韓國水資源學會論文集
第41卷 第3號 · 2008年 3月
pp. 325~340

DOI: 10.3741/JKWRA.2008.41.3.325

Bayesian 다중회귀분석을 이용한 저수량(Low flow) 지역 빈도분석

Regional Low Flow Frequency Analysis Using Bayesian Multiple Regression

김 상 욱* / 이 길 성**

Kim, Sang Ug / Lee, Kil Seong

Abstract

This study employs Bayesian multiple regression analysis using the ordinary least squares method for regional low flow frequency analysis. The parameter estimates using the Bayesian multiple regression analysis were compared to conventional analysis using the *t*-distribution. In these comparisons, the mean values from the *t*-distribution and the Bayesian analysis at each return period are not significantly different. However, the difference between upper and lower limits is remarkably reduced using the Bayesian multiple regression. Therefore, from the point of view of uncertainty analysis, Bayesian multiple regression analysis is more attractive than the conventional method based on a *t*-distribution because the low flow sample size at the site of interest is typically insufficient to perform low flow frequency analysis. Also, we performed low flow prediction, including confidence interval, at two ungauged catchments in the Nakdong River basin using the developed Bayesian multiple regression model. The Bayesian prediction proves effective to infer the low flow characteristic at the ungauged catchment.

keywords : Regional low flow frequency analysis, Uncertainty, Bayesian multiple regression, *t*-distribution, Ungauged catchment

요 지

본 연구는 저수량 지역 빈도분석(regional low flow frequency analysis)을 수행하기 위하여 일반최소자승법(ordinary least squares method)을 이용한 Bayesian 다중회귀분석을 적용하였으며, 불확실성측면에서의 효과를 탐색하기 위하여 Bayesian 다중회귀분석에 의한 추정치와 t 분포를 이용하여 산정한 일반 다중회귀분석의 추정치의 신뢰구간을 비교분석하였다. 각 재현기간별 비교결과를 보면 t 분포를 이용하여 산정된 평균 추정치와 Bayesian 다중회귀분석에 의한 평균 추정치는 크게 다르지 않았다. 그러나 불확실성 측면에서 평가해볼 때 신뢰구간의 상한추정치와 하한추정치의 차이는 Bayesian 다중회귀분석을 사용한 경우가 기존 방법을 사용한 경우보다 훨씬 작은 것으로 나타났으며, 이로부터 저수량(low flow) 지역 빈도분석을 수행하는 경우 Bayesian 다중회귀분석이 일반 회귀분석보다 불확실성을 표현하는데 있어서 우수하다는 결과를 얻을 수 있었다.

* 서울대학교 BK21 안전하고 지속가능한 사회기반건설 사업단 박사 후 연구원
Post-Doctor, Seoul National University BK21 SIR Group, Seoul National University, Seoul, 151-744, Korea
(e-mail: plethor1@snu.ac.kr)

** 서울대학교 공과대학 건설·환경공학부 교수
Professor, Dept. of Civil and Environmental Engineering, Seoul National University, Seoul, 151-744, Korea
(e-mail: kilselee@snu.ac.kr)

또한 낙동강 유역에 2개의 미계측 유역을 선정하고 구축된 Bayesian 다중회귀모형을 적용하여 불확실성을 포함한 미계측 유역에서의 저수량(low flow)을 추정하였으며 이와 같은 방법이 미계측 유역에서의 저수(low flow) 특성을 나타내는 데 있어서 효과적일 수 있음을 입증하였다.

핵심용어 : 저수량(low flow) 지역 빈도분석, 불확실성, Bayesian 다중회귀분석, *t* 분포, 미계측유역

1. 서 론

저수분석(low flow analysis)은 수자원의 관리 및 수공구조물의 설계에 있어서 중요한 요소 중의 하나이며, 저수량(low flow)특성을 나타낼 수 있는 여러 가지 지표들 중에서 빈도분석을 이용한 분석 결과가 주로 사용되어 진다. 국내에서는 355위 유량의 10년 빈도추정치에 해당되는 기준갈수량이 주로 사용되며 미국 등에서는 7일 지속기간 10년 빈도유량(7Q10)을 추정하여 수자원의 관리에 사용하고 있다. 그러나 수자원의 관리 측면에 있어서 위와 같은 빈도유량은 확정적인(deterministic) 추정치보다는 불확실성을 포함한 확률적인(probabilistic) 추정치가 사용될 필요가 있으며, 불확실성을 감안한 빈도유량을 사용함으로써 수자원 관리 측면에서 보다 효율적이고 다양한 관리 기법이 응용될 수 있다.

