*The Korean Journal of Applied Statistics* (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.6.947
**27**(6), 947–958

# A Bayesian Analysis of Return Level for Extreme Precipitation in Korea

Jeong Jin Lee[a] · Nam Hee Kim[a] · Hye Ji Kwon[b] · Yongku Kim[a,1]

[a]Department of Statistics, Kyungpook National University; [b]Statisitcs Korea

(Received July 29, 2014; Revised August 18, 2014; Accepted August 29, 2014)

**Abstract**

Understanding extreme precipitation events is very important for flood planning purposes. Especially, the $r$-year return level is a common measure of extreme events. In this paper, we present a spatial analysis of precipitation return level using hierarchical Bayesian modeling. For intensity, we model annual maximum daily precipitations and daily precipitation above a high threshold at 62 stations in Korea with generalized extreme value(GEV) and generalized Pareto distribution(GPD), respectively. The spatial dependence among return levels is incorporated to the model through a latent Gaussian process of the GEV and GPD model parameters. We apply the proposed model to precipitation data collected at 62 stations in Korea from 1973 to 2011.

Keywords: Bayesian analysis, daily precipitation, extremes, generalized extreme value distribution, generalized Pareto distribution, return level, spatial process.

## 1. 서론

최근 지구온난화와 같은 전 지구적인 기상현상으로 인한 강수의 패턴 분석과 예측에 관한 연구가 활발하게 이루어지고 있다. 예를 들어, 대기 중 이산화탄소($CO_2$) 농도 증가와 같은 기후조건 변화에 따른 대기 순환 모형 모의 결과는 온실 효과와 함께 강수 강도의 증가를 예측하고 있는데, 이는 지구온난화에 따른 강한 대류성 강수의 증가와 대규모 비대류성 강수의 감소에 기인한다고 해석되어지고 있다. 우리나라의 경우, 수문자료 분석에 있어 재현기간에 따른 최대 강수량 예측이 홍수조절, 관개용수 관리 등에 밀접한 관계를 가지고 있는데, 집중호우에 의한 홍수 피해가 증가함에 따라 국내 강수자료의 체계적인 분석을 통한 강수량의 비정상성에 대한 해석과 강수의 비정상성이 반영된 설계 강수량 산정 방법의 개발이 시급한 실정이다.

이상 기후와 같은 익스트림 이벤트란 일반적으로 아주 큰 값을 가지지만 그 발생 빈도는 상대적으로 아주 작은 사건을 말한다. 이러한 익스트림 이벤트는 자주 일어나지는 않지만 많은 관심을 갖게하는 이유은 우리에게 미치는 영향이나 피해가 일반적인 사건에 비해서 크기 때문이다. 최근에 익스트림 이벤트에 관한 다양한 연구가 진행되고 있고 특히 통계학 분야에서는 이러한 사건에 대한 분포와 그 분포에 관련된 추론에 관한 연구가 활발하게 진행되고 있다. 일반적으로 익스트림 이벤트에서의 주된 관심은 언

[1]Corresponding author: Department of Statistics, Kyungpook National University, 80 Daehakro, Daegu 702-701, Korea. E-mail: kim.1252@knu.ac.kr

제 이러한 사건이 일어날 것인지 그리고 얼마나 큰 사건이 일어날 지에 있는데, 특히 두번째 물음에 관한 답을 반환주기를 통해서 알아볼 수 있다. 반환주기는 지진이나 홍수와 같은 익스트림 이벤트의 발생 가능성에 대한 추정치이며, 일반적으로 장기간에 걸친 평균 반복 간격을 나타내는 기록 데이터에 기초하여 작성된 통계치이다. 따라서 특정 위험 구역에서 프로젝트를 진행할 때, 특정 반환주기과 이벤트를 견딜 수 있도록 구조를 설계한다. 여기에서 발생하는 이벤트의 확률은 시간에 걸쳐 변화하지 않는 것으로 가정하고 과거의 이벤트와는 독립이라고 가정한다. 이론적인 반환주기는 이벤트가 특정 수준을 초과할 확률의 역수로 표현된다. 예를 들어, 10년 홍수란 어느 한 해에 어떤 수준을 초과할 강수량이 발생할 확률이 0.1 (또는 10%)인 경우이고 50년 홍수는 그 확률이 0.02 (또는 2%)인 강수량을 나타낸다. 하지만 100년 홍수가 100년에 정기적으로 발생하거나, 100년 동안 한 번만 발생한다는 의미는 아니다. 즉, 이름이 "반환주기"임에도 불구하고 실제 어떤 주어진 100년의 기간 동안 한 번 이상 발생할 수도 있고 한 번도 발생하지 않을 수 있다.

실제 국내에서도 이미 비정상 강수 빈도 해석에 관한 활발한 연구가 진행 되고 있으며, Jang 등 (2011)은 연 최대 강수량의 회귀직선에 대한 잔차의 수문학적 빈도 해석을 바탕으로, 가까운 미래로 설정된 목표연도의 확률 강수량을 산정하는 방법을 제안하였으며, Lee (2010)는 우리나라(대한민국) 강수량계열의 특성 평가를 통한 경향성을 고려한 비정상성 빈도해석을 위한 모형을 제시하는 등 강수계열의 비정상성 평가 및 비정상성 빈도 해석 기법을 제시 하였다. Strupczewski 등 (2001)은 Akaike information criterion(AIC)의 관점에서 56개의 모델 중 비정상 홍수 빈도에 관한 최적화 모형을 비교분석하였다. Cunderlik 등 (2008)은 지점에 대한 홍수 빈도 모형에 대한 비정상 접근법을 제안하였다. 통계학 분야에서는 극단값 분포에 관한 연구가 주로 이루어져 왔는데, 통계적인 관점에서의 여러가지 응용에 대한 일반적인 이론과 분석 방법은 Embrechts 등 (1997)과 Coles (2001)에서 참고할 수 있다.

