한국환경과학회지 제20권(제3호), 329~340, 2011 DOI: 10.5322/JES.2011.20.3.329
Journal of the Environmental Sciences

Bayesian MCMC 및 Metropolis Hastings 알고리즘을 이용한 강우빈도분석에서 확률분포의 매개변수에 대한 불확실성 해석

서영민 · 박기범[1)]
영남대학교 토목공학과, [1)]동양대학교 철도토목과
(2010년 9월 27일 접수; 2010년 11월 23일 수정; 2011년 1월 13일 채택)

Uncertainty Analysis for Parameters of Probability Distribution in Rainfall Frequency Analysis by Bayesian MCMC and Metropolis Hastings Algorithm

Young Min Seo, Ki Bum Park*
Department of Civil Engineering, Yeungnam University Kyeungsan 712-749, Korea
[1)] Department of Railroad and Civil Engineering, Dongyang University, Kyeongbuk 750-711, Korea
(Manuscript received 27 September, 2010; revised 23 November, 2010; accepted 13 January, 2011)

Abstract

The probability concepts mainly used for rainfall or flood frequency analysis in water resources planning are the frequentist viewpoint that defines the probability as the limit of relative frequency, and the unknown parameters in probability model are considered as fixed constant numbers. Thus the probability is objective and the parameters have fixed values so that it is very difficult to specify probabilistically the uncertianty of these parameters.

This study constructs the uncertainty evaluation model using Bayesian MCMC and Metropolis -Hastings algorithm for the uncertainty quantification of parameters of probability distribution in rainfall frequency analysis, and then from the application of Bayesian MCMC and Metropolis- Hastings algorithm, the statistical properties and uncertainty intervals of parameters of probability distribution can be quantified in the estimation of probability rainfall so that the basis for the framework configuration can be provided that can specify the uncertainty and risk in flood risk assessment and decision-making process.

Key Words : Rainfall frequency analysis, Probability distribution, Parameter, Uncertainty, Bayesian MCMC, Metropolis-Hastings

1. 서 론

수자원 계획에 있어서 강우 또는 홍수빈도 분석시 가장 흔히 사용되는 통계학적 기법은 빈도학파 기법(frequentist method) 또는 전통적인 통계학적 기법으로 분류되며, 이러한 기법들은 미지의 매개변수들이 고정된 상수이고 제한된 상대빈도를 이용하여 확률을 정의할 수 있다고 가정한다. 이러한 가정들로 인해 확률은 객관적이며, 매개변수들이 고정된 값을 가지기 때문에 매개변수들에 대한 확률론적 설명은 매우 어렵다. 이에 비해 베이지안 기법(Bayesian

*Corresponding author : Ki Bum Park, Department of Railroad and Civil Engineering, Dongyang University, Kyeongbuk 750-711, Korea
Phone: +82-53-321-1517
E-mail: pkb5032@naver.com

method)은 대안적 접근방법을 제공한다. 베이지안 기법은 매개변수를 확률변수로 처리하며, 확률을 믿음의 정도(degrees of belief)로 정의한다. 즉, 어떤 사상에 대한 확률은 그 사상이 참이라고 믿는 정도를 나타내며, 이러한 점에서 빈도학파와는 다른 관점을 가진다.

어떤 확률밀도 $p(\boldsymbol{y}|\theta)$에 의해 설명되는 통계학적 모델을 이용하여 자료 $\boldsymbol{y}=\{y_1,\cdots,y_n\}$로부터 매개변수 벡터 θ를 추정하는데 관심이 있다고 할 경우 베이지안 관점에서 볼 때 θ는 정확하게 결정할 수 없으며, 매개변수에 대한 불확실성은 확률분포를 통하여 표현될 수 있다. 이러한 관점으로부터 확률분포형의 매개변수에 대한 불확실성을 추정하기 위해서 베이지안 추론(Bayesian inference)을 적용할 수 있으며, 베이지안 추론은 크게 3가지 단계로 구성된다. 먼저 매개변수 θ에 대한 확률분포를 $\pi(\theta)$로 설정하며, 이것은 사전분포(prior distribution)라고 알려져 있다. 사전분포는 자료를 적용하기 이전에 매개변수에 대한 분석자의 믿음, 즉 평균, 분산, 왜곡도 등을 나타낸다. 다음으로 관측자료 $\boldsymbol{y}$와 매개변수 벡터 θ가 주어졌을 때 $\boldsymbol{y}$의 분포를 나타내기 위하여 어떤 통계학적 함수 $p(\boldsymbol{y}|\theta)$를 선택한다. 마지막으로 사전분포로부터의 자료를 결합하고 사후분포(posterior distribution) $p(\theta|\boldsymbol{y})$를 계산함으로써 매개변수 벡터 θ에 대한 결과를 나타낸다.

따라서 본 연구에서는 강우빈도분석에서 확률분포형의 매개변수에 대한 불확실성을 해석하기 위하여 베이지안 해석을 적용하였으며, 이를 통해 현재 홍수위험관리 실무에서 홍수위험평가를 위한 모델선정 및 실행, 매개변수의 선택 또는 홍수량 추정치와 관련된 불확실성의 고려없이 통상 확정론적으로 모델링되고 있고 또한 공적토론, 정책결정 및 의사결정은 보통 이러한 확정론적 모델링 결과를 근거로 이루어지고 있는 실정에 있어서 홍수위험평가 및 의사결정시 위험도와 불확실성을 충분히 설명할 수 있는 방안을 제시하였다.

2. 베이지안 해석

2.1. 베이즈 정리

두 개의 사상 A와 B에 대하여 각각의 확률 $P(A)$와 $P(B)$가 0이 아니라면 사상 B가 주어졌을 때 사상 A에 대한 조건부 확률(conditional probability) $P(A|B)$의 정의는 식 (1)과 같이 나타낼 수 있다.

$$P(A|B)=P(A\cap B)/P(B) \tag{1}$$

여기서, $P(A\cap B)$는 A와 B가 모두 발생하는 사상에 대한 확률을 나타내며, $P(A|B)$는 사상 B가 이미 발생했다는 사실이 주어졌을 때 사상 A가 발생할 가능성(chance)을 나타낸다.