빈도분석을 수행하기 위해서는 추정결과를 얻고자 하는 지점에서 충분한 길이(약 30년 이상)의 과거 유량자료가 필요하고 충분한 과거 유량자료가 확보된 지점에서의 빈도분석은 점 빈도분석(at-site frequency analysis)을 수행하여 원하는 재현기간에서의 추정결과를 얻을 수 있다. 그러나 대부분의 수자원 신규 계획이나 수공 구조물의 설치 등은 과거 자료가 확보되어 있지 않거나 자료의 길이가 짧은 지점에서 수행되어야 하는 경우가 대부분이므로 이러한 경우에는 인근의 자료가 있는 지점에서의 유량자료를 사용하여 원하는 지점에서의 빈도유량을 추정하는 지역 빈도분석(regional frequency analysis)을 사용해야 한다.

지역 빈도분석은 자료가 충분한 인근 지점의 유량정보를 자료가 부족한 지점에서 사용하는 방법이므로 수문학적 특성이 비슷한 지점 또는 유역(hydrological homogeneous region)을 판별하는 작업이 가장 선행되어진다. 이와 같은 수문학적 동질성의 판별은 지역 빈도분석의 추정방법으로 index flood 방법을 적용하는 경우에는 필수적으로 수행해야 하고, 추정방법으로 회귀분석과 같은 다른 방법을 사용하는 경우에도 추정결과의 정확도 보장 측면에서 사전에 수문학적 동질성에 대한 판별이 수행될 필요가 있다. 수문학적 동질성에 대한 판별은 크게 지형적인 분할, 행정구역에 따른 분할, L-moment 방법에 인한 분할, 군집분석(cluster analysis)에 의한 분할 등에 의해 수행될 수 있으나 최종적인 단계에서는 각 방법에 따라 차이는 있지만 분할하고자 하는 주관적인 요인이 개입될 수 있다. 본 연구에서는 수문학적 동질성의 판별을 위하여 주관적인 요인이 가장 적게 포함되어지기 때문에 최근 들어 가장 많이 사용되고 있는 군집분석을 사용하여 낙동강 유역의 수문학적 동질성을 판별하였다.

수문학적 동질성의 판별이 완료되면 판별된 각 동질유역에 따라 추정방법을 적용한다. 주로 사용되는 추정방법으로는 회귀분석, 통계적 지표를 이용한 그래프를 이용하는 방법, 크리깅(kriging)과 같은 공간 보간(spatial interpolation)적 방법 등이 이용된다. 홍수량을 이용한 지역 빈도분석은 Hosking and Wallis (1997)이 제안한 L-moment를 이용한 Index flood 법이 주로 사용되어지나 Durrans and Tomic (1996)은 index flood 법을 저수량(low flow)에 적용하는 과정에서 나타나는 과대 또는 과소 추정의 문제를 제시한 바 있다. 그러므로 본 연구에서는 미국 등에서 저수량의 빈도분석을 위하여 가장 많이 사용되고 있는 다중회귀분석을 추정방법으로 사용하였다.

선정된 추정방법을 수문학적 동질유역에 각각 적용하면 회귀분석의 경우 구축된 회귀모형의 회귀계수(regression coefficient)를 얻을 수 있으며, 이를 이용하여 원하는 지점에서의 설명변수(explanatory variable)로부터 최종적인 종속변수(dependent variable)에 해당하는 추정치를 산정할 수 있다. 그러나 이와 같이 산정된 회귀계수와 그에 따른 종속변수의 추정치는 불확실성을 나타낼 수 있는 신뢰구간이 함께 산정되어 표현되어야 최종 결과인 빈도유량을 이용하여 확률적 개념의 수자원 관리 및 수공구조물의 설계를 수행할 수 있을 것이다. 이와 같은 불확실성을 표현하기 위한 신뢰구간의 산정은 기존에는 정규분포나 *t* 분포를 이용한 근사식으로부터 회귀계수의 신뢰구간을 표현하였다(Stedinger, 1983; Chowdhury and Stedinger, 1991; Stedinger *et al*., 1993; Ashkar and Quarda, 1998; Whitley and Hromadka II, 1999; Cohn *et al*., 2001). 그러나 최근 들어 Reis and Stedinger (2005)는 이와 같은 근사식을 이

용한 신뢰구간의 추정식은 기존 자료의 특성을 간과함으로써 불확실성을 과대하게 추정하여 현실적으로 사용하기 어려운 문제가 있음을 제시한 바 있다.

