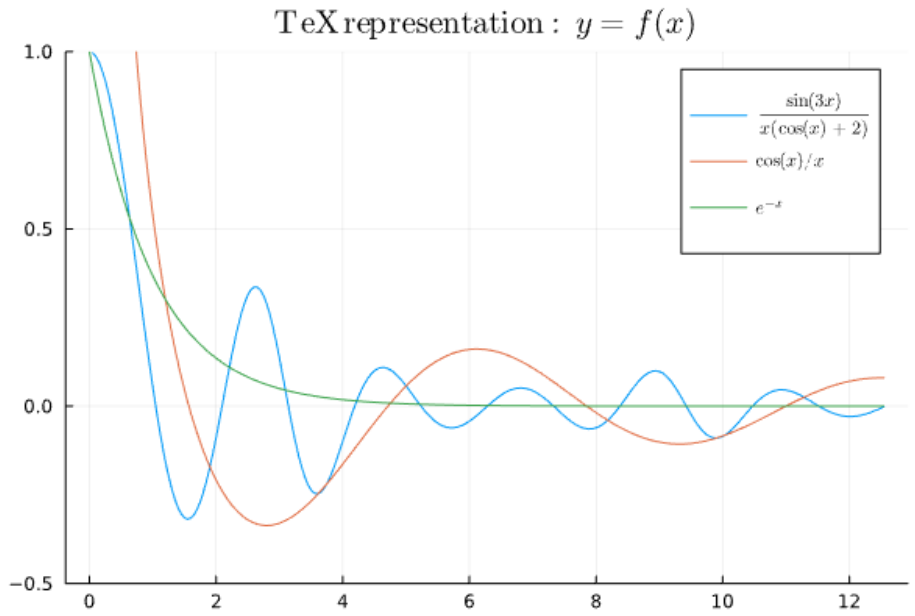
example)

```
μ = 3√6
Γ(x::Real) = √π*((x-1)/exp(1))^(x-1) * (8*(x-1)^3 + 4*(x-1)^2 + (x-1) + 1/30)^(1/6)
β = sum((x .- mean(x)) .* (y .- mean(y)))/sum((x .- mean(x)) .^ 2)
```

01. Introduction – Julia

LaTeX is a technique consisting of developing documents using clear text stylized with specific markup tags in a similar way to HTML/CSS or Markdown. This technique is mainly used for authoring scientific papers.

In Julia, we can leverage the capabilities of the packages “LaTexStrings”, “Latexify”, ... to convert a wide assortment of Julia objects to LaTeX-formatted strings.



01. Introduction – Generalized Extreme value distribution

Let $X_1, \ldots, X_n$ be a sequence of independent random variables having a common distribution function F.

$$M_n = \max\{X_1, \ldots, X_n\}$$

In applications, the X_i usually represent values of a process measured on a regular time-scale so that M_n represents the maximum of the process over n times units of observation.

If n is the number of observations in a year, then M_n corresponds to the annual maximum.

01. Introduction – Generalized Extreme value distribution

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \quad as\ n \to \infty$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$I : G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty$$

$$II : G(z) = -\exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, \quad z > b$$

$$III : G(z) = -\exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, \quad z < b$$

for parameters $a > 0,\ b$ and, in the case of families II and III,$\alpha > 0$.
I, II and III widely known as the **Gumbel**, **Frechet** and **Weibull** families respectively.

01. Introduction – Generalized Extreme value distribution

The Gumbel, Frechet and Weibull families can be combined into a single family of models having distribution functions of the form

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

μ : location parameter

σ : scale parameter

defined on the set $\{z : 1+\xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $-\infty < \mu < \infty,\ \sigma > 0$ and $-\infty < \xi < \infty$.

This is the **generalized extreme value** (GEV) family of distributions.

In hydrology the generalized extreme value (GEV) distribution is applied to extreme events such as annual maximum one-day rainfalls and river discharges.

01. Introduction – Generalized Extreme value distribution

The CDF of the Gumbel distribution is

$$P(X \leq x) = F(x) = \exp\left[-\exp\left\{-\frac{x-\mu}{\sigma}\right\}\right]$$

μ : location parameter

σ : scale parameter

Contents

02. A previous study

韓國水資源學會論文集
第45卷 第3號·2012年 3月
pp. 263~274

http://dx.doi.org/10.3741/JKWRA.2012.45.3.263

혼합 검벨분포모형을 이용한 확률강우량의 산정

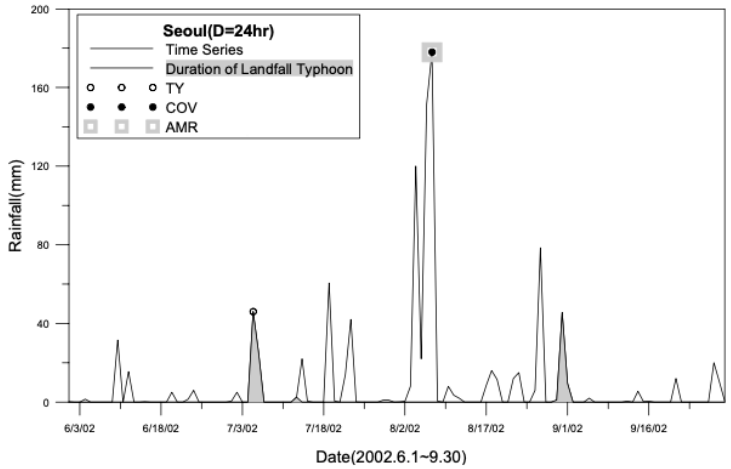
Estimating Quantiles of Extreme Rainfall Using a Mixed Gumbel Distribution Model

윤 필 용* / 김 태 웅** / 양 정 석*** / 이 승 오****

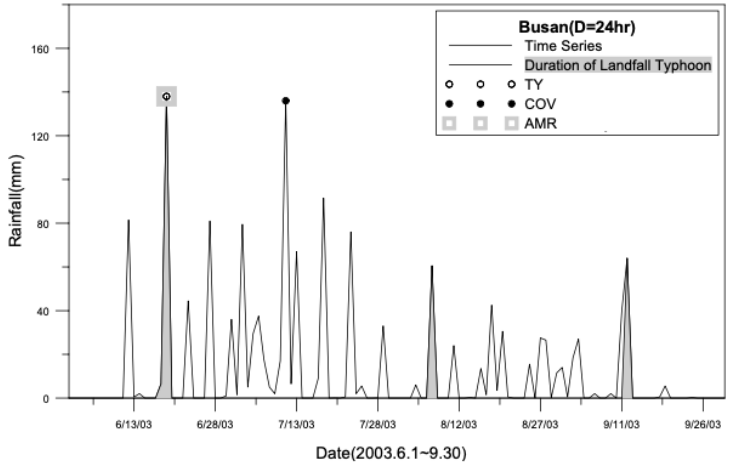
Yoon, Philyong / Kim, Tae-Woong / Yang, Jeong-Seok / Lee, Seung-Oh