만약 각각의 확률이 $P(A)\neq 0$, $P(B)\neq 0$이면

$$P(A|B)P(B)=P(A\cap B)=P(B|A)P(A) \tag{2}$$

이고 식 (2)를 정리하면 식 (3)과 같은 베이즈 정리를 얻을 수 있다.

$$P(A|B)=P(A)P(B|A)/P(B) \tag{3}$$

베이즈 정리는 $P(A)$를 곱하고 $P(B)$를 나눔으로써 간편하게 $P(B|A)$에서부터 $P(A|B)$로 변환할 수 있으며, 이러한 형식화에서 $P(A)$를 사전분포(prior distribution), $P(B|A)$를 우도함수(likelihood), $P(B)$를 정규화 인자(normalization factor), $P(A|B)$를 사후분포(posterior distribution)라고 한다. 대부분의 베이지안 해석에서 사전분포는 전문가들의 의견 또는 믿음으로부터 유도되며, 어떤 특정 증거가 고려되기 전에 분석자의 주관적 지식을 나타내기 위해 사용된다. 또한 사전분포는 애매한 선개념(amorphous preconception), 역학적 추론(mechanistic reasoning), 전문(hearsay) 또는 이러한 것들을 조합한 결과일 수 있다. 한편, 우도함수는 문제의 변수에 대하여 자료가 의미하는 것에 대한 어떤 모델을 나타내며, 또한 그것은 전문가의 주관적 지식으로부터 얻어질 수도 있다. 정규화 인자 $P(B)$는 종종 Eq. (4)와 같은 총확률법칙(law of total probability)을 이용하여 계산될 수 있으나 정규화 인자의 계산은 종종 베이즈 정리의

적용에 있어서 계산적 부담이 가장 큰 측면이다. 이러한 정규화 인자를 해석적으로 계산하기는 어려우나 공액쌍(conjugate pair)을 이용하여 문제를 상당히 단순화시킬 수 있다. 또한 이러한 계산기법을 적용할 수 없을 경우 수치적 기법을 적용하여 문제를 해결할 수도 있다.

$$P(B) = \sum_i P(B|C_i)P(C_i) \tag{4}$$

한편, 베이즈 정리는 어떤 이산표본공간(discrete sample space)에 대한 분할(partition)에 적용하여 일반적인 형태로 나타낼 수 있다. $\{A_i\}_{i=1}^n$을 표본공간에 대한 분할이라고 할 때, 여기서 n은 무한할 수 있으며, B를 어떤 집합이라고 하면, 각 i에 대하여 식 (5)와 같이 일반적인 형태의 베이즈 정리를 나타낼 수 있다.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)} \tag{5}$$

사상에 대한 베이즈 정리는 확률변수 및 분포함수에 대한 베이즈 정리를 정의하기 위하여 확장될 수 있다. 확장된 베이즈 정리는 사전분포와 우도함수를 결합하여 사후분포를 생성하는데 사용될 수 있으며, 사후분포는 위험도 해석에서 입력으로 사용될 수도 있다. 여기서, 각각의 확률사상을 연속확률밀도함수로 나타내면 베이즈 정리는 식 (6)과 같이 나타낼 수 있다.

$$\pi(\theta|x_1,x_2,\cdots,x_n) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)\pi(\theta)}{\int_\theta f(x_1|\theta)\cdots f(x_n|\theta)\pi(\theta)d\theta} \tag{6}$$

여기서, $\pi(\theta|x_1,x_2,\cdots,x_n)$은 사후분포, $\pi(\theta)$는 사전분포이며, 우변의 분모는 상수로서 주변분포이고 우변 분자의 $f(x_1|\theta)\cdots f(x_n|\theta)$는 발생할 수 있는 모든 가능성을 고려한 우도함수이다.

또한 연속표본공간에 대하여 어떤 이변량 확률벡터 (X, θ)를 고려할 때, 여기서 X는 표본벡터, θ는 매개변수 벡터를 나타내며, 이러한 이변량 확률벡터가 결합확률밀도함수 $\pi(X, \theta)$를 가진다고 하면, X에 대한 주변확률분포(marginal distribution) $f_X(x)$와 θ에 대한 주변확률분포 $\pi_\Theta(\theta)$는 다음과 같이 각각 나타낼 수 있다.

$$f_X(x) = \int \pi(x, \theta)d\theta \tag{7}$$

$$\pi_\Theta(\theta) = \int \pi(x, \theta)dx \tag{8}$$

여기서, $\pi_\Theta(\theta)$를 θ에 대한 사전확률밀도함수(prior density function)라 하기도 한다.

2.2. 베이지안 마코프 연쇄 몬테카를로 기법

사후분포에 대한 모의기법 중 가장 널리 사용되는 기법은 마코프 연쇄 몬테카를로 기법(Markov Chain Monte Carlo, MCMC)이다. 이것은 이산 마코프 연쇄의 연속값에 대한 일반화로서 MCMC 샘플링은 정상분포(stationary distribution)가 사후분포와 같은 기약적(irreducible), 비주기적(aperiodic) 마코프 연쇄를 구축한다.

MCMC 기법은 정규화 상수의 계산없이 사후분포로부터 매개변수의 값을 샘플링하는 기법으로서 널리 사용되는 알고리즘으로는 Gibbs sampler(Geman, 1984; Casella 와 George, 1992)와 Metropolis-Hastings 알고리즘(Metropolis 등., 1953; Hastings, 1970)이 있다. Gibbs sampler는 Metropolis-Hastings 알고리즘 보다 실행하기 쉽고 빠르며 튜닝이 필요하지 않지만 각 변수에 대한 조건부 분포가 알려져 있어야 하며, 이것은 항상 가능한 경우가 아니기 때문에 그 대신에 Metropolis-Hastings 알고리즘이 종종 사용된다. 따라서 마코프 연쇄를 구축하는 일반적인 방법은 Metropolis-Hastings 알고리즘을 사용하는 것이며, Metropolis-Hastings 알고리즘에 대한 두 가지 특별한 변형으로서 독립연쇄(independence chain)와 확률보행연쇄(random walk chain)가 널리 사용되고 있다.

2.2.1. 마코프 연쇄(Markov Chains)

사후분포 $p(\theta|y)$로부터 근사적으로 샘플링하기 위해 MCMC 기법은 마코프 연쇄에 의해 종속적 표본을 발생시킨다. $\theta^{(0)}$, $\theta^{(1)}$, $\cdots$를 확률변수에 대한 시퀀스(sequence)라고 할 때, 만약 식 (9)와 같은 관계가 성립한다면, $p(\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(T)})$를 마코프 연쇄라고 한다.

$$p(\theta^{(t)}|\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(t-1)}) = p(\theta^{(t)}|\theta^{(t-1)}) \tag{9}$$

식 (9)로부터 마코프 연쇄는 오직 바로 이전의 상태만이 현재의 상태와 관련됨을 확인할 수 있다.