이와 같이 신뢰구간의 추정에 있어서 기존 연구의 한계점들은 근사식을 사용하지 않고 신뢰구간을 얻을 수 있는 Bayesian 방법을 이용하여 개선되어 질 수 있다. 이 방법은 Vicens *et al.* (1975)에 의해 처음 수공학 분야에 적용되기 시작했으며 이후 Wood and Rodriguez-Iturbe (1975a, 1975b)에 의해 지역 빈도분석에 사용된 바 있다. 그러나 초기 연구에 있어서 구축된 Bayesian 모형의 복잡한 사후분포의 계산에 있어서 한계점이 있어 특정한 문제에만 적용되고 일반적인 문제에 적용되기에는 무리가 있다는 연구가 진행되면서 그간 관련 연구가 진행되어지지 않다가 최근 들어 복잡한 계산을 위한 하드웨어의 발전과 Bayesian 방법을 적용하기 위한 여러 가지 알고리즘의 개발로 인해 빈도분석 분야, 강우-유출모형의 매개변수 보정 분야, 유량의 확률 예측분야에 다시 활발히 적용되고 있는 실정이다(Krzysztofowicz, 1983a, 1983b; Kelly and Krizysztofowicz, 1994; Coles and Powell, 1996; Madsen and Rosbjerg, 1997; Kuczera and Parent, 1998; Kuczera, 1999; Zhang and Govindaraju, 2000; Thiemann *et al.*, 2001; Wang, 2001; O'Connell*et al.*, 2002; Vrugt *et al.*, 2003; Kingston *et al.*, 2005; Reis *et al.*, 2005; Reis and Stedinger, 2005; Kavetski *et al.*, 2006; Seidou *et al.*, 2006; Lee and Kim, 2007; 김상욱, 2007).

특히 지역 빈도분석과 관련하여 Madsen and Rosberg (1997)는 홍수량을 대상으로 빈도분석을 수행하면서 Bayesian 방법을 적용하여 지역 빈도분석을 수행한 바 있으며, 또한 Reis *et al.* (2005)은 일반화 회귀분석(generalized least square regression, GLS)을 구축하면서 Bayesian 방법을 이용하여 모형의 분산오차에 따른 홍수량 빈도분석 모형의 Bayesian GLS모형을 제시한 바 있다. 그러나 수자원의 관리측면에서 저수량(low flow)의 중요성에도 불구하고 위와 같은 방법은 주로 홍수량만을 대상으로 적용되고 있으므로, 본 연구에서는 근사적 방법을 사용하지 않고 불확실성을 나타낼 수 있는 Bayesian 방법을 저수량을 대상으로 지역 빈도분석을 수행하고 구축된 결과를 이용하여 미계측 유역에서의 저수량을 예측하는 연구를 수행하였다.

2. 다중회귀분석과 신뢰구간의 산정

다중회귀분석(multiple regression analysis)은 회귀모형의 설명변수(또는 독립변수)가 2개 이상인 회귀모형에 대한 분석이며, 이를 행렬을 이용하여 표현하면 다음과 같다. 다음의 다중회귀분석과 관련된 내용은 Chatterjee and Price (1977)의 저서를 간단히 요약하였다.

$$\mathrm{y} = \mathrm{X}\beta + \epsilon \tag{1}$$

여기서, y는 종속변수, X는 설명변수, β는 회귀계수이고 ϵ은 잔차항을 나타낸다.

회귀분석의 최종목적은 종속변수와 설명변수간의 관계를 합리적으로 표현할 수 있는 회귀계수를 추정하는 것으로, 주로 잔차항이 평균 0과 일정한 분산 σ^2을 가지고 이상적이고 균일하게 분포되어 있다는 가정 하에 일반 최소자승법(ordinary least squares, OLS)을 사용하여 회귀계수를 추정하게 된다. 위의 가정 하에 OLS에 의한 평균과 회귀계수의 추정치를 나타내면 다음과 같다.

$$E[\mathrm{y}] = \mathrm{X}\beta \tag{2}$$

$$\hat{\beta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y} \tag{3}$$

그러나 잔차항을 검토한 결과가 등분산적(homoscedastic)으로 분포되어 있지 않고 반등분산적(heteroscedastic)으로 분포되어 있거나 잔차간에 자기상관성(autocorrelation)이 존재하는 경우에는 일반화최소자승법을 사용하여 회귀분석을 수행해야 좋은 결과를 얻을 수 있다. 또한 잔차의 분포는 등분산적이지만 자기상관성이 존재하는 경우에는 OLS와 GLS의 중간단계인 가중최소자승법(weighted least squares, WLS)을 사용하여 회귀분석을 수행할 수 있다.

본 연구에서는 잔차항의 특성이 OLS의 적용에 합리적인지를 확인하기 위하여 White의 테스트를 수행하여 적용성을 판단하였다. White의 테스트는 F 테스트의 일종으로서 귀무가설로 회귀모형의 잔차항들의 분포가 등분산적임을 가정한 후, F 테스트에 의한 검정치와 유의확률을 비교함으로써 귀무가설을 채택할 것인지 기각할 것인지를 판정하는 방법이다. 회귀분석을 위한 회귀모형을 구축하는 과정에 있어서 또 다른 검토 대상은 설명변수의 선택이라고 할 수 있다. 특히 다중회귀모형은 설명변수가 여러 개이므로 이를 선정함에 있어서 많은 주의를 기울여야 할 필요가 있다. 선정된 설명변수들은 각 변수 간 일정 정도의 상관관계가 존재하지만 이 상관관계가 지나치게 높은 경우에는 다중공선성(multicollinearity)이 발생하게 되어 회귀계수를 추정하

는 데 있어서 정확도가 감소하게 되어 최종적인 다중회귀모형의 정확도가 많이 낮아지는 문제가 발생하게 된다. 이와 같은 문제를 포함하여 구축된 회귀모형은 그 안정성을 평가하기 위하여 여러 가지 통계시험을 이용하게 된다.