Abstract

Fig. 1. Rain Gauge Stations Used in this Study



(a) Seoul Station (6.1~9.30)



(b) Busan Station (6.1~9.30)

Fig. 2. Illustration of TY, COV, AMR

02. A previous study

Table 5. Parameters for Probability Distributions

Station	Mixed Gumbel				Gumbel		p
	TY		COV		AMR		
	α_1	β_1	α_2	β_2	α	β	
Gangneung	72.98	83.01	28.46	93.65	54.52	121.71	0.388
Seoul	34.77	45.42	46.85	118.99	46.92	125.81	0.143
Incheon	38.19	43.65	42.09	99.64	44.72	106.76	0.204
Ulleungdo	35.57	43.26	20.83	71.64	26.37	78.85	0.245
Chupungnyeong	37.84	46.84	24.21	78.51	31.14	88.33	0.306
Pohang	55.53	57.24	19.71	72.57	36.93	92.98	0.469
Daegu	41.84	53.14	23.80	68.65	33.61	80.82	0.388
Jeonju	33.99	38.33	29.19	88.14	32.92	94.69	0.184
Ulsan	54.35	70.75	27.52	82.86	42.99	105.40	0.429
Gwangju	46.86	49.79	36.44	94.94	42.54	108.19	0.265
Busan	50.54	61.31	40.75	105.79	48.02	120.15	0.306
Mokpo	45.14	47.13	24.86	86.94	30.80	97.39	0.245
Yeosu	53.13	64.19	28.45	97.18	43.72	112.54	0.326
Jeju	68.28	79.51	44.38	83.48	54.21	116.21	0.449
Seogwipo	49.96	54.54	36.51	117.51	41.47	126.71	0.224