이전까지는 주로 단변량 극단값에 관한 여러가지 통계모형 및 분석방법이 주로 연구되어져 왔다면 최근에는 이러한 모형이나 방법을 다변량 극단값 또는 공간구조를 가지는 극단값으로 확장하는 연구가 활발하게 진행되고 있다. 다변량 극단값 또는 공간구조를 가지는 극단값에 관련된 연구는 주로 이들 극단값들 사이의 종속성을 어떻게 설명하는지에 중점을 두고 진행되어 왔는데 이러한 연구 결과는 집중호우와 같은 극단기후사건들의 모형에 적용될 수 있다. 다변량 극단값 분포에 관한 연구는 de Haan (1985), Coles 등 (1999), Schlather와 Tawn (2003), Heffernan과 Tawn (2004)에서 참고할 수 있다. 최근에는 베이지안 모형에 관한 연구가 활발해지면서 베이지안 계층구조를 이용하여 다변량 극단분포를 구성하는 방법에 관한 연구가 진행되었는데, Cooley 등 (2006)은 베이지안 계층구조를 가지는 Generalized Extreme Value 모형을 소개하였고, Casson과 Coles (1999)는 어떤 지점에서의 특정값 이상의 극단값에 대한 공간모형을 연구하였다. 최근 Cooley 등 (2007)은 극단값분포의 모수에 공간구조를 적용하고 이를 통한 공간구조를 가지는 극단값의 분포를 구성하고 이를 통해 집중호우에 관한 반환주기에 대한 공간 구조를 구현하였다. 특히 반환주기는 극단값 분포의 모수들의 함수 형태로 표시되는데 베이지안 모형을 이용하여 이 반환주기의 분포를 구할 수 있다.

국내 강우자료를 이용한 연구에서는 Kwak과 Kim (2014)가 공간구조를 가지는 GLM weather generator 모형을 소개하였고 그 밖에 여러가지 공간구조를 가지는 확률적 강우모형에 관한 연구가 시도되었지만 공간구조를 가지는 극단값 분포나 반환주기에 관한 연구는 여전히 미진한 상황이다. 본 논문에서는 1973년부터 2011년까지 39년 동안의 우리나라 62개 기상 관측소에서 관측된 일일 강우량 자료를 사용하여 대표적인 두가지 극단값 분포인 generalized extreme value(GEV) 분포와 generalized Pareto(GP) 분포에 적합시키고 계층적 베이지안 모형을 이용하여 이들 분포의 모수들에 공간구조를 소개한 후 이를 통해 우리나라 전 지역에 대한 강우 극단값에 대한 반환주기에 대한 지도를 완성하였다.

## 2. 극단값 분포

극단값의 분포이론은 주로 확률분포의 꼬리부분에 대한 통계모형을 제공하는데 일반적인 단별량 극단값에 대한 분포에 대한 이론은 잘 알려져 있고 이는 근사적인 분포인 generalized extreme value(GEV) 분포를 이용해서 설명된다. 이는 독립적이고 동일한 분포를 가진 극단값을 무작위로 샘플링 했을 때, 그 분포가 Gumbel, Frechet 그리고 Weibull 분포 중에 하나를 따른다는 극단값이론과 그 맥을 같이 한다. 일반적으로 극단값이론은 확률분포의 꼬리의 형태에 대한 통계적 이론으로 극단적인 사건에 적합한 보다 나은 분포를 제공하기 때문에 위험분석에 있어서는 좀 더 유연성을 가질 수 있다. 여러분야에서 응용되는 극단값 이론모형은 기본적으로 두 개의 모형으로 나눌 수 있다. 우선 동등한 크기로 전체 표본을 나누어 각 블록의 최대값(block maxima) 혹은 최소값으로 구성된 관측치들의 분포를 추정하는 모형과 높은 임계값(threshold)을 초과한 모든 관측값에 대한 모형이다. 일반적으로 전자의 관측값에 대해서는 GEV 분포를 가정하고 후자의 관측값에 대해서는 generalized Pareto 분포를 가정한다. GEV 분포는 Gumbel, Frechet 그리고 Weibull 분포를 통합적인 모형으로 표현한 것으로 분포함수는 다음과 같이 표현할 수 있다.

$$F(z) = P(Z < z) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, & \xi > 0, \\ \exp\left\{-\exp\left(\frac{z-\mu}{\sigma}\right)\right\}, & \xi = 0, \end{cases} \tag{2.1}$$

여기에서 $\mu$는 위치모수이고, $\sigma$는 척도모수이며, $\xi$는 형상모수이다. 특히 GEV 분포는 $\xi > 0$일 때 Frechet 분포가 되고, $\xi < 0$일 때 Weibull 분포가 되며, $\xi = 0$일 때 Gumbel 분포가 된다.

높은 임계값을 초과하는 모든 관측값에 대한 GPD 분포는 일반적으로 다음과 같은 조건부 초과분포함수의 형태로 표현된다.

$$P(Z > z + u | Z > u) = \left(1 + \xi\frac{z}{\sigma_u}\right)_+^{-\frac{1}{\xi}}, \tag{2.2}$$

여기에서 척도모수인 $\sigma_u > 0$는 일반적으로 $\sigma + \xi(u - \mu)$의 형태로 표현되기도 하며, 형상모수인 $\xi$는 분포의 꼬리 부분이 제한되는지($\xi < 0$), 얇은지($\xi \to 0$), 아니면 두꺼운지($\xi > 0$)를 결정한다. 일반적으로 임계값 초과치 자료를 GPD 모형 적용에 앞서 각 시계열에 적합한 임계값을 산정하는 것은 중요한 문제이다. Pickands-Balkema-de Hann 정리를 만족하기 위해서는 임계값이 높게 설정되어야 한다. 그러나 너무 큰 임계값은 자료의 수가 너무 작아서 모수 추정시 정확하게 되지 않는다. 또한 너무 작은 임계값을 선택하면 관찰치가 늘어나서 모수의 추정치가 좀 더 정확해질 수 있으나 분포의 중심에 가까운 값을 사용하게 되어 추정된 모수들의 편의가 증가하게 된다. 임계값을 산정하기 위하여 평균초과함수(Mean Excess Function), Hill 방법 및 Oliveria 등 (2006)의 방법을 고려할 수 있다. 본 연구에서는 전기간 강우량 계열의 상위 98%에 해당하는 값으로 임계값으로 가정하고 적합성을 검토한 후 연구에 이용하였다 (Lee, 2010). 참고로 Figure 2.1에서는 기상청 (Korea Meteorological Administration, KMA)으로부터 얻은 1973년 1월 1일부터 2011년 12월 31일까지로 39년동안 서울지역에서 측정된 일별 강우량 관측자료를 사용하여 연간 블록 최대값과 임계값 초과치 자료를 표시한 것이다.