본 연구에서는 설명변수 간 다중공선성의 문제를 근본적으로 피하기 위하여 각 변수 간 상관계수를 산정하여 상관계수가 지나치게 높은 경우(0.9 이상)에는 선정된 상관계수를 삭제하거나, 다른 물리량을 이용하여 변환하여 설명변수를 선정하였다. 또한 다중공선성의 검증을 위하여 분산팽창계수(variation inflation factor)를 산정하여 모형의 안정성을 확인하였으며, 이외에도 결정계수와 조정 결정계수(coefficient of determination, R^2 and adjusted R^2), 자기상관성의 확인을 위한 Durbin-Watson 통계치, leverage 통계치, Cook의 거리, 잔차플롯을 이용하여 구축된 다중회귀모형의 안정성을 검증하였다. 안정성이 검토된 다중회귀분석모형과 OLS를 이용하여 추정된 회귀계수는 지역 빈도분석의 불확실성을 표현하기 위하여 신뢰구간을 산정해야할 필요가 있다. 분석하고자 하는 일정 유의수준, α에서의 회귀계수에 대한 $100(1-\alpha)$ % 신뢰구간은 t 분포를 이용하여 다음과 같이 나타낼 수 있다.

$$\hat{\beta}_i \pm \ se(\hat{\beta}_i) \cdot t_{(\alpha/2,\ n-p-1)} \tag{4}$$

$$se(\hat{\beta}_i) = \sqrt{c_{ii} \cdot MSE} \tag{5}$$

여기서, p 는 회귀모형의 자유도를 나타내고, c_{ii}는 $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$의 대각행렬 요소값, MSE는 평균제곱오차(mean square error)를 나타낸다. 그러나 위에서 산정된 회귀계수의 신뢰구간의 상하한값들은 회귀계수의 불확실성을 나타낼 뿐이고, 지역 빈도분석결과에 대한 불확실성은 주어진 설명변수들에 대한 평균반응치(mean response)에 대한 신뢰구간을 산정함으로서 나타낼 수 있다. 각 설명변수 $\mathrm{X}=\mathrm{a}$ 인 경우, $100(1-\alpha)$ %에 대한 신뢰구간은 Eq. (6)과 같이 나타낼 수 있으며, 이를 이용하여 각 재현기간에 해당되는 빈도유량의 불확실성을 표현할 수 있다.

$$\mathrm{a}^{\mathrm{T}}\hat{\beta} \pm t_{(\alpha/2,\ n-p-1)}\sqrt{MSE \cdot \mathrm{a}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{a}^{\mathrm{T}}} \tag{6}$$

3. Bayesian 통계학과 Bayesian 다중회귀분석

베이즈의 정리는 A가 먼저 발생하고 그 후에 B가 발생하는 두 개의 사건 A, B가 서로 종속적일 경우 A의 사건에 의해 B 사건의 확률이 달라진다는 것이다. 베이즈의 정리를 수식으로 나타내면 Eq. (7)과 같고, 여기서 각각의 확률 사건을 연속 확률밀도함수(probability density function)로 나타내면 베이즈의 정리는 Eq. (8)과 같이 표현될 수 있다(Sorensen and Gianola, 2002).

$$\Pr(B_j|A) = \frac{\Pr(B_j)\Pr(A|B_j)}{\sum_{j=1}^{n}\Pr(A|B_j)\Pr(B_j)} \tag{7}$$

$$\pi(\theta|x_1, x_2, ..., x_n) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)\pi(\theta)}{\int_{\theta} f(x_1|\theta)\cdots f(x_n|\theta)\pi(\theta)d\theta} \tag{8}$$

Eq. (8)에서 좌변의 $\pi(\theta|x_1, x_2, ..., x_n)$는 사후분포(posterior distribution), 우변 분자의 $\pi(\theta)$는 사전분포(prior distribution)라 명명되며, 우변의 분모는 상수로서 주변분포(marginal distribution)이고, 우변 분자의 $f(x_1|\theta)\cdots f(x_n|\theta)$는 발생할 수 있는 모든 가능성을 고려한 우도함수 (likelihood function)이다. 그러므로 Eq. (8)로부터 사후분포는 우도함수와 사전분포의 곱에 비례하게 됨을 알 수 있다. 분석하고자 하는 자료를 나타낼 수 있는 확률밀도함수가 결정되면 이로부터 우도함수를 유도할 수 있고, 적절한 사전분포를 부여함으로써 사후분포로부터 확률밀도함수의 매개변수를 추출하고 매개변수의 불확실성을 탐색할 수 있다. Bayesian 방법을 이용한 매개변수의 추정은 매개변수를 미지의 상수로 간주하는 것이 아니라 미지의 난수로 간주하게 됨으로써 추정의 관심이 되는 매개변수의 불확실성의 정도를 확률 모형을 이용하여 표현할 수 있게 된다. 결국 Bayesian 방법을 이용한 매개변수의 추정은 자료로부터 얻은 매개변수에 대한 정보와 매개변수에 대한 과거의 경험 또는 주관을 사전분포로 표현함으로써 보다 정확한 매개변수의 불확실성에 대한 탐색에 그 목적이 있다고 할 수 있다.