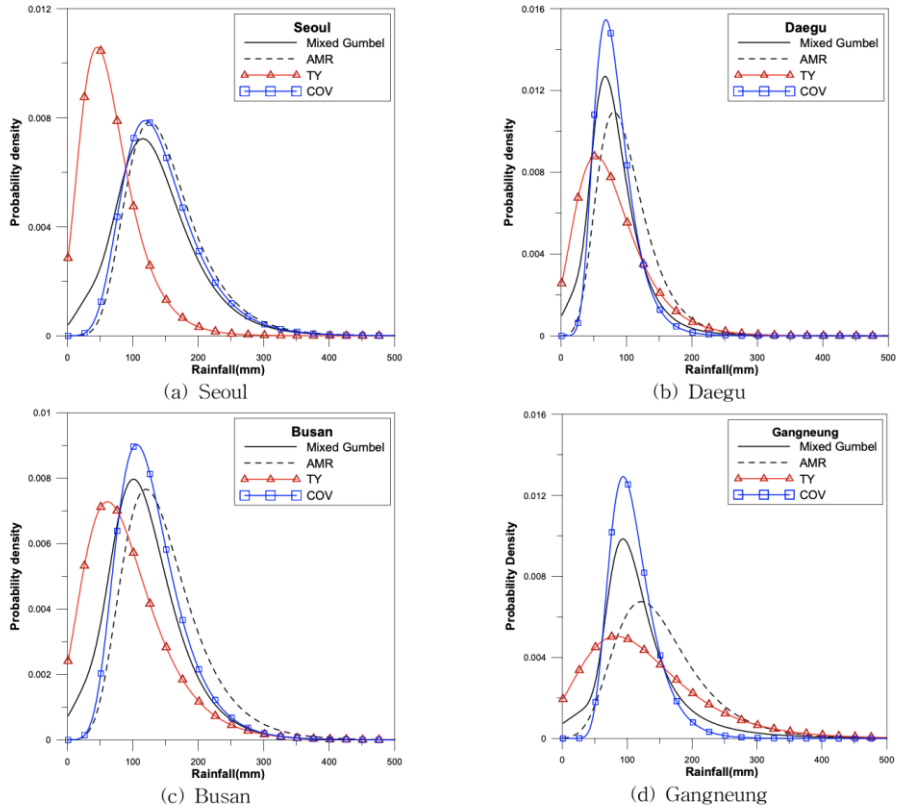


Fig. 4. Examples of Probability Density Function for Selected Stations

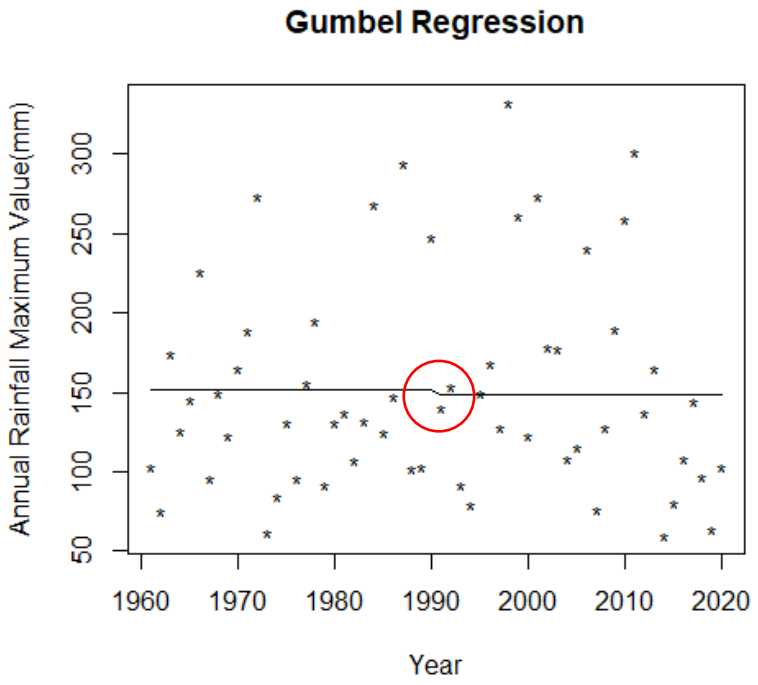
Contents

03. Mixture modeling for analyzing a rainfall pattern

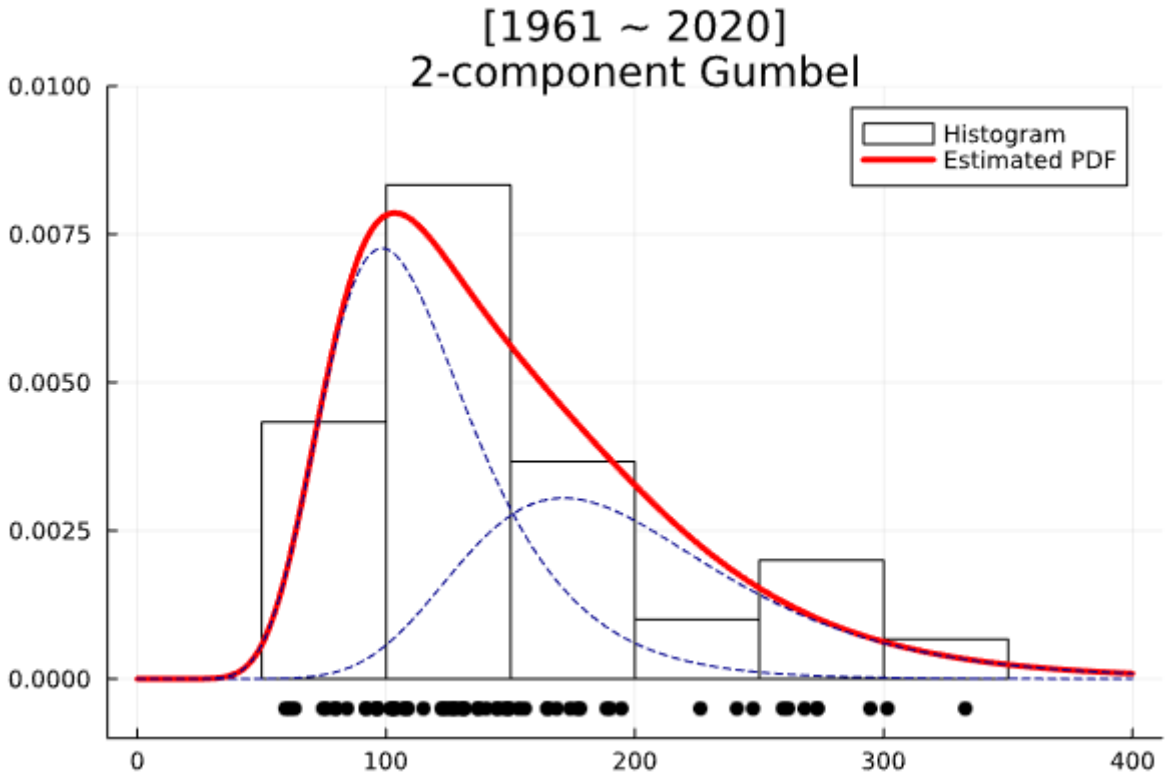
VGAM package

```
> gumbel(llocation = "identitylink", lscale="loglink",
+        iscale=NULL, R=NA, percentiles = c(95,99),
+        mpv=FALSE, zero=NULL)
Family:  gumbel
Informal classes: gumbel, vextremes

Gumbel distribution for extreme value regression
Links:    location, loglink(scale)

> gumbelff(llocation = "identitylink", lscale = "loglink",
+          iscale = NULL, R = NA, percentiles = c(95, 99),
+          zero = "scale", mpv = FALSE)
Family:  gumbelff

Gumbel distribution (multiple responses allowed)

Links:    Identity: location, Log: loglink(scale)
Mean:     location + scale*0.5772..
Variance: pi^2 * scale^2 / 6

> fit1 <- vglm(y ~ D, gumbelff(perc=NULL), data=data,trace=T)
VGLM    linear loop  1 :  loglikelihood = -330.82591
VGLM    linear loop  2 :  loglikelihood = -329.68364
VGLM    linear loop  3 :  loglikelihood = -329.641
VGLM    linear loop  4 :  loglikelihood = -329.6404
VGLM    linear loop  5 :  loglikelihood = -329.64039
VGLM    linear loop  6 :  loglikelihood = -329.64039
```

```
> coef(fit1)
(Intercept):1 (Intercept):2             D
   123.100176      3.883757     -2.727949
```