$\theta^{(t)}$가 이산상태공간(discrete state space) $S=\{s_1, s_2, \cdots\}$ 상에 정의된다고 할 때 마코프 연쇄에서는 식 (10)과 같은 t단계 전이행렬(t-step transition matrix) $Q_{i,j}(t)$를 가지는 식 (11)과 같은 일정한 1단계 전이행렬 $Q_{i,j}$에 의해 $p(\theta^{(t)}|\theta^{(t-1)})$가 정의된다고 가정된다.

$$Q_{i,j} = \Pr(\theta^{(t)} = s_j \mid \theta^{(t-1)} = s_i) \tag{10}$$

$$Q_{i,j}(t) = \Pr(\theta^{(t)} = s_j \mid \theta^{(0)} = s_i) \tag{11}$$

일정한 1단계 마코프 연쇄로부터의 샘플링은 만약 연쇄에 대한 추가조건, 즉 기약성(irreducibility), 비주기성(aperiodicity) 및 positive recurrence가 만족된다면 요구되는 정상분포(stationary distribution), 즉 $\pi(\theta) = p(\theta|y)$에 수렴한다(Roberts, 1996; Norris, 1997). 이러한 조건들을 만족시키는 연쇄에 대한 샘플링은 식 (13)과 같은 완전균형조건(full balance condition)을 만족시키는 Eq. (12)와 같은 유일한 정상분포를 가지게 된다.

$$\lim_{t\to\infty} Q_{i,j}(t) = \pi_{(j)} \tag{12}$$

$$\pi_{(j)} = \sum_i \pi_{(i)} Q_{i,j} \tag{13}$$

2.2.2. Metropolis-Hastings 알고리즘

Metropolis-Hastings 알고리즘은 Metropolis 등.(1953)에 의해 처음으로 소개되었으며, 이후 Hastings (1970)에 의해 일반화되었다. 이 알고리즘은 제안분포(proposal distribution) $q(.|\boldsymbol{X}_t)$에 기초한 마코프 연쇄의 구축을 가능케 하며, 또한 이 알고리즘은 목표분포(target distribution) $\pi(\cdot)$의 밀도에 대한 평가 가능성만을 요구하는 기각 샘플링 알고리즘(rejection sampling algorithm)이다. 이 알고리즘은 제안분포로부터 한 표본을 추출하며, 수용확률(acceptance probability)에 따라 그 표본을 수용한다. 만약 표본이 기각되면 그 연쇄는 상태 $t+1$에서 변하지 않은 채로 남아있게 된다. 제안분포는 $\pi(\cdot)$와 유사해야 하지만 현재 표본 $\boldsymbol{X}_t$에 의존할 수 있으며, 이러한 제안분포의 선택으로서 일반적으로 다변량 정규분포(multivariate normal distribution)가 많이 추천되고 있다.

Metropolis-Hastings 알고리즘은 제안분포(proposal distribution) $q(.|\boldsymbol{X}_t)$로부터 후보값(candidate point) $\boldsymbol{Y}$를 샘플링함으로써 $t+1$에서의 연쇄의 상태를 얻을 수 있다. 이것은 오직 이전 상태 $\boldsymbol{X}_t$에만 의존하며, 정칙조건(regularity condition)에 종속되는 어떤 형태를 가질 수 있다(Roberts, 1996). 제안분포 $q(.|\boldsymbol{X}_t)$의 예로서 평균 $\boldsymbol{X}_t$와 고정된 공분산 행렬을 가지는 다변량 정규분포가 있으며, 제안분포 $q(.|\boldsymbol{X}_t)$를 선택할 때 그 제안분포를 이용하여 쉽게 샘플링할 수 있어야 한다는 사실에 주의해야 한다. 제안분포 $q(.|\boldsymbol{X}_t)$에 대하여 요구되는 정칙조건은 기약성(irreducibility)과 비주기성(aperiodicity)이며(Chib 와 Greenberg, 1995), 기약성은 마코프 연쇄가 모든 시작점들로부터 어떤 공집합이 아닌 집합에 이를 수 있는 양의 확률이 존재함을 의미하고 비주기성은 마코프 연쇄가 다른 상태 집합 사이에서 진동하지 않음을 의미한다. 만약 제안분포가 목표분포와 동일한 서포트(support)에서 양의 밀도를 가진다면 일반적으로 이러한 조건들은 만족될 수 있으며, 또한 목표분포가 어떤 제한된 서포트를 가질 경우 이러한 조건들은 만족될 수 있다. 한편, 연쇄의 다음 상태가 다음과 같은 확률을 가진다면 후보점은 수용되며, 만약 점 $\boldsymbol{Y}$가 수용되지 않는다면, 그 연쇄는 이동하지 않으며, $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$가 된다.

$$\alpha(\boldsymbol{X}_t,\ Y) = \min\left\{1,\ \frac{\pi(\boldsymbol{Y})q(\boldsymbol{X}_t|\boldsymbol{Y})}{\pi(\boldsymbol{X}_t)q(\boldsymbol{Y}|\boldsymbol{X}_t)}\right\} \quad (14)$$

이러한 Metropolis-Hastings 알고리즘은 다음과 같은 단계에 따라 실행될 수 있다(Martinez 등., 2002).

- **Step 1**. 마코프 연쇄에 대한 초기값 $\boldsymbol{X}_0$ 설정하고 $t=0$로 설정
- **Step 2**. 제안분포 $q(.|\boldsymbol{X}_t)$로부터 후보값 $\boldsymbol{Y}$를 발생
- **Step 3**. 균등분포 $Uniform(0,\ 1)$로부터 U를 발생
- **Step 4**. 만약 $U \le \alpha(\boldsymbol{X}_t,\ \boldsymbol{Y})$이면 $\boldsymbol{X}_{t+1} = \boldsymbol{Y}$이고, 그렇지 않으면 $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$
- **Step 5**. $t = t+1$로 설정하고 Step 2~5를 반복

3. 결과 및 고찰

3.1. 적용유역 및 분석개요

본 연구에서는 강우빈도해석에서 확률분포의 매개변수에 대한 불확실성을 정량화하기 위하여 베이지안 MCMC 및 Metropolis-Hastings 알고리즘을 이용한 불확실성 평가모델을 구축하였다. 그리고 베이지안 MCMC 및 Metropolis-Hastings 알고리즘의 적용을 통하여 확률강우량 산정시 확률분포의 매개변수에 대한 통계학적 특성 및 불확실성 구간을 정량화하였다.