또한 다중회귀분석은 종속변수에 영향을 미치는 독립변수가 두 개 이상인 경우에 독립변수에 따른 종속변수간의 변화를 나타내기 위하여 사용된다. 본 연구에서 사용된 회귀모형은 Eq. (9)와 같은 비선형 모형이지만 회귀분석을 위하여 양변에 로그변환을 취하여 Eq. (10)과 같은 선형 회귀식으로 변환함으로써 선형회귀분석을 수행하였다.

$$y = \beta_0(x_1^{\beta_1})(x_2^{\beta_2}).....(x_p^{\beta_p}) \tag{9}$$

$$\log y = \log \beta_0 + \beta_1 \log x_1 + ...\beta_p \log x_p \qquad (10)$$

위 식에서 y는 종속변수, x는 설명변수들, β는 회귀상수 및 계수들이다.

Bayesian 회귀분석은 최소자승법을 회귀분석에 적용하는 과정에서 확률적 개념을 이용하는 것으로부터 시작된다. 즉 이는 최소자승법에 의해 표현되는 회귀분석 모형의 오차를 평균 0과 분산 σ^2을 가지는 각각의 정규분포(normal distribution)에 대한 조건부확률을 이용하여 표현할 수 있다고 가정하는 것이고 이를 수식으로 나타내면 다음과 같다(Martz and Waller, 1982).

$$p(\epsilon|\sigma^2) \approx N(\epsilon|0,\sigma^2) = (\frac{1}{2\pi\sigma^2})^{1/2}(\exp - \frac{1}{2\sigma^2}\epsilon^2) \qquad (11)$$

그러므로 설명변수와 회귀계수가 주어지는 경우 이에 대한 종속변수의 조건부 확률은 최소자승법의 특성과 Eq. (11)을 이용하여 Eq. (12)로 나타낼 수 있으며, Eq. (12)에서 $z = y - X\beta$ 로 놓고 발생할 수 있는 모든 경우를 나타내는 우도함수를 구하면 Eq. (13)과 같다.

$$p(\mathbf{y}\,|\mathbf{X},\ \boldsymbol{\beta},\ \sigma^2) = (\frac{1}{2\pi\sigma^2})^{1/2}\exp(-\frac{1}{2\sigma^2}|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2)\} \qquad (12)$$

$$L(\mathbf{y}\,|\boldsymbol{\beta},\ \sigma^2) = (\frac{1}{2\pi\sigma^2})^{n/2}\exp(-\frac{1}{2\sigma^2}\,\mathbf{z}^{\mathrm{T}}\mathbf{z})\Big\} \qquad (13)$$

Eq. (8)의 연속분포에 대한 베이즈의 정리를 주어진 변수 β와 σ^2에 대하여 다시 표현하면 다음과 같이 표현할 수 있으며, Eq. (14)에서 $\pi(\beta,\sigma^2)$가 사전분포이고 분모인 주변확률분포는 적분하여 임의의 상수로 표현될 수 있다.

$$\pi(\beta,\ \sigma^2|y\ ) = \frac{L(y|\beta,\ \sigma^2)\pi(\beta,\sigma^2)}{\int_{\sigma^2}\int_{\beta} L(y|\beta,\ \sigma^2)\pi(\beta,\sigma^2)d\beta d\sigma^2}\Big\} \qquad (14)$$

Eq. (14)에서 사전분포를 적절히 선정하는 것은 Bayesian 방법을 이용하여 지역 빈도분석을 수행하는데 있어서 가장 중요한 부분이라 할 수 있다. 사전 분포는 크게 자료에 기반한 사전분포와 자료에 기반하지 않은 사전분포로 구분할 수 있는데, 본 연구에서 적용되는 회귀분석의 경우에는 회귀계수들에 대한 자료에 기반한 사전분포를 각 대상 유역별로 구축하는 것이 불가능하다.