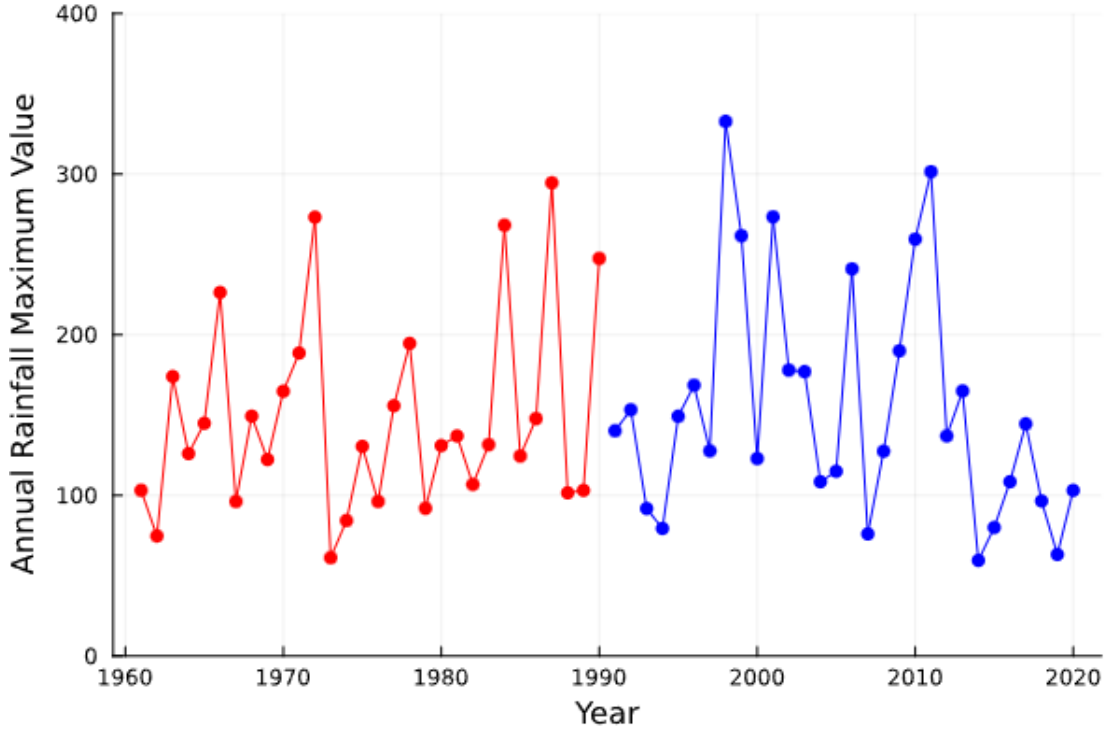
Annual Maximum Rainfall Data in Seoul

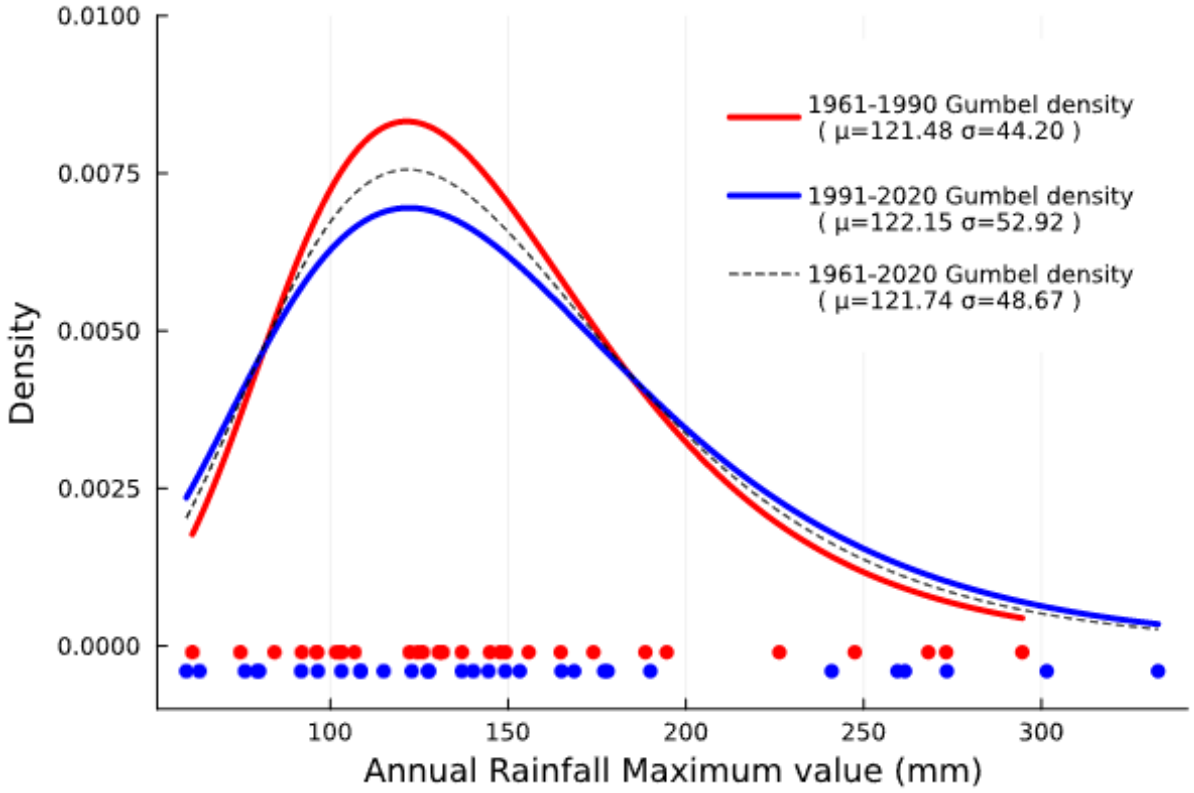


03. Mixture modeling for analyzing a rainfall pattern

Annual Maximum Rainfall Data in Seoul

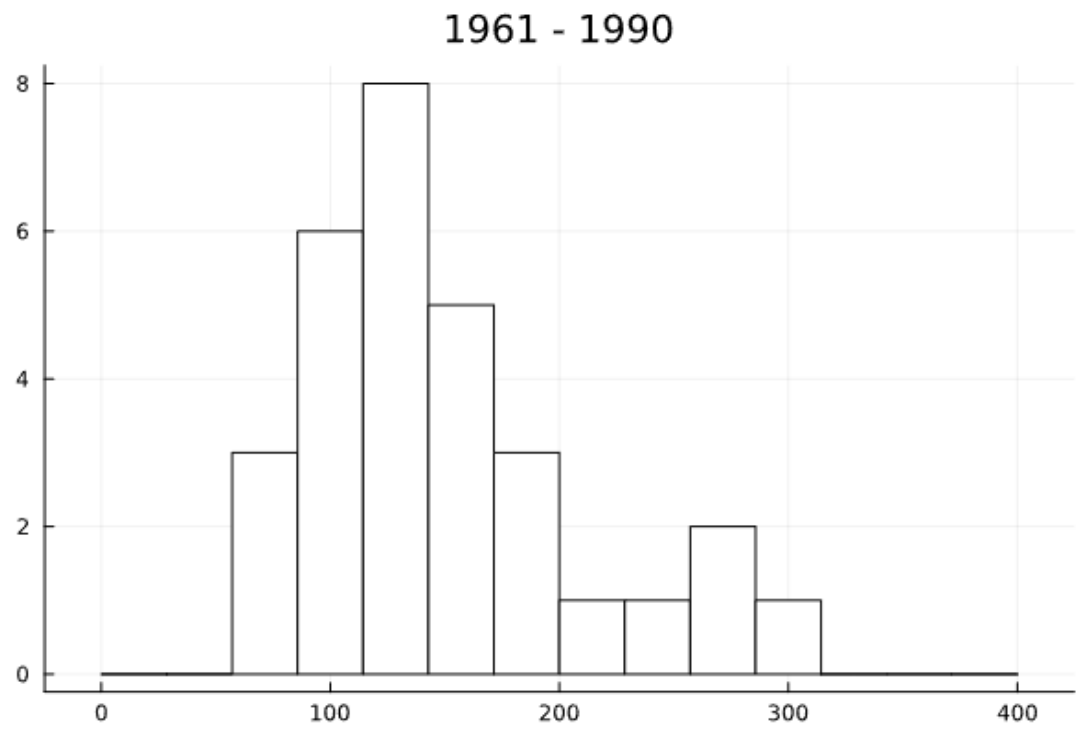
Series plot

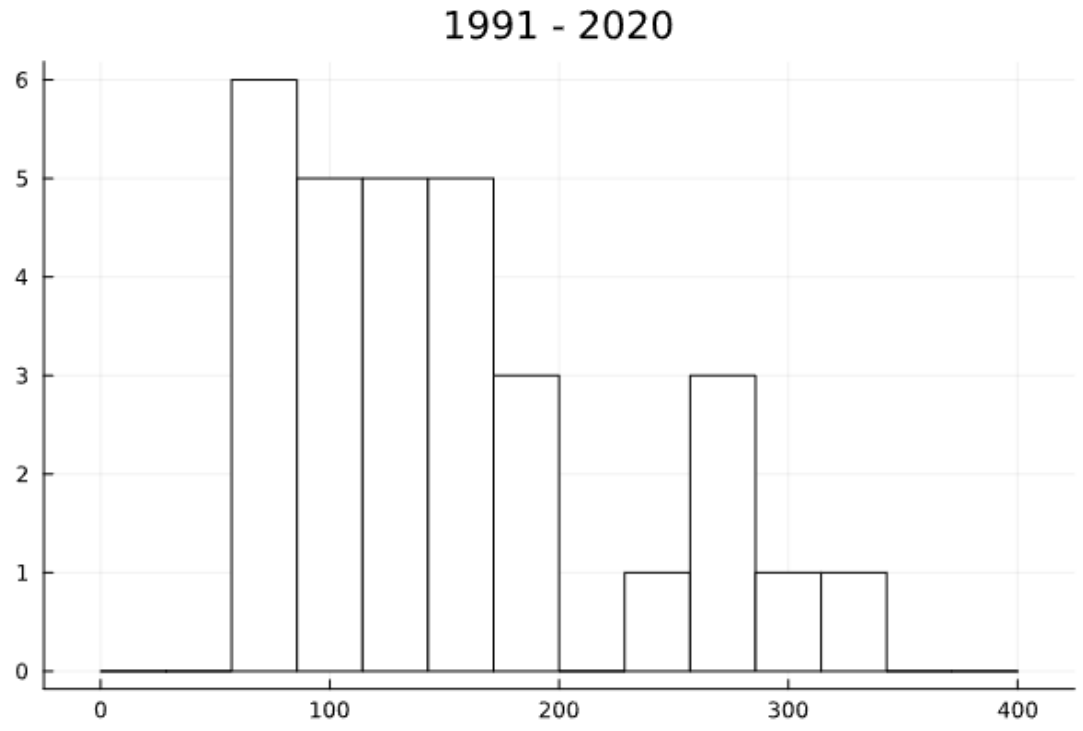




Annual Maximum Rainfall Data in Seoul