앞에서 소개한 극단값 분포에 대한 반환주기는 각 극단값 분포의 모수의 함수형태로 표현되는데 $r$년 반환주기는 이 특정 수준을 초과하는 이벤트가 일어날 확률이 $1/r$인 경우이다. 일반적으로 GEV 분포에 대한 $r$년 반환주기는 다음과 같이 표현된다.

$$z_r = \mu - \frac{\sigma}{\xi}\left\{1 - \left[-\log\left(1 - \frac{1}{r}\right)^{-\xi}\right]\right\}, \tag{2.3}$$

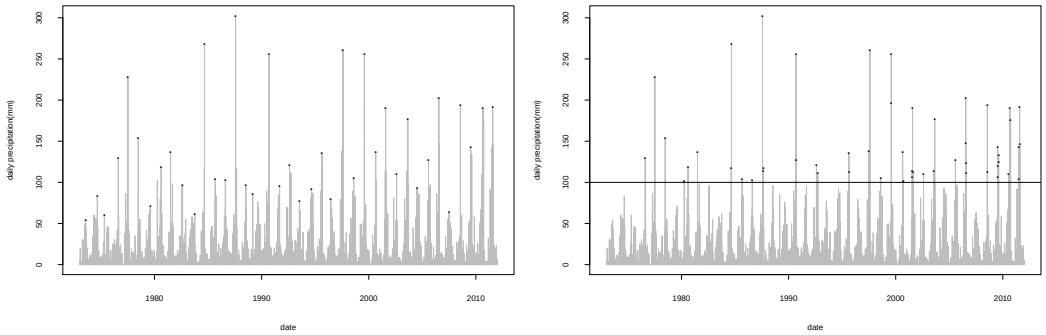


**Figure 2.1.** Plots of annual maximum daily precipitations (left) and daily precipitations of threshold exceedances (right) in Seoul.

여기에서 $\mu$, $\sigma$와 $\xi$는 GEV 분포의 모수이다. GPD 분포의 경우에는 식 (2.2)는 다음과 같이 표현할 수 있다.

$$P(Z > z + u) = p_u \left(1 + \xi \frac{z}{\sigma_u}\right)^{-\frac{1}{\xi}}, \tag{2.4}$$

여기에서 $p_u = P(Z > u)$이고 $P(Z > z_r) = 1/rn_y$이다. 참고로 $n_y$는 1년 중에 관측된 관측값의 수이다. 따라서 GPD 분포에 대한 $r$년 반환주기는

$$z_r = u - \frac{\sigma_u}{\xi}\left\{(rn_y p_u)^{\xi} - 1\right\}, \tag{2.5}$$

여기에서 $\sigma_u$와 $\xi$는 GPD 분포의 모수이다. 따라서 극단값 분포에 대한 반환주기는 각 블록의 최대값(block maxima)이나 높은 임계값(threshold)을 초과한 모든 관측값의 분포를 추정하여 추정된 각 극단값 분포의 모수를 이용하여 구할 수 있다.

## 3. 베이지안 모형

2장에서 소개된 기존의 반환주기에 공간구조를 적용하기 위해서 베이지안 모형에서 많이 사용되는 계층구조를 이용하였다. 반환주기의 분포에 대한 모수적인 정보가 부족하므로 반환주기에 직접 공간구조를 적용하는 대신에 각 극단값 분포의 모수에 공간구조를 가지는 다변량 정규분포를 가정함으로써 간접적으로 반환주기에 공간적 연관성을 소개하였다. 즉, 관측지점 $s$의 $r$년 반환주기 $z_r(s)$는 각 관측지점 $s$의 관측 극단값들에 대한 극단값 분포의 모수들인 위치모수 $\mu(s)$, 척도모수 $\sigma(s)$ 그리고 형상모수 $\xi(s)$의 함수로 표현된다. 따라서 이들 극단값 분포의 모수들에 공간구조를 가지는 사전분포를 가정하고자 한다. 위치모수 $\mu(s)$에 대해서는 평균이 $\mathbf{m}_\mu$이고 공분산 함수로 $C_\mu(\cdot,\cdot)$를 가지는 정규확률과정을 가정하였고, 척도모수 $\sigma(s)$와 형상모수 $\xi(s)$에 대해서는 log 변환한 모수값들에 대해서 평균이 $\mathbf{m}_\sigma$와 $\mathbf{m}_\xi$이고 공분산 함수로 $C_\sigma(\cdot,\cdot)$와 $C_\xi(\cdot,\cdot)$를 각각 가지는 정규확률과정을 가정한다. 즉,

$$\mu(\,\cdot\,) \sim \mathrm{MVN}(\mathbf{m}_\mu, \mathbf{\Sigma}), \tag{3.1}$$

$$\log\sigma(\,\cdot\,) \sim \mathrm{MVN}(\mathbf{m}_\sigma, \mathbf{\Psi}), \tag{3.2}$$

와

$$\log\xi(\,\cdot\,) \sim \mathrm{MVN}(\mathbf{m}_\xi, \mathbf{\Omega}). \tag{3.3}$$

참고로 $\mathbf{\Sigma}$, $\mathbf{\Psi}$ 그리고 $\mathbf{\Omega}$는 공간구조를 가지는 공분산 행렬이고 분산값과 상관행렬의 곱의 형태로 나타난다. 즉, $\mathbf{\Sigma} = \sigma^2 R_{\rho,\nu}(s, s')$. 평균 $\mathbf{m}_\mu$, $\mathbf{m}_\sigma$ 그리고 $\mathbf{m}_\xi$는 다른 보조변수를 이용하여 모형화할 수 있는데, 예를 들어,

$$\mathbf{m}_\mu = f(\alpha, \text{covariates}(s)) = \alpha_0 + \alpha_1(\text{elevation}).$$

가장 일반적으로 사용되는 공분산함수는 Matérn covariance function인데 다음의 형태를 가진다 (Matérn, 1986).

$$R_{\rho,\nu}(s, s') = \frac{\left(\frac{d(s,s')}{\rho}\right)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu \left(\frac{d(s, s')}{\rho}\right), \tag{3.4}$$