강우빈도해석시 확률분포의 매개변수에 대한 불확실성 평가를 위한 베이지안 MCMC의 적용을 위하여 분석유역으로서 위천유역을 선정하였으며, 위천 유역 내 의성 관측소(기상청)를 대상으로 확률밀도함수 및 우도함수를 구축하고 의성 관측소 인근 18개 관측소(Fig. 1)로부터 확률분포의 매개변수에 대한 정보적 사전분포를 구축하였다. 다음으로 이러한 우도함수 및 사전분포로부터 Metropolis-Hastings 알고리즘을 적용하였으며, 이로부터 확률분포의 매개변수에 대한 불확실성 구간을 추정하였다.

강우빈도분석은 수자원 분야에서 일반적으로 널리 적용되고 있는 FARD를 이용하였으며, 베이지안 MCMC 및 Metropolis-Hastings 알고리즘은 통계계산 및 그래픽을 목적으로 GNU 프로젝트로 개발된 프로그래밍 언어인 R을 이용하여 코딩하였다. 본 연구의 대상유역 및 강우관측소에 대한 현황은 Fig. 1과 같다.

Fig. 1. Study Area and Location of Raingauges.

3.2. 확률밀도함수의 선정 및 우도함수

베이지안 MCMC 기법을 이용하여 강우빈도분석에서 확률분포의 매개변수에 대한 불확실성을 평가하기 위해서는 먼저 적합한 확률분포형을 선정해야 한다. 본 연구에서는 위천 유역 내 의성 관측소을 대상으로 하였으며, 의성 관측소의 1973~2008년 동안의 시우량 자료를 수집하여 먼저 이상치 및 결측치 등을 보정하고 지속시간별 최대강우량을 산정하였다. 확률분포의 매개변수 추정기법으로는 L-moment법을 적용하였으며, 적합도 검정을 위해 Chi-Square, Kolmogorov-Smirnov, Cramer Von Mises 및 PPCC 검정을 적용하였다. 이로부터 적정 확률분포형으로 General Logistic 분포를 채택하였으며, 이 분포에 대한 확률분포함수 및 확률밀도함수는 각각 식 (15)와 식 (16)과 같다.

$$F(x) = \left[1 + \left\{1 - k\left(\frac{x-\varepsilon}{\alpha}\right)\right\}^{1/k}\right]^{-1} \quad (15)$$

$$f(x) = \frac{1}{\alpha}\left[1-k\left(\frac{x-\varepsilon}{\alpha}\right)\right]^{\left(\frac{1}{k}-1\right)}\left[1+\left\{1-k\left(\frac{x-\varepsilon}{\alpha}\right)\right\}^{1/k}\right]^{-2} \quad (16)$$

Table 1. Results of Goodness-of-fit Test

Duration(hr)		1	2	3	4	6	9	12	15	18	24	48	72
χ^2	Com.	2.54	3.57	1.51	0.49	1.51	5.97	3.23	4.60	1.17	1.51	2.54	2.54
	Tab.	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
KS	Com.	0.09	0.12	0.08	0.06	0.05	0.08	0.08	0.10	0.07	0.10	0.06	0.08
	Tab.	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22
CVM	Com.	0.05	0.08	0.04	0.02	0.01	0.06	0.03	0.07	0.04	0.04	0.03	0.03
	Tab.	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
PPCC	Com.	0.99	0.98	0.99	0.99	0.99	0.97	0.98	0.99	0.99	0.98	0.99	0.99
	Tab.	0.94	0.95	0.94	0.93	0.94	0.93	0.93	0.92	0.92	0.93	0.93	0.94

※ χ^2 : Chi-Square test, KS : Kolmogorov-Smirnov test, CVM : Cramer Von Mises test, PPCC : Probability-Plot Correlation Coefficient, Com.: Computed value, Tab.: Table value.

여기서, ε는 위치매개변수(location parameter), α는 규모매개변수(scale parameter)이고 k는 형상매개변수(shape parameter)이다. 그리고 General Logistic 분포에 대한 적합도 검정결과는 Table 1과 같다.

한편, 베이지안 MCMC 기법을 이용하여 강우빈도분석에서 확률분포형에 대한 매개변수의 불확실성을 평가하기 위해서는 앞서 선정한 확률분포형인 General Logistic 분포에 대한 우도함수가 필요하며, General Logistic 분포에 대한 우도함수를 나타내면 식 (17)과 같다(Rao 와 Hamed, 2000).

$$L(x \mid \alpha, \varepsilon, k) = \frac{1}{\alpha^N}\prod_{i=1}^{N}\left[1-k\left(\frac{x_i-\varepsilon}{\alpha}\right)\right]^{1/k-1}\prod_{i=1}^{N}\left[1+\left\{1-k\left(\frac{x_i-\varepsilon}{\alpha}\right)\right\}^{1/k}\right]^{-2} \quad (17)$$

여기서, L은 우도함수, α는 규모매개변수, ε은 위치매개변수, k는 형상매개변수, N은 자료수이다.

3.3. 사전분포의 선정

적절한 사전분포의 구축은 베이지안 해석에서 가장 논쟁의 여지가 많은 부분이다. 사전분포는 크게 정보적 사전분포(informative prior)와 무정보적 사전분포(non-informative prior)로 구분할 수 있다. 사전분포를 갱신하는데 있어서 많은 양의 자료가 가용하고 분석자의 사전적 믿음이 상대적으로 불확실할 경우 사후분포에 대한 자료의 영향은 사전분포로부터의 어떤 영향을 압도하는 경향이 있다. 따라서 정보적 사전분포를 개발하는데 있어서 많은 시간과 노력을 소비하는 것이 유용하지 않을 수도 있으며, 이러한 상황에서는 불확실한 지식의 상태를 나타내기 위해 수학적으로 구축되는 무정보적 사전분포가 매우 유용하다. 그러나 자료가 거의 없을 경우에는 이러한 무정보적 사전분포의 사용은 정당화되기 더욱 어렵다는 점에 주의해야 한다. 만약 추정문제가 위험도에 대한 중대한 기여를 포함한다면 그것은 결과에 중요한 영향을 미칠 수 있기 때문에 분석자는 어떤 정보적 사전분포를 구축하기 위해 노력해야 한다. 확률론적 위험도 해석은 사전분포의 구축에 있어서 반복적 절차를 가지는 경향이 있기 때문에 무정보적 사전분포로부터 시작할 수 있으며, 그 다음으로 위험도에 상당한 기여를 하는 매개변수에 대하여 정보적 사전분포가 구축될 수 있다. 한편, 사전분포를 구축하기 위한 이러한 반복적인 절차에서 가능한한 갱신시 사용되는 자료와는 독립이어야 한다는 점에 주의해야 한다.