Histogram

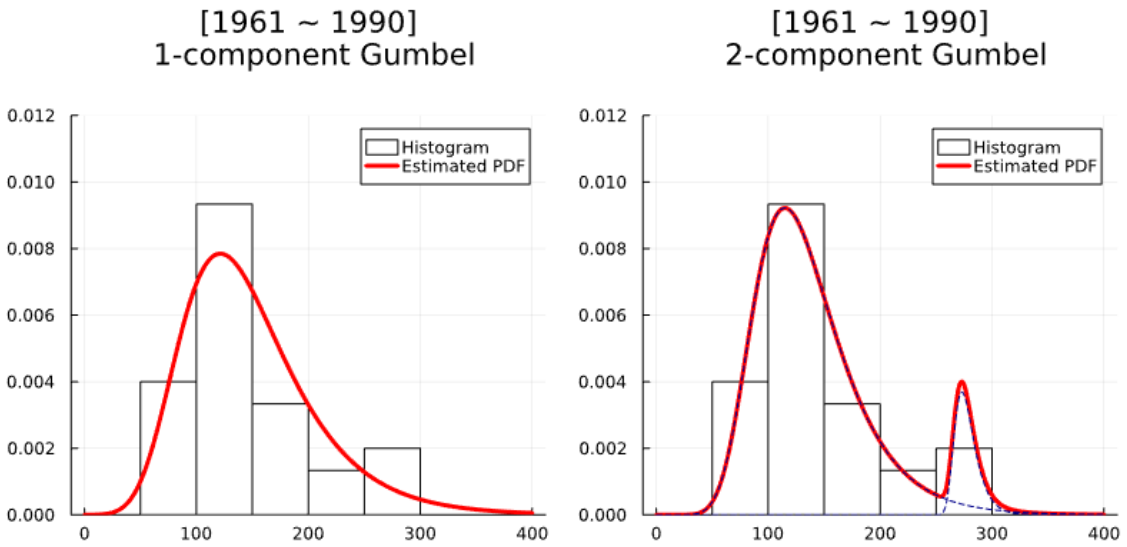


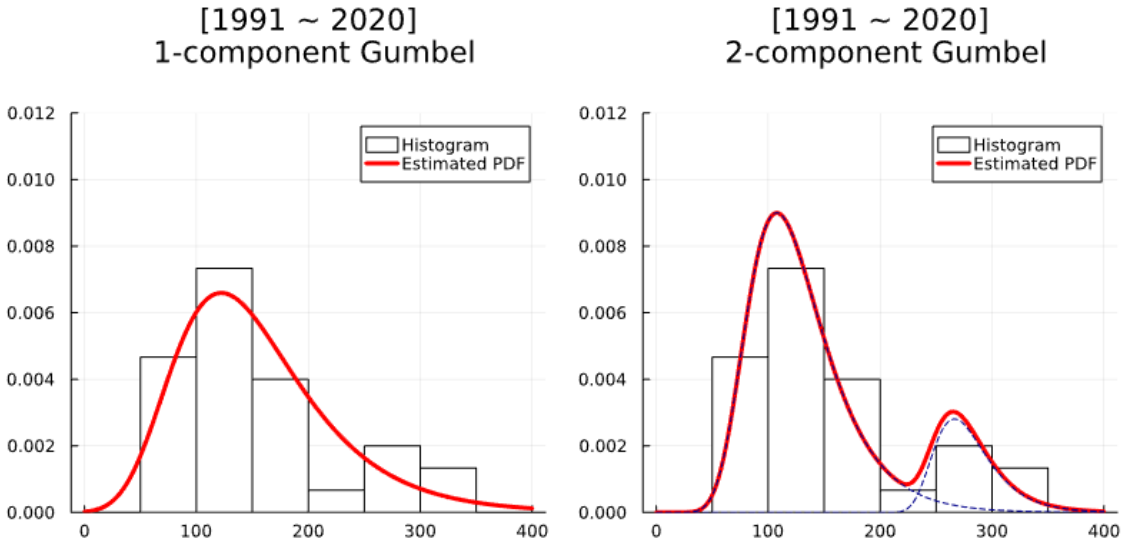


03. Mixture modeling for analyzing a rainfall pattern

Gumbel distribution

		1-component	2-component
[1961 ~ 1990]	-2LL	324.02	318.56
	AIC	328.02	328.56
	BIC	330.83	335.56
[1991 ~ 2020]	-2LL	334.83	329.23
	AIC	338.83	339.23
	BIC	341.64	346.23