여기에서 $\rho$는 척도모수이고 $\nu$은 smoothness의 정도를 나타내는 모수이다. 이 모수를 적절히 조절하면 Matern의 특수한 형태의 공분산 구조를 얻을 수 있다. 특히 $\nu = 0.5$일 경우에는 일반적으로 잘 알려진 지수모형을 가진다. 더 일반적인 시공간모형을 고려하기 위해서는 다음과 같은 공분산 함수를 고려할 수도 있다.

$$C_\mu^t(s_1, s_2) \propto \exp\left(\frac{- \parallel s_1 - s_2 \parallel}{A(t)}\right). \tag{3.5}$$

$\parallel s_1 - s_2 \parallel$는 $s_1$과 $s_2$ 지점 사이의 거리이다. $A(t) = \exp(\alpha_0 + \alpha_1 C_t + \alpha_2 S_t)$이다. 일반적으로 공간적인 랜덤효과를 보기 위해서 $\mu(\cdot)$, $\sigma(\cdot)$와 $\xi(\cdot)$에 관한 추론에 초점을 맞추어진다.

본 연구에서 적용되는 모형은 다음과 같은 계층적 베이지안 모형으로 정리할 수 있다.

- Data model: $[z(s)|\mu(s), \sigma(s), \xi(s)]$ for $s = s_1, \ldots, s_n$

$$z(s) \overset{\text{ind}}{\sim} \text{GEV}\left(\mu(s), \sigma(s), \xi(s)\right) \quad \text{or} \quad z(s) \overset{\text{ind}}{\sim} \text{GPD}\left(\mu(s), \sigma(s), \xi(s)\right).$$

- Process model
  - $[\mu(s_1), \ldots, \mu(s_n)|\mathbf{m}_\mu, \mathbf{\Sigma}(\boldsymbol{\theta}_\mu)] : (\mu(s_1), \ldots, \mu(s_n))' \sim \text{MVN}(\mathbf{m}_\mu, \mathbf{\Sigma}(\boldsymbol{\theta}_\mu))$.
  - $[\sigma(s_1), \ldots, \sigma(s_n)|\mathbf{m}_\sigma, \mathbf{\Psi}(\boldsymbol{\theta}_\sigma)] : (\log\sigma(s_1), \ldots, \log\sigma(s_n))' \sim \text{MVN}(\mathbf{m}_\sigma, \mathbf{\Psi}(\boldsymbol{\theta}_\sigma))$.
  - $[\xi(s_1), \ldots, \xi(s_n)|\mathbf{m}_\xi, \mathbf{\Omega}(\boldsymbol{\theta}_\xi)] : (\xi(s_1), \ldots, \xi(s_n))' \sim \text{MVN}(\mathbf{m}_\xi, \mathbf{\Omega}(\boldsymbol{\theta}_\xi))$.
- Prior model: $[\mathbf{m}_\mu], [\mathbf{m}_\sigma], [\mathbf{m}_\xi], [\boldsymbol{\theta}_\mu], [\boldsymbol{\theta}_\sigma], [\boldsymbol{\theta}_\xi]$

일반적으로 $\mathbf{m}_\mu$, $\mathbf{m}_\sigma$ 와 $\mathbf{m}_\xi$에 대해서는 서로 독립인 무정보 사전분포를 사용될 수 있고, 위치모수 $\mu$, 척도모수 $\sigma$ 그리고 형상모수 $\xi$에 대한 공분산함수의 모수인 $[\boldsymbol{\theta}_\mu], [\boldsymbol{\theta}_\sigma], [\boldsymbol{\theta}_\xi]$에 대해서는 균일분포를 사용할 수 있다. 단, 여기에서는 분석의 용이를 위해서 초모수 값 $\mathbf{m}_\mu$, $\mathbf{m}_\sigma$, $\mathbf{m}_\xi$와 공분산함수의 분산 부분을 고정하였다. 즉 본 연구에서는 초모수에 대한 보조변수를 이용한 모형은 고려하지 않았다.

## 4. 전국 62개 지역의 강우량 자료에 기초한 모형 적합

이 논문에서는 기상청(KMA)로부터 얻은 1973년 1월 1일부터 2011년 12월 31일까지로 39년동안 우리나라 76개 지역에서 측정한 일별 강우량 관측자료를 사용하였다. 그 중에서 100개 이상의 결측치가 포함된 지역 14군데를 제외한 나머지 62개 지역의 자료를 사용하였다. 여기서는 복잡한 데이터구조와

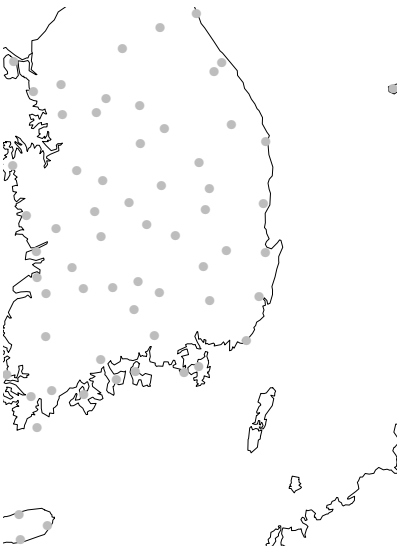

**Figure 4.1.** Locations of 62 observing stations

계산을 고려하여 1년을 365일로 적용하였고 윤달의 경우 2월 29일을 제외하였다. 관측지점은 Figure 4.1와 같이 주어진다.