본 연구에서는 앞서 선정된 Generalized Logistic 분포의 매개변수들에 대한 사전분포를 구축하기 위하여 정보적 사전분포를 적용하였다. Generalized Logistic 분포의 각 매개변수들에 대한 사전분포를 구축하기 위해 본 연구의 대상 관측소인 의성 지점 인근의 18개 관측소를 대상으로 시우량을 수집한 후 지속시간별 최대강우량을 산정하고 Generalized Logistic 분포에 적합시켰으며, 각 매개변수별로 각 관측소에 대한 매개변수값을 다시 확률분포에 적합시켜 각 매개변수, 즉 위치, 규모 및 형상매개변수에 대한 확률분

Table 2. Summary Statistics for Location Parameter

Duration (hr)	Min.	Max.	Mean	Mode	Med.	Std.	Skew	Kur.	Percentile			
									10%	25%	75%	90%
1	23.312	37.792	29.901	30.580	30.233	3.971	0.147	3.017	23.312	27.389	31.956	36.735
2	34.701	54.315	42.135	44.638	41.630	4.711	0.741	4.417	36.013	39.794	44.548	48.051
3	41.326	57.366	49.631	48.537	49.193	4.677	0.015	2.249	43.536	45.450	52.830	57.292
4	46.129	69.087	56.495	55.760	56.413	5.597	0.229	3.317	49.416	51.529	60.575	63.431
6	53.668	81.157	67.573	68.574	68.153	6.829	0.031	2.984	59.319	62.803	71.197	76.263
9	63.573	93.416	79.342	76.958	78.675	8.247	0.003	2.180	70.337	72.225	87.715	89.666
12	69.382	105.432	88.225	81.123	86.262	10.002	0.183	2.301	77.081	80.899	96.292	102.869
15	74.222	114.253	95.077	92.032	92.886	10.144	0.067	2.781	83.573	89.079	106.094	107.246
18	82.077	118.964	101.043	97.641	98.291	9.714	0.153	2.472	90.250	94.588	110.355	113.704
24	90.337	129.570	110.700	106.111	106.874	10.809	0.141	2.201	98.146	105.086	120.478	126.463
48	114.924	160.214	132.238	134.762	131.757	13.792	0.406	2.034	115.939	118.790	145.421	148.755
72	123.690	172.677	144.165	125.036	143.915	14.867	0.332	2.065	125.522	133.421	159.261	165.938

※ Min.: Minimum, Max.: Maximum, Med.: Median, Std.: Standard deviation, Skew: Skewness, Kur.: Kurtosis.

Table 3. Summary Statistics for Scale Parameter

Duration (hr)	Min.	Max.	Mean	Mode	Med.	Std.	Skew	Kur.	Percentile			
									10%	25%	75%	90%
1	3.681	11.112	7.648	8.706	7.999	1.747	-0.434	3.557	5.350	6.286	8.689	9.466
2	5.979	13.251	9.338	7.539	8.877	2.198	0.261	1.909	6.289	7.533	11.308	12.468
3	6.886	15.019	10.591	12.702	10.006	2.425	0.353	2.201	6.961	8.888	12.815	14.507
4	7.376	16.750	11.810	10.455	10.803	2.770	0.355	2.020	8.298	9.791	14.004	15.873
6	9.833	20.782	13.996	16.378	12.899	3.267	0.529	2.172	10.327	11.403	16.430	18.548
9	8.689	24.919	16.552	13.910	16.453	3.731	0.212	3.686	13.095	13.974	18.761	21.502
12	10.681	27.634	18.987	19.708	19.041	4.109	0.263	3.417	14.948	15.999	20.963	25.897
15	12.920	31.207	21.093	21.249	21.035	4.772	0.603	3.238	15.611	17.649	22.530	29.778
18	14.791	33.863	23.331	23.360	22.920	5.117	0.577	3.296	16.090	20.971	23.952	32.986
24	15.850	39.101	27.106	27.114	26.568	5.926	0.517	3.343	20.768	23.605	27.666	37.436
48	21.350	50.844	33.533	29.573	32.852	6.907	0.713	4.285	25.547	29.647	37.364	43.658
72	19.309	60.662	36.349	37.857	36.790	9.050	0.813	5.363	25.729	32.881	38.784	48.280

Table 4. Summary Statistics for Shape Parameter

Duration (hr)	Min.	Max.	Mean	Mode	Med.	Std.	Skew	Kur.	Percentile			
									10%	25%	75%	90%
1	-0.387	0.236	-0.141	-0.199	-0.174	0.153	0.947	3.967	-0.293	-0.239	-0.087	0.098
2	-0.373	0.154	-0.120	-0.366	-0.120	0.150	-0.134	2.786	-0.363	-0.167	-0.035	0.110
3	-0.480	0.188	-0.159	-0.188	-0.177	0.171	0.037	2.605	-0.362	-0.333	-0.004	0.014
4	-0.543	0.186	-0.170	-0.155	-0.173	0.169	-0.104	3.821	-0.415	-0.304	-0.069	0.034
6	-0.532	0.137	-0.170	-0.215	-0.198	0.159	-0.142	3.892	-0.400	-0.235	-0.070	0.090
9	-0.580	0.135	-0.185	-0.286	-0.194	0.189	-0.324	2.936	-0.502	-0.288	-0.038	0.063
12	-0.595	0.154	-0.188	-0.311	-0.202	0.204	-0.205	2.647	-0.514	-0.305	-0.049	0.091
15	-0.591	0.108	-0.203	-0.142	-0.162	0.198	-0.298	2.481	-0.531	-0.353	-0.063	0.072
18	-0.589	0.075	-0.221	-0.374	-0.198	0.186	-0.320	2.294	-0.497	-0.374	-0.065	0.009
24	-0.596	0.028	-0.235	-0.158	-0.175	0.190	-0.481	1.995	-0.520	-0.415	-0.067	-0.018
48	-0.566	0.221	-0.240	-0.130	-0.164	0.254	0.102	1.665	-0.564	-0.520	-0.064	0.073
72	-0.639	0.239	-0.249	-0.631	-0.158	0.300	0.017	1.484	-0.629	-0.521	-0.053	0.144