자료에 기반한 사전분포를 구성하기 위해서는 점 빈도분석의 경우 인근 유량자료로부터 분석하고자 하는 지점의 사전분포를 유도할 수 있으나, 회귀분석의 경우에는 회귀계수에 대한 인근 자료를 이용할 수가 없어 자료에 기반한 사전분포를 사용하는 것이 불가능하므로 본 연구에서는 Sorensen and Gianola (2002)가 제안한 다음과 같은 균일분포를 사용하였다. 이와 같은 균일분포는 회귀계수에 대한 사전정보를 전혀 알 수 없다는 것을 반영한 것으로써 사전분포가 모형의 분산에만 관련되어짐을 나타낸 것이다.

$$\pi(\boldsymbol{\beta},\ \sigma^2) \propto \frac{1}{\sigma^2} \qquad (15)$$

앞서 언급한 바와 같이 주변분포는 적분하여 상수가 되므로, 제안된 우도함수와 사전분포를 이용하여 Eq. (14)를 나타내면 다음과 같다.

$$\pi(\beta,\ \sigma^2|y\ ) \propto (\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\,z^{\mathrm{T}}z)\sigma^{-2} \qquad (16)$$

수식의 전개를 쉽게 하기 위하여 $\beta = [\ \beta_0, \beta_1]$인 경우만을 고려하면, 다음 Eq. (17)과 같이 정리되고 이를 다시 행렬표기로 나타내면 Eq. (18)과 같다.

$$\begin{aligned} z^{\mathrm{T}}z &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1)^2 \\ &= (n-2)s^2 + n(\beta_0 - \hat{\beta}_0)^2 + (\beta_1 - \hat{\beta}_1)^2\sum_{i=1}^{n}x_i^2 \qquad (17) \\ &\quad + 2(\beta_0 - \hat{\beta}_0)(\beta_1 - \hat{\beta}_1)\sum_{i=1}^{n}x_i \\ &= (n-2)s^2 + (\beta - \hat{\beta})^{\mathrm{T}}(X^{\mathrm{T}}X)(\beta - \hat{\boldsymbol{\beta}}) \qquad (18) \end{aligned}$$

Eq. (18)을 Eq. (16)에 대입하면 Eq. (19)로 정리될 수 있으며, Eq. (19)가 균일분포가 적용된 Bayesian 다중회귀분석을 위한 수식이라 할 수 있다.

$$\begin{aligned} \pi(\beta,\sigma^2|y) &\propto (\sigma^2)^{-n/2}\exp\left[-\frac{(n-2)s^2}{2\sigma^2}\right] \\ &\times \sigma^{-2}\exp\left[-\frac{(n-2)s^2}{2\sigma^2}(\beta-\hat{\beta})^{\mathrm{T}}(X^{\mathrm{T}}X)(\beta-\hat{\beta})\right] \end{aligned} \qquad (19)$$

Eq. (19)의 우변의 전항을 $p(\sigma^2|y)$로 표기하고 후항

을 $p(\beta,\sigma^2|y)$로 간략화하면 최종적인 수식은 다음과 같이 정리될 수 있다.

$$\pi(\beta,\ \sigma^2|y\ ) \propto p(\sigma^2|y) \bullet p(\beta,\sigma^2|y) \tag{20}$$

또한 매개변수 α,γ를 가지는 역감마분포(inverse gamma distribution)는 다음의 식으로 나타낼 수 있다.

$$f(x) \propto x^{-(\alpha+1)} \exp[-(\frac{\gamma}{x})], \quad 0 < x < \infty \tag{21}$$

그러므로 Eq. (21)에서 $x=\sigma^2,\ \alpha=(n-2)/2,\ \gamma=(n-2)s^2/2$로 두면, Eq. (20)의 우변의 전항은 Eq. (22)와 같이 역감마분포를 따르는 것을 알 수 있으며, 후항은 Eq. (23)과 같이 정규분포를 따르는 것을 알 수 있다.

$$\sigma^2|y \sim IG(\frac{n-2}{2},\ \frac{(n-2)s^2}{2}\Big\} \tag{22}$$

$$\beta,\sigma^2|y \sim N(\hat{\beta},\ \sigma^2(X^{T}X)^{-1}) \tag{23}$$

그러므로 구하고자 하는 회귀계수, β 를 추정하기 위해서는 먼저 $\hat{\beta}$, $(X^{T}X)^{-1}$, s^2을 자료로부터 산정한 후 이 값들을 이용하여 Eq. (22)로부터 σ^2을 생성하고, 생성된 σ^2을 이용하여 Eq. (23)으로부터 최종적으로 β 를 생성시킴으로서 회귀계수를 얻을 수 있으며 이로부터 회귀계수의 평균추정치와 원하는 유의수준에서의 신뢰구간을 산정할 수 있다.

4. 다중회귀분석의 수행 및 결과

본 연구에서는 위의 이론적 배경에서 제시된 OLS를 이용한 Bayesian 다중회귀분석을 수행함으로써 낙동강 유역에서의 지역 빈도분석을 수행하고 각 재현기간에서의 빈도유량을 산정한 후, 불확실성을 표현할 수 있는 신뢰구간의 상한값과 하한값을 추정하였다. 또한 Bayesian 다중회귀분석의 결과를 기존 방법의 결과와 비교하기 위하여 다중회귀분석을 수행하고 t 분포를 이용한 신뢰구간으로부터 회귀계수와 종속변수의 불확실성을 산정하였다.