베이지안 모수추정에서는 관측값 $\mathbf{z}$이 주어졌을 때, 관심모수 $\theta$의 사후분포를 통해서 이루어 지는데, 이는 베이즈 정리에 의해서 구해질 수 있다. 즉,

$$\pi(\theta|\mathbf{z}) \propto p(\mathbf{z}|\theta)\pi(\theta), \tag{4.1}$$

여기에서 $p(\mathbf{z}|\theta)$는 모형에 대한 확률분포함수이고 $\pi(\theta)$는 모수에 대한 사전분포이며 이를 이용해서 사후분포인 $\pi(\theta|\mathbf{z})$를 구할 수 있다. 본 논문에서는 보다 복잡한 계층구조의 베이지안 모형을 제안했는데 이 경우에도 유사한 방법을 통해서 사후분포를 구할 수 있다. 예를 들어,

$$\pi(\theta_1, \theta_2|\mathbf{z}) \propto p(\mathbf{z}|\theta_1)\pi(\theta_1|\theta_2)\pi(\theta_2), \tag{4.2}$$

여기에서 $\pi(\theta_1|\theta_2)$는 관측값의 모형에 직접 관련이 있는 모수에 관한 사전분포이며 이 분포는 모수 $\theta_2$에 의해서 결정된다. 그리고 모수 $\theta_2$에 대한 사전분포 $\pi(\theta_2)$가 정의된다. 예를 들어, 극단값 분포인 GEV 분포함수나 GPD 분포함수는 $p(\mathbf{z}|\theta_1)$에 해당하며 $\theta_1$는 극단값 분포의 위치모수 $\mu(s)$, 척도모수 $\sigma(s)$ 그리고 형상모수 $\xi(s)$가 된다. 식 (3.1)–식 (3.3)과 같은 극단값 분포의 모수들에 대한 공간모형은 $\pi(\theta_1|\theta_2)$단계에 해당한다. 마지막으로 $\pi(\theta_2)$는 공간모형에서 사용된 평균이나 공분산함수에 관련된 모수에 관한 사전분포가 된다.

본 논문에서 제안한 모형의 관심모수에 대한 결합사후확률분포를 구하는 것은 어려우므로 각 변수의 조건부사후확률분포로부터 랜덤표본을 반복적으로 생성하면 적절한 조건하에서 이들의 극한분포가 결합사후확률밀도함수가 된다는 사실에 근거하여 표본 추출이 용이한 깁스 표집기를 형성한다. 깁스 표집기를 가동하기 위한 각 확률 변수의 조건부 사후밀도함수를 구할 수 있는데 깁스 표집기를 위한 대부분의 조건부 분포가 잘 알려진 분포족에 속하지 않는 경우가 많아서 정규화 상수의 계산에도 무리가 따른다. 이 경우 깁스 표집기 내부에 정규화 상수 없이도 목표함수를 따르는 표본 추출이 가능한 메트로폴

리스-해스팅스 알고리즘(Metropolis-Hastings algorithm)을 삽입하여 모수의 추정을 시도한다. 이러한 방법에 기초하여 여러 관심모수에 대한 사후분포를 추정할 수 있다.

전국 62개 지역의 강우량 자료를 위한 계층적 베이지안 모형에 대한 결합사후분포는 다음과 같다.

$$\begin{aligned}\pi\left(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\xi}|z(s_1),\ldots,z(s_n)\right) \propto & \prod_{i=1}^{n}\left[1+\xi(s_i)\left(\frac{z(s_i)-\mu(s_i)}{\sigma(s_i)}\right)\right]^{-\frac{1}{\xi(s_i)}} \\ & \times |\boldsymbol{\Sigma}(\boldsymbol{\theta}_\mu)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_\mu)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_\mu)(\boldsymbol{\theta})(\boldsymbol{\mu}-\mathbf{m}_\mu)\right) \\ & \times |\boldsymbol{\delta}\boldsymbol{\Psi}\left(\boldsymbol{\theta}_\sigma\right)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\delta}-\mathbf{m}_\sigma)'\boldsymbol{\Psi}^{-1}(\boldsymbol{\theta}_\sigma)(\boldsymbol{\delta}-\mathbf{m}_\sigma)\right) \\ & \times |\boldsymbol{\Omega}(\boldsymbol{\theta}_\xi)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\xi}-\mathbf{m}_\xi)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}_\xi)(\boldsymbol{\xi}-\mathbf{m}_\xi)\right).\end{aligned} \tag{4.3}$$

여기에서 $\boldsymbol{\delta} = (\log(\sigma(s_1),\ldots,\log(\sigma(s_n))'$이다. 참고로 GPD분포의 경우에도 동일한 방법으로 구해질 수 있다. 위의 결합사후분포에 기초한 각 모수의 조건부사후분포는 다음과 같이 구해진다.

- $\boldsymbol{\mu}\Big|\text{rest} \propto \prod_{i=1}^{n}\left[1+\xi(s_i)\left(\frac{z(s_i)-\mu(s_i)}{\sigma(s_i)}\right)\right]^{-\frac{1}{\xi(s_i)}}\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_\mu)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_\mu)(\boldsymbol{\theta})(\boldsymbol{\mu}-\mathbf{m}_\mu)\right)$.
- $\boldsymbol{\sigma}\Big|\text{rest} \propto \prod_{i=1}^{n}\left[1+\xi(s_i)\left(\frac{z(s_i)-\mu(s_i)}{\sigma(s_i)}\right)\right]^{-\frac{1}{\xi(s_i)}}|\boldsymbol{\delta}\boldsymbol{\Psi}(\boldsymbol{\theta}_\sigma)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\delta}-\mathbf{m}_\sigma)'\boldsymbol{\Psi}^{-1}(\boldsymbol{\theta}_\sigma)(\boldsymbol{\delta}-\mathbf{m}_\sigma)\right)$.
- $\boldsymbol{\mu}\Big|\text{rest} \propto \prod_{i=1}^{n}\left[1+\xi(s_i)\left(\frac{z(s_i)-\mu(s_i)}{\sigma(s_i)}\right)\right]^{-\frac{1}{\xi(s_i)}}\exp\left(-\frac{1}{2}(\boldsymbol{\xi}-\mathbf{m}_\xi)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}_\xi)(\boldsymbol{\xi}-\mathbf{m}_\xi)\right)$.
- $\boldsymbol{\theta}_\mu|\text{rest} \propto |\boldsymbol{\Sigma}(\boldsymbol{\theta}_\mu)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_\mu)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_\mu)(\boldsymbol{\theta})(\boldsymbol{\mu}-\mathbf{m}_\mu)\right)$.
- $\boldsymbol{\theta}_\sigma|\text{rest} \propto |\boldsymbol{\delta}\boldsymbol{\Psi}(\boldsymbol{\theta}_\sigma)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\delta}-\mathbf{m}_\sigma)'\boldsymbol{\Psi}^{-1}(\boldsymbol{\theta}_\sigma)(\boldsymbol{\delta}-\mathbf{m}_\sigma)\right)$.
- $\boldsymbol{\theta}_\xi|\text{rest} \propto |\boldsymbol{\Omega}(\boldsymbol{\theta}_\xi)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\xi}-\mathbf{m}_\xi)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}_\xi)(\boldsymbol{\xi}-\mathbf{m}_\xi)\right)$.