Table 5. Distributions and Estimates for GLO Parameters

Duration	Location Para. ε (Normal Dist.)		Scale Para. α (Logistic Dist.)		Shape Para. k (Triangular Dist.)		
(hr)	Mean, μ	Std., σ	m	a*	Min., a	Most Likely, c	Max., b
1	29.901	3.971	7.751	0.955	-0.425	-0.239	0.294
2	42.135	4.711	9.247	1.292	-0.459	-0.117	0.214
3	49.631	4.677	10.453	1.400	-0.544	-0.179	0.251
4	56.495	5.597	11.647	1.615	-0.602	-0.167	0.246
6	67.573	6.829	13.785	1.914	-0.579	-0.210	0.213
9	79.342	8.247	16.458	2.048	-0.661	-0.116	0.204
12	88.225	10.002	18.862	2.250	-0.685	-0.112	0.236
15	95.077	10.144	20.763	2.600	-0.724	0.108	0.108
18	101.043	9.714	22.977	2.754	-0.711	0.075	0.075
24	110.700	10.809	26.664	3.173	-0.711	0.028	0.028
48	132.238	13.792	33.138	3.664	-0.566	-0.566	0.359
72	144.165	14.867	35.907	4.605	-0.639	-0.639	0.436

포를 구축하였다. 그 결과 각 매개변수에 대한 통계학적 특성은 Tables 2~4와 같고 적합된 확률분포형은 Table 5와 같다.

따라서 본 연구에서 적용된 사전분포는 위치, 규모 및 형상매개변수에 대하여 각각 정규분포, Logistic 분포 및 삼각형 분포이며, 각 매개변수의 사전분포를 나타내면 식 (18)~(20)과 같다.

$$\pi(\varepsilon|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(\varepsilon-\mu)^2/(2\sigma^2)} \tag{18}$$

$$\pi(\alpha|a^*,m)=\frac{1}{a^*}e^{\left(\frac{\alpha-m}{a^*}\right)}\left[1+e^{\left(\frac{\alpha-m}{a^*}\right)}\right]^{-2} \tag{19}$$

$$\pi(k\,|\,a,\,b,\,c)=\begin{cases}\dfrac{2(k-a)}{(b-a)(c-a)} & \text{for } a\le k<c\\ \dfrac{2(b-k)}{(b-a)(b-c)} & \text{for } c\le k\le b\\ 0 & otherwise\end{cases} \tag{20}$$

그리고 독립성을 가정하여 세 매개변수에 대한 사전분포는 식 (21)과 같이 나타낼 수 있다.

$$\pi(\varepsilon,\alpha,k)=\pi(\varepsilon|\mu,\sigma)\pi(\alpha|a^*,m)\pi(k|a,b,c) \tag{21}$$

따라서 식 (17)과 식 (21)로부터 본 연구에서 구축한 사후분포는 식 (22)의 형태로 나타낼 수 있다.

$$\pi(\varepsilon,\alpha,k|x)=\frac{L(x|\varepsilon,\alpha,k)\pi(\varepsilon,\alpha,k)}{\iiint L(x|\varepsilon,\alpha,k)\pi(\varepsilon,\alpha,k)d\varepsilon d\alpha dk} \tag{22}$$

3.4. Metropolis-Hasting 알고리즘 적용 및 고찰

MCMC 기법은 정규화 상수의 계산없이 사후분포로부터 매개변수의 값을 샘플링하는 기법으로서 널리 사용되는 알고리즘으로는 Gibbs sampler와 Metropolis-Hastings 알고리즘이 있다. 여기서 Metropolis-Hastings 알고리즘의 경우 제안분포가 정상분포 $\pi(\cdot)$와 유사해야 하지만 현재표본 X_t에 의존할 수 있으며, 이러한 제안분포의 선택으로서 일반적으로 식 (23)과 같은 다변량 정규분포가 많이 추천되고 있다.

$$\begin{aligned}f(\vec{x})&=f(x_1,\cdots,x_p)\\&=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})'\Sigma^{-1}(\vec{x}-\vec{\mu})\right]\end{aligned} \tag{23}$$

여기서, $\vec{x}=[x_1, x_2, \cdots, x_p]$이고 $\vec{x}\sim N_p(\vec{\mu}, \Sigma)$이며, $\vec{\mu}$는 평균벡터(mean vector), Σ는 공분산행렬(covariance matrix)이다. 본 연구에서는 전절에서 구

축한 사후분포로부터 매개변수의 값을 샘플링하기 위해 Metropolis-Hastings 알고리즘을 적용하였으며, 여기에 필요한 제안분포로서 삼변량 정규분포(trivariate normal distribution)를 사용하였다.

한편, 번인(burn-in)은 사후분포 추정에 있어서 초기값의 영향을 최소화시키기 위해 마코프 연쇄의 초기부분값들을 버리는 것으로서 본 연구에서는 최소번인을 결정하기 위하여 도식적인 방법을 이용하였으며, 초기값의 영향을 최소화하기 위해 시작값으로 각 매개변수에 대한 평균값을 사용하였다. 이러한 평균값의 사용은 연쇄의 시작을 정상분포의 서포트 집합 내에서 시작할 수 있기 때문에 초기값의 영향이 감소되는데 필요한 단계를 대폭 감소시킴으로 해서 번인 크기를 크게 감소시킬 수 있으나 Gilks 등.(1998)에 의하면 심지어 연쇄가 즉각적으로 수렴하는 것으로 나타나더라도 그 연쇄에 있어서 시작값의 영향이 감소될 만큼 충분히 긴 번인을 실행하는 것이 필수적인 것으로 연구된 바 본 연구에서는 101,000회의 모의를 실시하여 이중 1,000회를 번인으로 하고 100,000개의 표본을 분석에 사용하였으며, 각 매개변수에 대하여 발생된 100,000개의 표본의 경우 사후분포를 추정하기 위한 마코프 연쇄의 양호한 혼합상태를 확인할 수 있으며, 이에 대한 trace plot을 나타내면 Fig. 2～4와 같다.