4.1 자료의 선정 및 대상유역의 수문학적 동질성 파악

구축된 Bayesian 회귀분석모형은 Fig. 1과 같은 낙동강 유역에 대하여 적용되었다. 낙동강 유역은 유역면적 23,702 km^2 으로 수자원단위지도상에서 중권역 22개와 표준유역 191개로 구성되어져있다. 본 연구에서 사용된 낙동강 유역의 유역분할은 본류지점 및 진동지점 하류유역을 제외한 10개 중권역을 대상으로 하였으며, 미계측 유역에의 적용을 위하여 8번과 10번 유역은 수자원단위지도상의 2개씩의 중권역을 1개의 중권역으로 구성한 뒤 사용하였다. 특히 본류지점은 이후 설명될 자연유량의 산정을 위한 수문모형의 적용 시 제외되어져 지역빈도분석에서 제외되었으며, 진동지점 이후의 유역은 낙동강 본류를 포함하는 유역이므로 위와 같은 이유를 고려하여 연구에서 제외하였다.

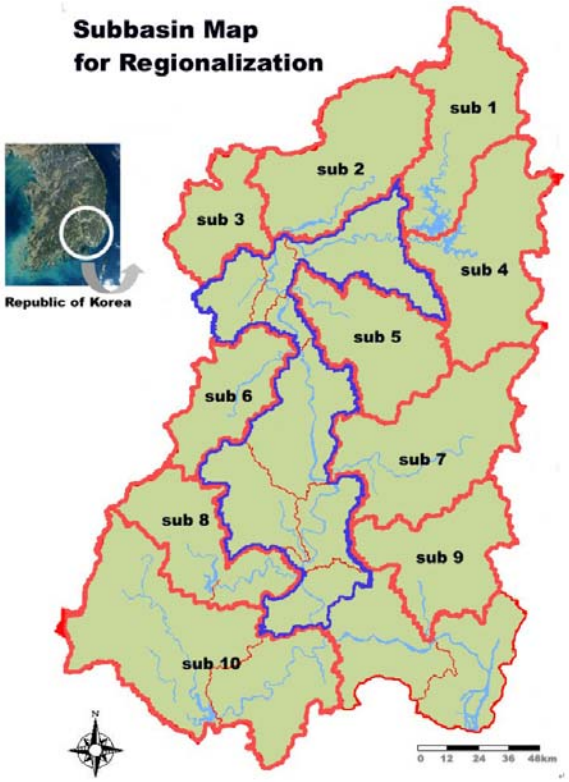

Fig. 1. The selected basin and sub-basin map

저수량(low flow) 빈도분석을 수행하고자 하는 경우 가장 먼저 산정해야 하는 것은 자연유량(natural flow)이다. 자연유량이란, 유역 내 인위적인 유량이 조절이 없는 자연 상태의 유역으로부터 발생되는 유역의 고유한 유량으로써 실측자료를 자연유량으로 환산하기 위해서는 유역 내 존재하는 댐으로 인한 조절량(유입량, 방류량, 수면 증발량 등), 실제 취수량과 같은 인위적 조절 유량의 일자료가 필요하다. 그러나 국내 축적된 자료를 이용하여 임의 지점에서 자연유량을 산정하는 것은 필요 자료의 부족으로 인하여 올바른 자연유량이 산정되기 매우 어려운 실정이다. 이와 같은 자료의 부족

및 부정확성에 대한 문제는 저수량(low flow) 빈도분석과 그 결과값을 이용한 유지유량 또는 환경유량 등의 설정에 매우 큰 영향을 미칠 수 있으므로 시급하게 해결되어져야 하는 문제라고 할 수 있다. 일반적으로 취수 허가량은 실제 취수량보다 큰 값을 가지는 것으로 추정되고 있으며, 특히 농업용수의 경우 계절별로 허가량과 실제 취수량과의 차이가 매우 크므로 자연유량 산정 시에는 필히 실제 취수량을 사용할 필요가 있다. 현재 건교부에서는 이를 위하여 '하천유수사용 실적관리 시스템'을 지속적으로 운영하고 있으며, 조만간 이 시스템을 이용하여 얻어진 실제 취수량을 자연유량의 산정에 이용할 수 있을 것으로 판단된다.

위와 같은 자료의 한계성으로 인하여 본 연구에서는 10개 소유역의 출구 유출량에 대한 자연유량을 구하기 위하여 건교부와 한국수자원공사(2006)가 수행한 낙동강 유역의 유역조사사업 결과를 이용하였다. 이 연구에서는 수자원단위지도상의 중권역별 유출량에 대한 자연유량을 산정하기 위하여 PRMS (precipitation-runoff modeling system)모형을 사용함으로써 자연유량을 모의하였다. 수문모형의 매개변수의 보정은 자연유량을 산정하는 데 있어서 가장 어렵고도 중요한 문제라 할 수 있는데 유역조사사업의 경우에는 댐이 없는 유역, 취수시설이 적은 유역, 자료의 길이가 충분한 유역을 조건으로 하여 안동댐 유역, 임하댐 유역, 합천댐 유역과 더불어 한강 유역의 도암댐 유역과 괴산댐 유역을 추가적으로 선정하여 이들 유역의 유량자료를 이용하여 모형의 매개변수를 산정한 후, 보정된 매개변수를 수문학적 특성이 유사한 낙동강 유역의 타 중권역으로 전이하여 최종적으로 자연유출량을 산정한 바 있으며, 본 연구에서는 위와 같은 과정으로 모의된 자연유량의 1966년부터 2001년까지의 36년간의 자료를 이용하여 미국 등에서 주로 이용되고 있는 7일 지속기간 년 최소유량(7Q)을 각 소유역별로 먼저 산정하였다.