여기에서 모든 관심모수의 조건부사후분포가 일반적으로 잘 알려진 분포족에 속하지 않음을 알 수 있다. 메트로폴리스-해스팅스 알고리즘을 위한 일반적인 후보 생성 밀도 함수의 선택은 $q(y \mid x) = q_1(\mid y-x \mid)$로 주어지는 확률 보행(random walk) 연쇄이며, 본 연구의 경우 목표함수가 가지는 경험적 밀도함수와 유사한 형태의 절단정규분포를 후보생성 밀도함수로 설정한다.

$$q_1(y \mid x) = \frac{\phi\left(y \mid x,\sigma^2\right)}{1-\Phi\left(0 \mid x,\sigma^2\right)}, \tag{4.4}$$

여기서 $\phi(y \mid x,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-x)^2\right\}$이며 $\Phi(0 \mid x,\sigma^2) = \int_{\infty}^{0}\phi(y \mid x,\sigma^2)\,dy$이다. 목표 함수가 가지는 확률변수의 범위는 0보다 크거나 같은 구간임을 감안하여 절단 정규 분포의 하한 값을 0으로 결정한다. 이때 분산 $\sigma^2$을 후보 생성 밀도 함수의 조율 모수(tuning parameter)라 정의 하며, 조율 모수의 조정을 통해 적절한 채택-기각률을 가지는 효율적인 메트로폴리스-해스팅스 알고리즘을 형성한다.

표본 추출을 위한 깁스 표집기의 효율성과 체인의 수렴도를 확인하기 위해 병렬적 구조를 갖춘 시뮬레이션 체인을 형성 하였으며, 메트로폴리스-해스팅스 알고리즘의 기각률과 겔만-루빈통계량(Gelman-Rubin statistics 또는 G-R statistics)을 각각 확인하였다. 관심모수는 매 시간단위에서 독립이며, 각

**Table 4.1.** Gelman-Rubin statistics and rejection rate in Markov chain Monte Carlo simulation.

| 모수 | $\mu$'s | $\sigma$'s | $\xi$'s | $\theta_\mu$ | $\theta_\sigma$ | $\theta_\xi$ |
|---|---|---|---|---|---|---|
| G-R통계량 | $1.12 \sim 1.24$ | $1.19 \sim 1.28$ | $1.26 \sim 1.38$ | 1.19 | 1.27 | 1.36 |
| 기각률(%) | $32.34 \sim 39.21$ | $39.12 \sim 44.32$ | $42.34 \sim 49.21$ | 40.25 | 43.04 | 44.93 |

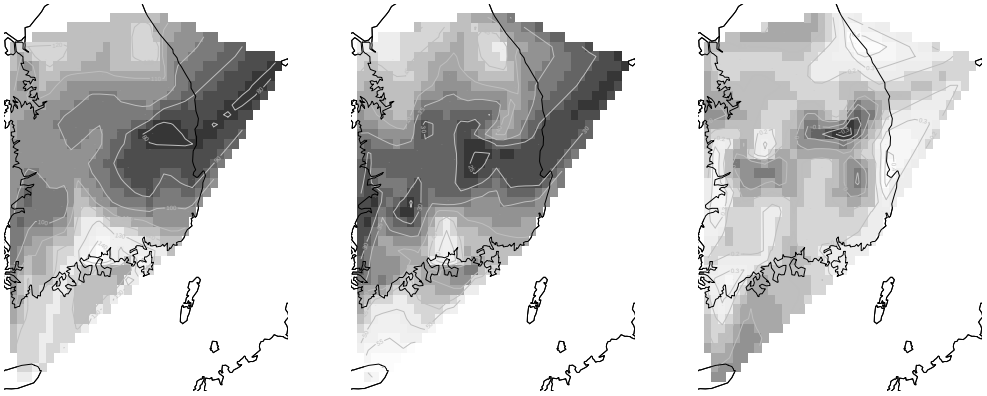

**Figure 4.2.** Plots of location (left), scale (middle) and shape (right) parameters of GEV distribution fitted based on observed annual maximum daily precipitations.

단위시간별 관심모수를 1차원적 모수로 간주하여 메트로폴리스-해스팅스 알고리즘을 통한 추출이 가능하다. 본 논문의 메트로폴리스-해스팅스 알고리즘이 가지는 기각률은 전체적으로 30-50%로 나타나는 것으로 확인할 수 있다 (Table 1 참조). 본 연구에서는 3개의 병렬 체인을 이용하여 체인의 수렴 여부를 확인하였고 50,000개의 표본을 생성한 후, 초기값의 영향을 제거하기 위하여 처음의 25,000개의 표본을 burn-in 과정으로 제거하였다.

Figure 4.2에서는 우리나라 62개 지역에서 39년간 측정된 일별 강우량 관측자료를 기초한 연간 최고 강우량을 적용한 GEV 분포의 위치모수 $\mu(s)$, 척도모수 $\sigma(s)$ 그리고 형상모수 $\xi(s)$의 사후평균값들이다. 사후평균과 함께 추정오차도 함께 구해질 수 있는데, 추정오차는 모든 계수가 비슷한 공간적 특성을 보임에 반해서 추정 평균값들은 서로 다른 공간적 특성을 보임을 알 수 있다. Figure 4.3에서는 동일한 기간동안의 일별 강우량 관측자료를 기초한 임계값 초과치 자료를 적용한 GPD 분포의 척도모수 $\sigma(s)$ 그리고 형상모수 $\xi(s)$의 사후평균값들이다. 이렇게 얻어진 공간정보를 이용하여 반환주기에 대한 공간모형을 구성한다. 더 나아가 관측자료가 없는 지역에 대한 반환주기를 추정할 수 있다. 즉, 특정지역의 반환주기는 위치모수 $\mu(s)$, 척도모수 $\sigma(s)$ 그리고 형상모수 $\xi(s)$의 공간구조를 이용하여 그 지점의 모수값들을 추정하고 이를 이용해서 그 지점의 반환주기를 추정한다.