Table 6 및 Fig. 5～7은 베이지안 MCMC 모의결과에 대한 분포적합결과를 정리한 것으로서 Table 6은 GLO 분포의 각 매개변수에 대한 지속시간별 모의결과로부터 적합된 분포와 그 분포의 매개변수를 정리한 것이고 Fig. 5～7은 Table 6의 결과중 지속시간 6, 12 및 24시간에 대한 적합결과를 그래프로 나타낸 것이다. GLO 분포의 위치매개변수 및 규모매개변수는 정규분포로 적합되었으며, 형상매개변수는 3변수 Weibull 분포로 적합되었다. 예를 들어, 지속시간 6시간의 경우 GLO 분포의 위치매개변수는 평균이 67,61, 표준편차가 0.696인 정규분포, 규모매개변수의 경우 평균이 13.52, 표준편차가 0.669인 정규분포

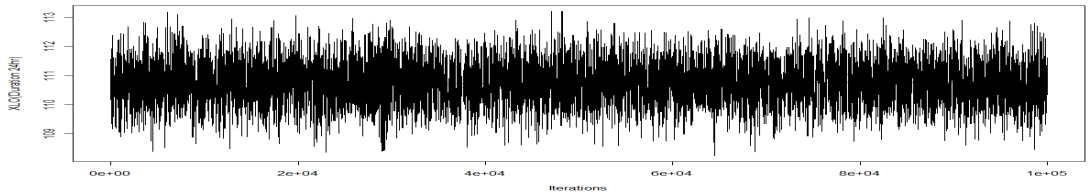


Fig. 2. Trace Plot for Location Parameter. (Duration 24 hr)

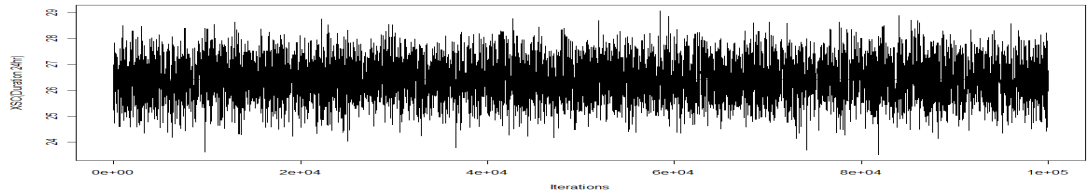


Fig. 3. Trace Plot for Scale Parameter. (Duration 24 hr)

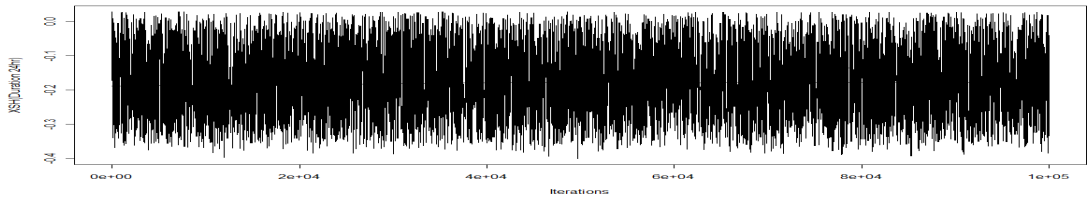


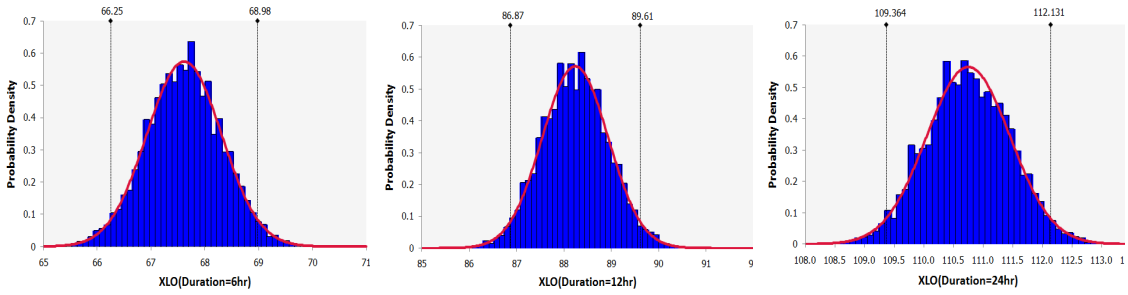
Fig. 4. Trace Plot for Shape Parameter. (Duration 24 hr)

Table 6. Distribution Fitting for Bayesian MCMC Simulation

Parameters		XLO		XSC		XSH		
Distribution Types & Parameters		Normal Distribution		Normal Distribution		Weibull Distribution		
		μ	σ	μ	σ	ζ	β	δ
D(hr)	1	30.01	0.638	6.91	0.585	-0.413	0.215	1.898
	2	42.21	0.684	8.87	0.609	-0.504	0.410	3.355
	3	49.71	0.688	10.19	0.627	-0.524	0.354	2.886
	4	56.51	0.681	11.41	0.643	-0.562	0.346	3.130
	6	67.61	0.696	13.52	0.669	-0.439	0.254	2.433
	9	79.38	0.695	16.17	0.678	-0.397	0.219	2.238
	12	88.24	0.699	18.63	0.676	-0.402	0.228	2.260
	15	95.10	0.693	20.51	0.682	-0.378	0.213	2.020
	18	101.07	0.696	22.73	0.682	-0.389	0.221	2.183
	24	110.75	0.708	26.41	0.698	-0.441	0.294	2.890
	48	132.31	0.700	32.93	0.691	-0.513	0.372	2.717
	72	144.25	0.710	35.73	0.696	-0.618	0.496	3.598

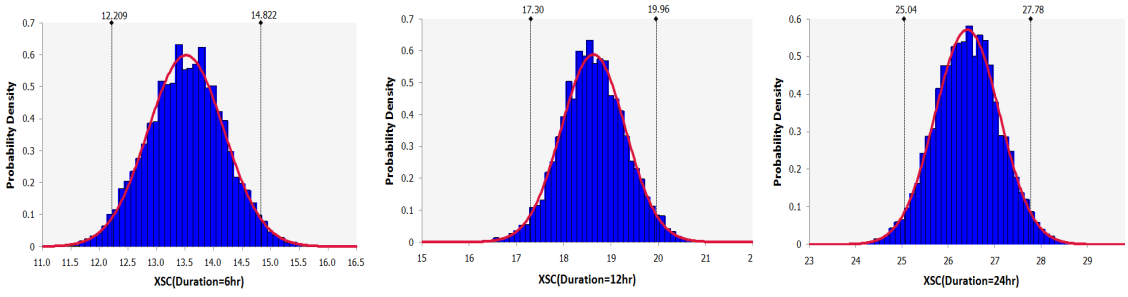
※ μ: mean, σ: standard deviation, ζ: location parameter of Weibull distribution, β: scale parameter of Weibull distribution, δ: shape parameter of Weibull distribution

로 확률적으로 설명될 수 있으며, GLO 분포의 형상매개변수의 경우 위치매개변수가 -0.439, 규모매개변수가 0.254, 형상매개변수가 2.433인 3변수 Weibull 분포로 확률적으로 설명될 수 있다.