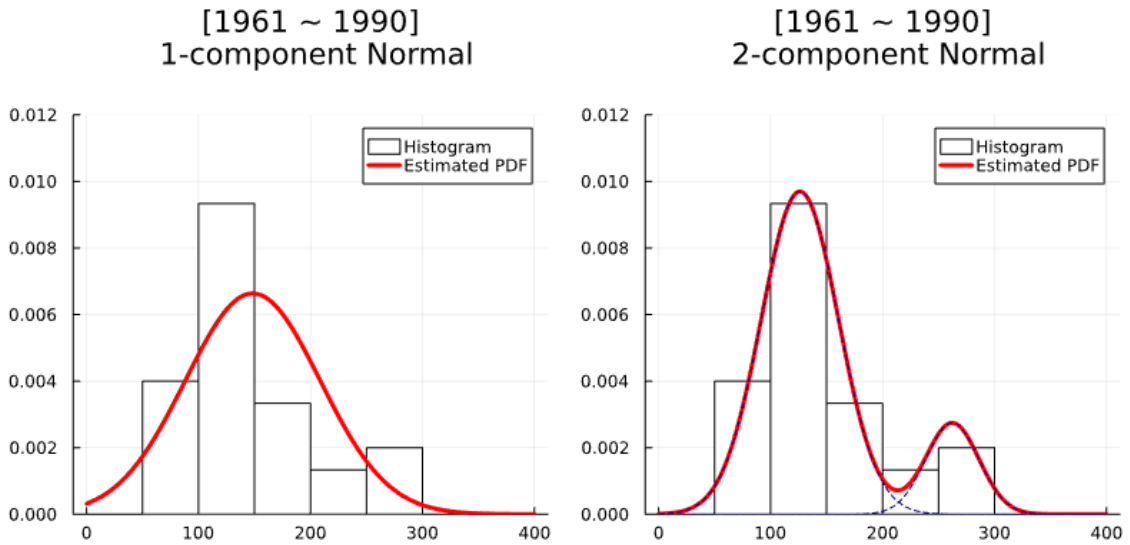


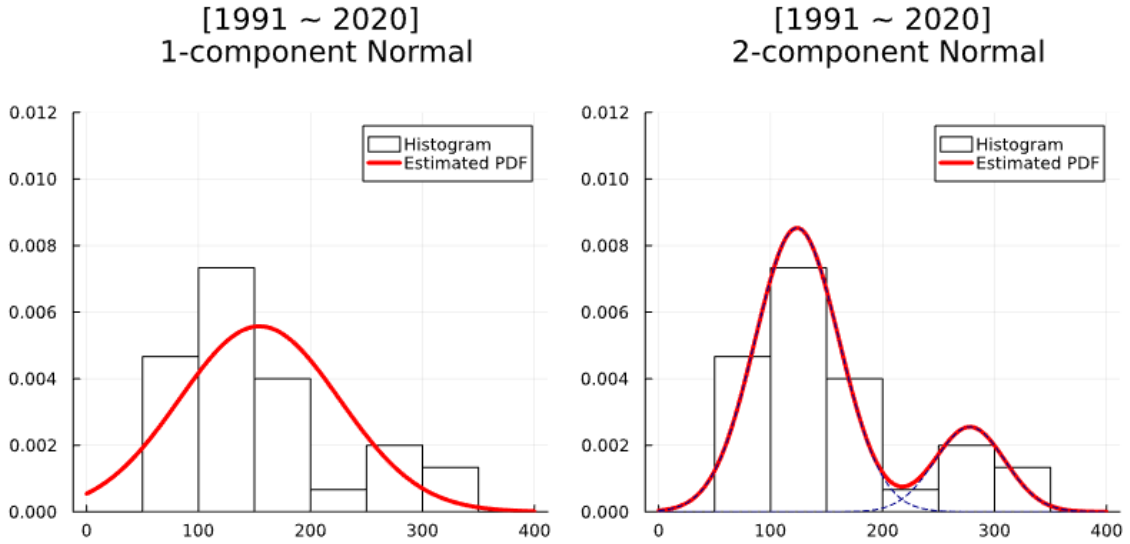


03. Mixture modeling for analyzing a rainfall pattern

Normal distribution

		1-component	2-component
[1961 ~ 1990]	-2LL	330.94	319.17
	AIC	334.94	329.17
	BIC	337.74	336.18
[1991 ~ 2020]	-2LL	341.38	329.39
	AIC	345.38	339.39
	BIC	348.18	346.39