이제 추정된 극단값 분포의 모수들에 대한 사후분포를 이용하여 반환주기에 대한 사후분포를 알아보았다. 정확한 관심모수의 사후분포가 알려지지 않았으므로 반환주기의 사후분포도 몬테카를로 방법을 이용하여 구할 수 있다. 즉, 관심모수의 사후분포로부터 추출된 메트로폴리스 해이스팅 알고리즘을 통한 표본을 이용하여 반환주기를 구성하고 이렇게 얻어진 반환주기값들을 이용하여 반환주기에 대한 사후분포를 얻을 수 있다. Figure 4.4와 Figure 4.5에서는 이렇게 얻어진 GEV 분포와 GPD 분포에 기초한 반환주기의 사후분포의 평균과 표준편차의 공간적 분포를 보여준다. 추정된 모형에 기반하여 대구 및 경북지역에 대한 반환주기 값이 다른 지역에 비하여 상대적으로 높음을 GPD와 GEV 모두에서 확인할 수 있었고 이들 반환주기 사이에 유의한 공간적 관련성을 확인할 수 있다. 또한 반환주기 값이 높은 지역에

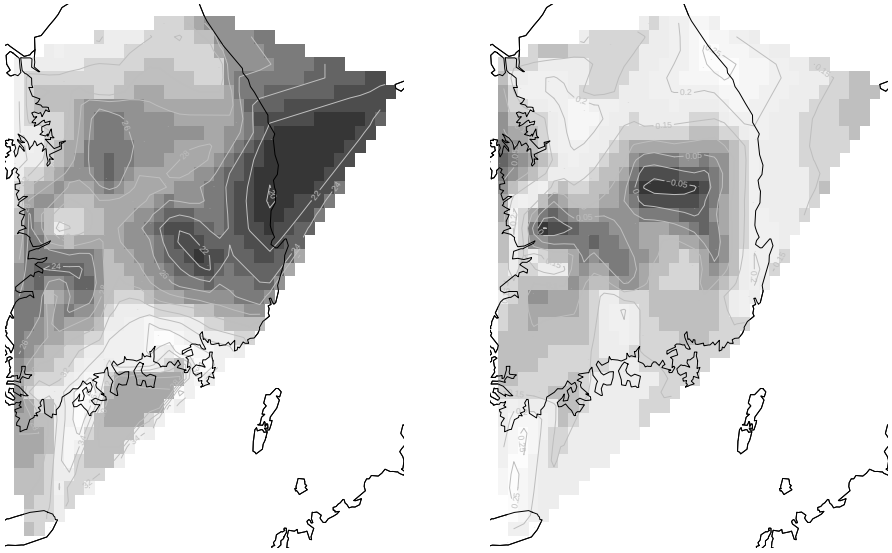

**Figure 4.3.** Plots of scale(left) and shape (right) parameters of GPD distribution fitted based on observed threshold exceedances.

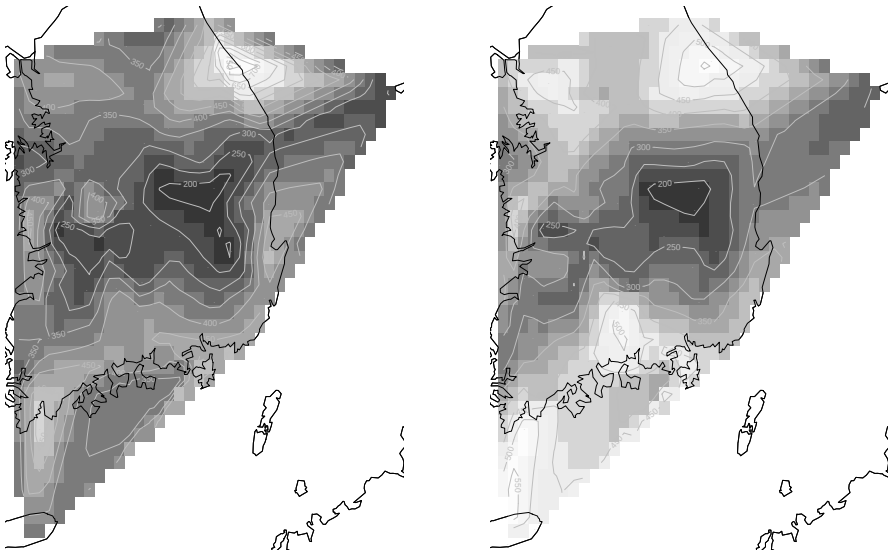

**Figure 4.4.** Posterior means of return leves based on GEV (left) and GPD (right) distribution.

서 불확실성 또한 상대적으로 높음을 알 수 있다. GEV 분포와 GPD 분포에 기초해 얻어진 반환주기는 서로 다른 공간적 특성를 보여주지만 각각의 평균값과 표준편차는 서로 비슷한 공간적 특성을 공유함을 알 수 있다. 또한 관측지점이 밀집된 지역의 추정값에 대한 표준편차가 관측지점이 성긴 지역에 비해서 상대적으로 작음을 확인할 수 있었다.