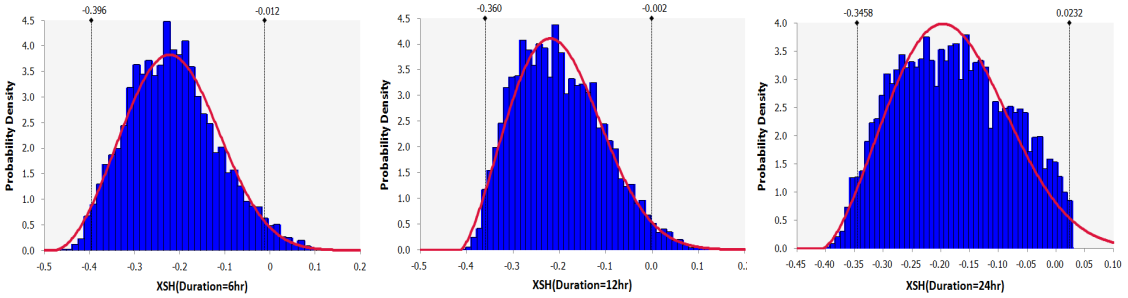
(a) Durarion 6 hr (b) Duration 12 hr (c) Duration 24 hr

Fig. 5. Distribution Fitting of Location Parameter



(a) Durarion 6 hr (b) Duration 12 hr (c) Duration 24 hr

Fig. 6. Distribution Fitting of Scale Parameter.



(a) Durarion 6 hr (b) Duration 12 hr (c) Duration 24 hr

Fig. 7. Distribution Fitting of Shape Parameter.

한편, 본 연구에서는 GLO 분포의 각 매개변수에 대한 불확실성을 정량화하기 위하여 HPD (highest posterior density interval) 구간을 추정하였으며, 그 결과는 Table 7과 같다. Table 7로부터 지속시간 6시간의 경우 위치매개변수, 규모매개변수 및 형상매개변수에 대한 불확실성 구간이 각각 66.285~69.015, 12.197~14.790, -0.4008~-0.0223로 추정되었으며, 지속시간 12시간의 경우 각각 86.884~89.599, 17.312~20.003, -0.3661~-0.0241로 추정되었다.

4. 결 론

본 연구에서는 강우빈도분석에서 확률분포의 매개변수에 대한 불확실성의 정량화를 통해 확률강우량의 산정에 불확실한 범위를 제시하여 홍수위험평가 및 의사결정과정에서 불확실성 및 위험도를 설명할 수 있는 프레임워크 구성을 위한 기초를 마련하고자 베이지안 해석을 적용하였다. 강우빈도해석에서 확률분포의 매개변수에 대한 불확실성을 해석 및 정량화하기 위한 방법으로 베이지안 MCMC 및 Metropolis-Hastings 알고리즘을 이용한 불확실성 평가모델을 구축하였다.

본 연구에서 베이지안 MCMC 및 Metropolis-Hastings 알고리즘을 확률분포함수의 매개변수 추정의 적용에서 적합도 검정등을 통하여 GLO 분포가 가장 적합한 분포형으로 분석되었다.

GLO 분포의 매개변수의 불확실성을 평가한 결과 위치매개변수(XLO), 규모매개변수(XSC) 그리고 형상매개변수(XSH)의 불확실성 구간을 산정하였으며, 확률강우량 산정시 확률분포의 매개변수에 대한 통계학적 특성 및 불확실성 구간을 정량화하여 제시할 수 있었다.

Table 7. HPD Interval (GLO distribution)

duration	XLO		XSC		XSH	
	lower	upper	lower	upper	lower	upper
1	28.830	31.289	5.804	8.075	-0.4096	-0.0199
2	40.903	43.558	7.631	10.049	-0.3703	0.0919
3	48.368	51.024	8.997	11.430	-0.4380	0.0120
4	55.201	57.827	10.261	12.757	-0.4561	-0.0466
6	66.285	69.015	12.197	14.790	-0.4008	-0.0223
9	77.955	80.669	14.916	17.530	-0.3640	-0.0303
12	86.884	89.599	17.312	20.003	-0.3661	-0.0241
15	93.712	96.418	19.067	21.783	-0.3524	-0.0026
18	99.698	102.427	21.441	24.050	-0.3531	-0.0122
24	109.345	112.062	25.068	27.801	-0.3516	-0.0017
48	130.922	133.660	31.560	34.276	-0.4215	0.0487
72	142.842	145.594	34.362	37.101	-0.4296	0.0739

향후 불확실성을 고려한 확률분포형을 결정하여 확률강우량을 산정함에 있어 확률분포함수의 매개변수 추정의 불확실성을 산정하여 제시함으로써 확률분포함수의 불확실성 범위가 가장 작은 확률분포함수를 채택하여 확률강우량을 산정할 수 기준을 제시할 수 있을 것으로 판단된다.

참 고 문 헌

서영민, 지홍기, 이순탁, 2010, 강우빈도분석에서 확률분포식을 결정하는 과정에서의 매개변수에 대한 불확실성 해석: Bayesian MCMC 및 Metropolis-Hastings 알고리즘을 중심으로, 한국수자원학회 학술발표회 초록집, 293.

Casella, G., George, E. I., 1992, Explaining the Gibbs Sampler, American Statistical Association, 46(3), 167-174.

Chib, S., Greenberg, E., 1995, Understanding the Metropolis-Hastings Algorithm, The American Statistician, 49, 327-345.

Geman, S., Geman, D., 1984, Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721-741.

Gilks, D. J. S., Walter, R., 1998, Markov Chain Monte Carlo in Practice, First CRC Press, 1st ed.

Hastings, W. K. 1970, Monte Carlo Sampling Methods Using Markov Chains and Their Applications, Biometrika, 57, 97-109.

Martinez, W. L., Martinez, A. R., 2002, Computational Statistics Handbook with MATLAB, Chapman & Hall/CRC.

Metropolis, N., Rosenbluth, A. W., Teller, A. H., Teller, E., 1953, Equations of state calculations by fast computing machines, Journal of Chemical Physics, 21, 1087-1092.

Norris, J., 1997, Markov Chains, Cambridge University Press: Cambridge.

Rao, A. R., Hamed, K. H., 2000, Flood Frequency Analysis, CRC Press LLC.

Roberts, G., 1996, Markov Chain Concepts Related to Sampling Algorithms, in: Gilks, W., Richardson, S., Spiegelhalter, D. (eds.), Markov Chain Monte Carlo in Practice, Chapman & Hal, London, 45-59.