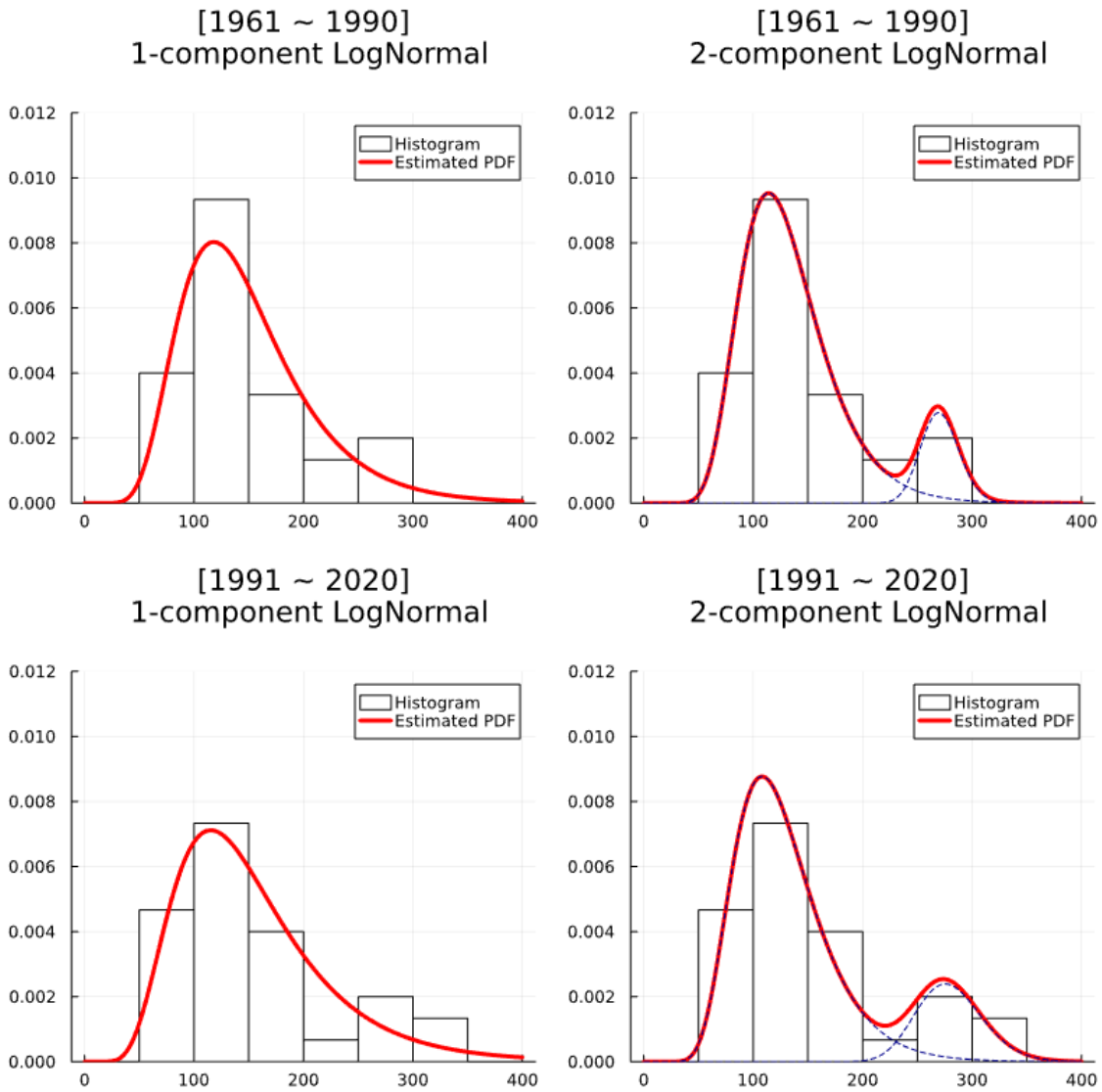




03. Mixture modeling for analyzing a rainfall pattern

Lognormal distribution

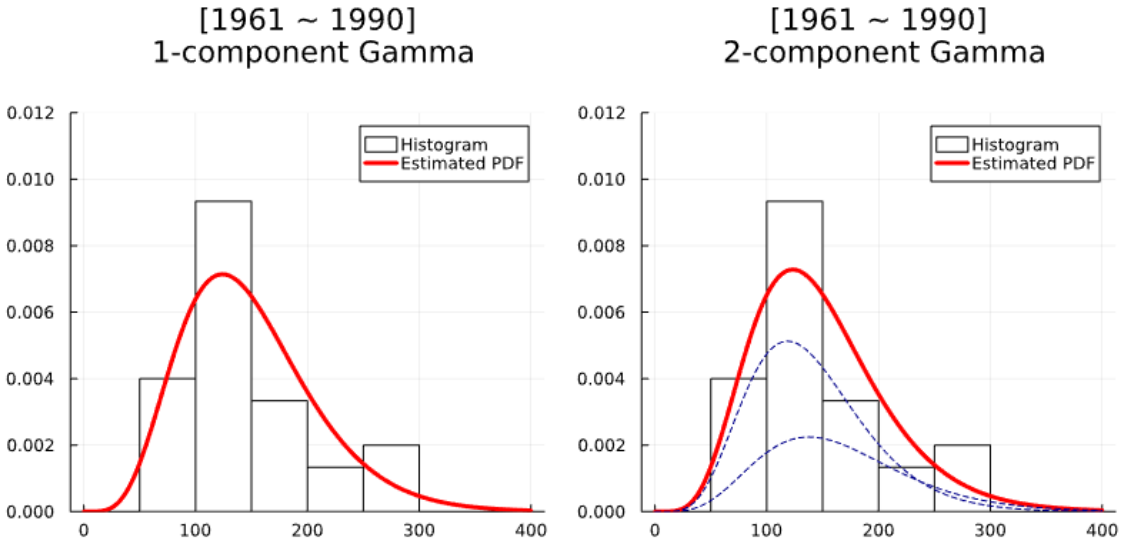
		1-component	2-component
[1961 ~ 1990]	-2LL	323.49	318.33
	AIC	327.49	328.33
	BIC	330.29	335.34
[1991 ~ 2020]	-2LL	333.73	329.34
	AIC	337.73	339.34
	BIC	340.53	346.35

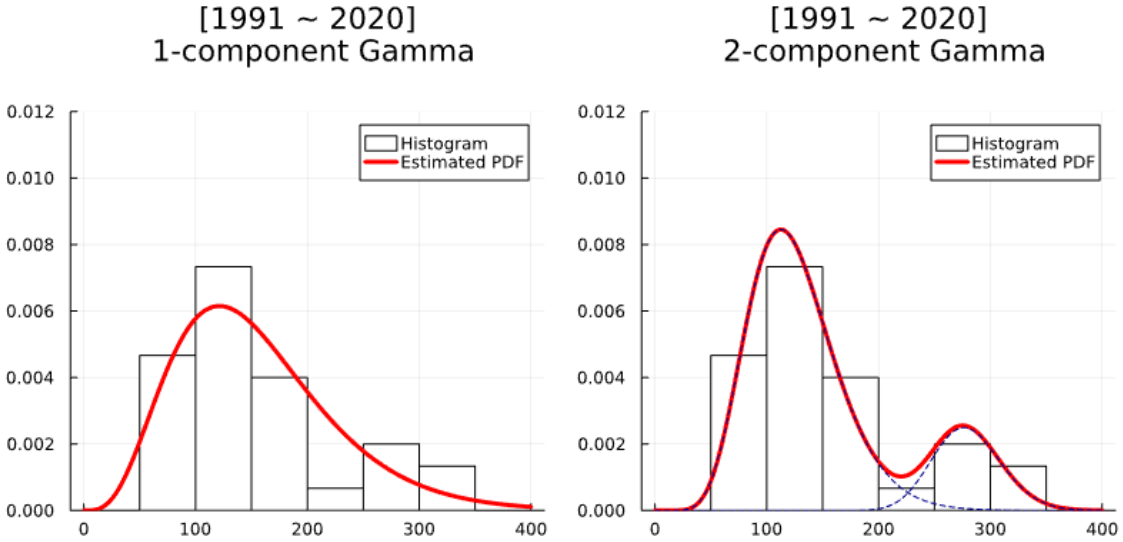


03. Mixture modeling for analyzing a rainfall pattern

Gamma distribution

		1-component	2-component
[1961 ~ 1990]	-2LL	324.96	324.96
	AIC	328.96	334.71
	BIC	331.76	341.72
[1991 ~ 2020]	-2LL	334.86	328.94
	AIC	338.86	338.94
	BIC	341.66	345.95