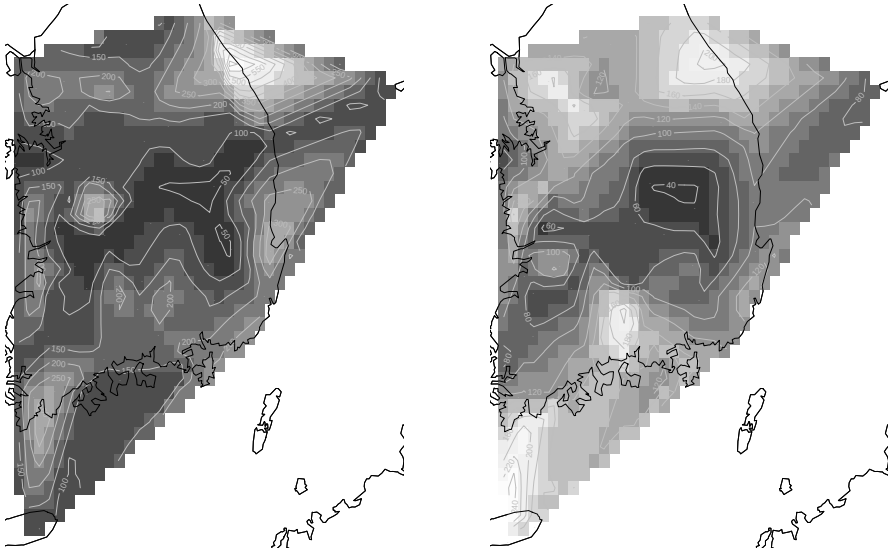

**Figure 4.5.** Posterior standard deviation of return leves based on GEV (left) and GPD (right) distribution.

## 5. 결론

본 논문에서 두가지 극단값 분포와 이를 이용한 공간구조를 가지는 반환주기에 대한 계층적 베이지안 모형을 소개하였고, 우리나라 62개 지역에서 39년간 측정된 일별 강우량 관측자료를 제안된 모형에 적용하였다. 효과적으로 관심모수값인 반환주기의 공간모형을 구성하기 위해서 계층적 구조를 이용하여 각 극단값 분포의 모수들에 공간구조를 가지는 사전분포를 가정하였다. 이를 통해서 간접적으로 반환주기에 공간적 관련성을 설명하였다. 사후평균의 측면에서는 두 가지 극단값 분포가 서로 유사한 경향을 보여주었지만 사후분포의 오차 측면에서는 서로 다른 경향을 보여줌을 알 수 있었다. 이에 대한 추가적인 연구가 필요할 것으로 보인다. 일반적으로 실제 모수의 변화에 따른 극치 강수량의 경우 형상모수보다 척도모수의 영향을 많이 받지만 추론의 측면에서는 형상모수에 더 민감하게 반응하는 것을 알 수 있다. 그리고 모형의 적합성과 효율성을 알아보기 위해서 교차타당성을 평가한 결과, 관측지점이 밀집되어 있는 지점들에서는 상당히 효과적인 예측값을 제공하지만 주변에 관측값이 없는 지점에 대해서는 예측의 효과가 상대적으로 떨어짐을 알 수 있다. 추가적으로 본 논문에서는 여러지점의 극단값 모수의 사전평균값을 동일하게 가정하였지만, 각 지점에 관한 정보를 모형에 활용하는 것도 예측의 편차를 줄이는데 도움을 줄 수 있을 것으로 기대된다.

## References

Casson, E. and Coles, S. (1999). Spatial Regression Models for Extremes, *Extremes*, **1**, 449–468.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*, London: Springer-Verlag.

Coles, S., Heffernan, J. and Tawn, J. (1999). Dependence measures for extreme value analysis, *Extremes*, **2**, 339–365.

Cooley, D., Naveau, P., Jomelli, V, Rabatel, A. and Grancher, D. (2006). A Bayesian hierarchical extreme value model for lichenometry, *Environmetrics*, **17**, 555–574.

Cooley, D., Nychka, D. and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels, *Journal of the American Statistical Association*, **102**, 824–840.

Cunderlik, J. M., Burn, D. H., Posbherg, D., Robinson, B. A. and Zyvoloski, G. A. (2008). Generalized likelihood uncertainty estimation(GLUE) using adaptive Markov Chain, *Journal of Hydrology*, **276**, 210–223.

de Haan, L. (1985). *Extremes in Higher Dimensions: The Model and Some Statistics*, in proceedings of the $45^{\text{th}}$ session of the International Statistical Institute.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Berlin: Springer-Verlag.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values, *Journal of the Royal Statistical Society, Series B*, **66**, 497–546.

Lee, J. J. (2010). Assessment of nonstationarity in precipitation and development of nonstationary frequency analysis, Ph.D. Thesis, Chonbuk National University.

Jang, S. W., Seo, L., Kim, T. W. and Ahn, J. H. (2011). Non-stationary rainfall frequency analysis based on residual analysis, *Journal of the Korean Society of Civil Engineers*, **31**, 449–457.

Kwak, M. and Kim, Y. (2014). Multi-site stochastic weather generator for daily rainfall in Korea, *The Korean Journal of Applied Statistics*, **27**, 475–485.

Matérn, B. (1986). *Spatial Variation*, $2^{\text{nd}}$ edition, Springer-Verlag.

Oliveria, O. A., Gomes, M. I. and Alves, M. I. F.(2006). Improvements in the estimation of a heavy tail, *REVSTAT - Statistical Journal*, **4**, 81–109.

Schlather, M. and Tawn, J. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference, *Biometrika*, **90**, 139–156.

Strupczewski, W. G., Singh, V. P. and Mitosek, H. T. (2001). Non-stationary approach to at-site flood frequency modeling III : Flood analysis of Polish rivers, *Journal of Hydrology*, **248**, 152–167.

# 한국지역 집중호우에 대한 반환주기의 베이지안 모형 분석

이정진[a] · 김남희[a] · 권혜지[b] · 김용구[a,1]

[a]경북대학교 통계학과, [b]통계청

(2014년 7월 29일 접수, 2014년 8월 18일 수정, 2014년 8월 29일 채택)

**요 약**

집중호우의 특성을 이해하는 것은 수문관리 및 재해방재 등에서 매우 중요하다. 특히 반환주기는 이러한 집중호우의 특성을 나타내는 측정치로 자주 사용된다. 본 논문에서는 베이지안 계층적 모형을 이용하여 강우의 반환주기에 대한 공간구조를 분석하였다. 먼저 국내 62개 지점에서 측정한 강우 강도을 기초로 하여 연간 일일 최대강우량과 특정한 수준을 초과하는 강우량에 대해서 generalized extreme value(GEV)와 generalized Pareto distribution(GPD)를 각각 가정하여 추정하였다. 집중호우 반환주기에 대한 공간구조는 이 GEV 분포와 GPD 분포의 모수에 공간구조를 가지는 다변량 정규분포를 이용하여 설명하였다. 제안된 모형을 국내 76개 지역에서 39년간 측정된 일별 강우량 관측자료에 적용하였다.

주요용어: 베이지안 분석, 강우모형, 극단값, generalized extreme value 분포, generalized Pareto 분포, 반환주기, 공간모형.

[1]교신저자: (702-701) 대구광역시 북구 대학로 80, 경북대학교 통계학과. E-mail: kim.1252@knu.ac.kr