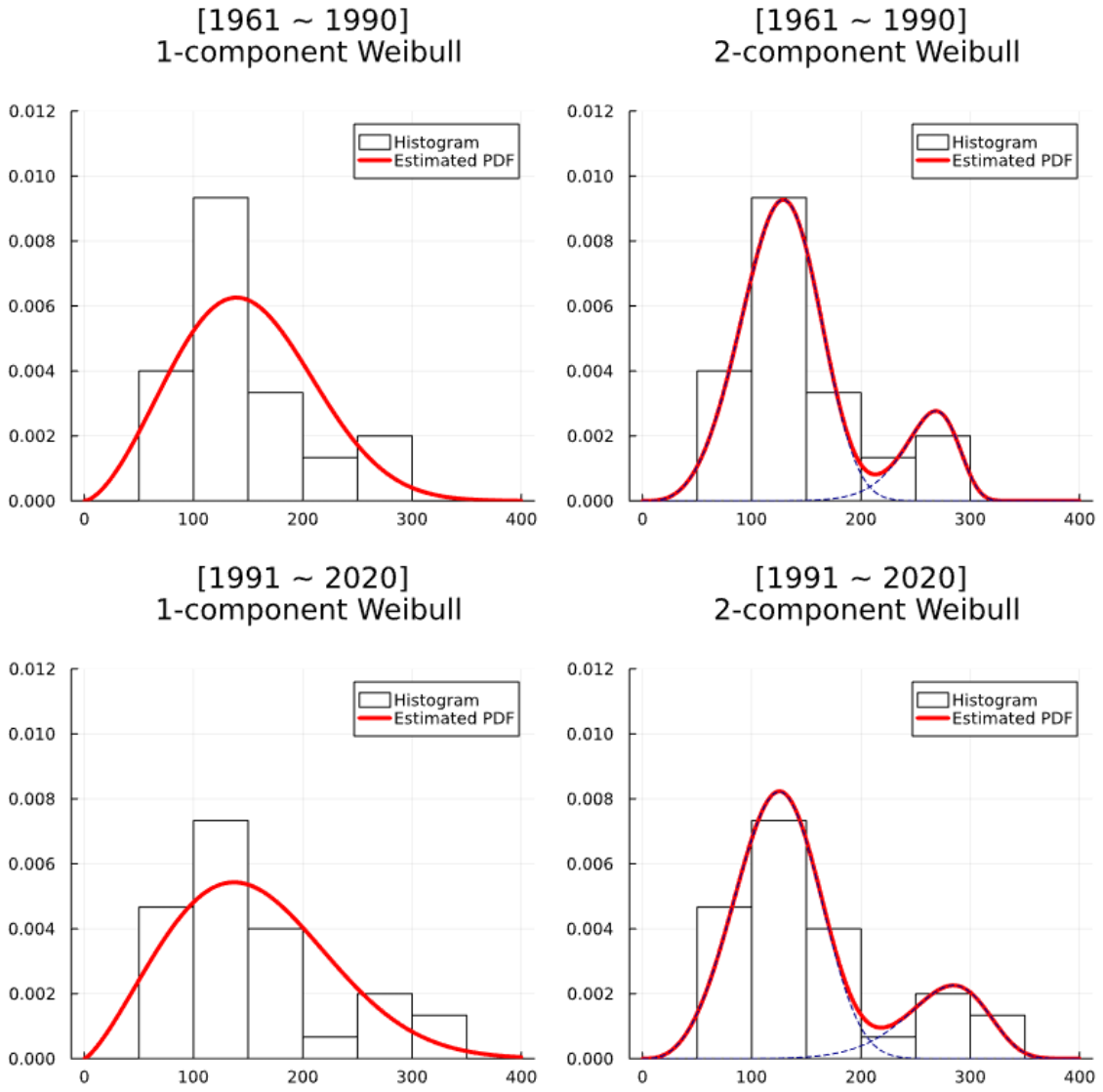




03. Mixture modeling for analyzing a rainfall pattern

Weibull distribution

		1-component	2-component
[1961 ~ 1990]	-2LL	328.6	319.1
	AIC	332.6	329.1
	BIC	335.4	336.1
[1991 ~ 2020]	-2LL	337.6	329.4
	AIC	341.6	339.4
	BIC	344.4	346.4



Contents

04. Discussion

In a previous studies in which the annual maximum rainfall were fit with the 2-component Gumbel mixture. These studies modeled at one time with about half a century of data.

Alternately, we modeled each in a slightly different way by dividing the data in half. When fitting in half, the mean of distribution slightly increased and the variance greatly increased, indicating that the extreme values had larger values than in the previous period.

The AIC and BIC values were smaller when the model fitting results were generally fit with one distribution than the mixture distribution. This seems to be due to the small numbers of data.

Also, the **Lognormal** distribution tended to fit better than the Gumbel distribution used in previous studies.

04. Discussion

And while using the Julia language, We could feel the advantages of the convenience and speed of the Julia Language. As mentioned earlier, it is faster than R and has a syntax that is as easy as Python.

In the future, we will try to the mixture model with various combinations of distributions such as Lognormal-Gumbel mixture, as well as fitting only two identical distributions.

Because the mixture models were fit by dividing the dataset in half, there was a lack of research on where the structural change appeared. Therefore, further research will be conducted in this point.

References

1. 윤필용 외, 「혼합 Gumbel 분포를 이용한 태풍의 설계강우량에 미치는 영향 평가」, 한국 방재학회 학술대회논문집, (2011), p.45
2. 윤필용 외, 「혼합 검벨분포모형을 이용한 확률강우량의 산정」, 한국수자원학회논문집, vol.45(2012), p.263-274
3. 최홍근 외, 「Bayesian 기법을 이용한 혼합 Gumbel 분포 매개변수 추정 및 강우빈도해석 기법 개발」, 대한토목학회논문집(국문), vol.38(2018), p.249-260
4. 6. Danjuma, I. 1Yahaya, A. and 2Asiribo, O. E. (2020). Fitting and Comparing Gamma, Lognormal and Weibull Distributions Using Nigeria Rainfall Intensity Data. Journal of Climate Studies, Volume 6 ~ Issue 5, p.18-24
5. Wei Lun Tan, Woon Shean Liew, and Lloyd Ling. (2021). Statistical modelling of extreme rainfall in Peninsular Malaysia. ITM Web of Conferences, vol.36(2021)
6. Gökçen ERYILMAZ TÜRKKAN, Tuğçe HIRCA. (2019)Statistical Modeling of Annual Maximum of Precipitation Data: A Case Study of Bayburt Province. 3rd International Conference on Advanced Engineering Technologies
7. Md Ashraful Alam et al. (2018). Best-Fit Probability Distributions and Return Periods for Maximum Monthly Rainfall in Bangladesh. Cliimate.
8. Gökçen ERYILMAZ TÜRKKAN, Tuğçe HIRCA. (2019). Statistical Modeling of Annual Maximum of Precipitation Data: A Case Study of Bayburt Province. 3rd International Conference on Advanced Engineering Technologies
9. M. T. Amin, M. Rizwan, A. A. Alazba.(2015). A best-fit probability distribution for the estimation of rainfall in northern regions of Pakistan. Special Issue on CleanWAS. P.432- 3440


Thank you for your attention