지역 빈도분석을 수행하기 위한 첫 번째의 절차로써 Fig. 1의 10개의 소유역에 대한 수문학적 동질성을 규명하기 위하여 K-means 알고리즘을 이용한 군집분석을 수행하였으며, 그 결과를 Kaufmann and Rousseeuw (1990)이 제시한 실루엣 추정치(silhouette value)를 산정하여 동질유역을 구분하였다. 군집분석을 위한 자료로는 10개 소유역에 대해 PRMS 모형을 이용한 모의자료를 이용하였으며, 홍수기로 7월부터 9월의 유량자료를 제외한 일유량 자료를 사용하였다. Fig. 2는 10개 소유역에 대한 군집분석에 대한 결과를 나타낸 것으로써, 10개의 소유역을 하나의 군집으로 분석한 경우 모든 유역에 대한 평균 실루엣 추정치가 0.6 이상으로 나타났으므로 10개의 소유역이 수문학적으로 동질하다는 가정을 수립할 수 있었고 이로부터 10개 소유역을 대상으로 하는 1개의 회귀모형을 구축하는 것이 무리가 없다는 것을 예측할 수 있었다.

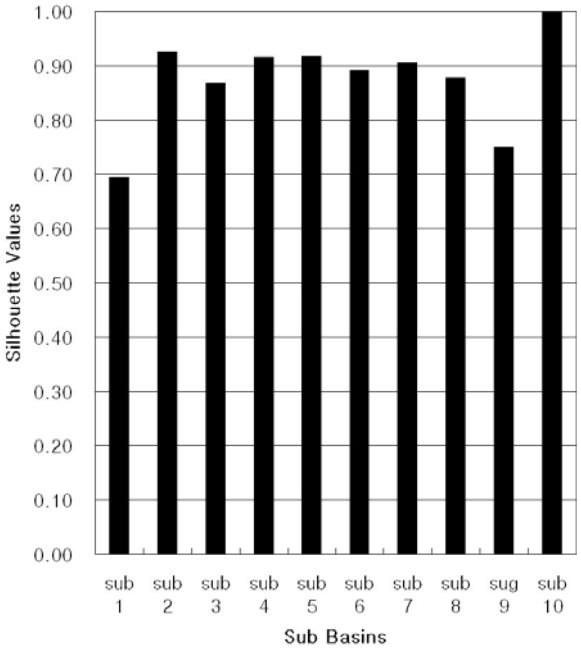

Fig. 2. Results of identification in the Nakdong River basin

4.2 다중회귀모형의 구축과 회귀모형의 적정성 검토

본 연구에서 처음으로 고려된 다중회귀모형의 설명변수로는 면적 평균 강우량, 유역면적, 삼림면적 비율, 유역밀도, 유역평균경사, 유역둘레, 유역 평균폭, 표고별 누가면적이 선정되었으나, 이 중에서 유역 평균둘레, 유역 평균폭, 표고별 누가면적은 유역면적과의 상관계수가 0.9를 넘어 다중공선성이 발생할 가능성이 크기 때문에 분석에서 제외하고 5개만을 설명변수로 선택하였다. 선정된 5개의 상관계수는 Table 1과 같으며, 선정된 설명변수의 각 값들은 수자원종합관리시스템을 통해 수집하여 Table 2에 표시하였다.

1) 7월부터 9월사이의 강우량을 제외한 36년간 전체 면적 평균 강우량(P)
2) 유역면적(A)
3) 삼림면적 비율(PBF)
4) 유역밀도(=유로연장/유역면적, D)
5) 유역평균경사(BS)

또한 본 연구에서 제안된 회귀모형의 종속변수는 빈도유량이 사용되어져야 하므로 김상욱(2008)이 수행한

Table 1. Result of Correlation Coefficient between Explanatory Variables

Variables	P (mm)	A (km^2)	PBF (%)	D	BS
P (mm)	1.0000	0.3878	-0.0259	-0.2415	0.2007
A (km^2)	0.3878	1.0000	0.1699	-0.8427	-0.0349
PBF (%)	-0.0259	0.1699	1.0000	-0.1002	0.5483
D	-0.2415	-0.8427	-0.1002	1.0000	-0.1570
BS	0.2007	-0.0349	0.5483	-0.1570	1.0000

Table 2. The used Values for Explanatory Variables

Sub-basin	P (mm)	A (km^2)	PBF (%)	D	BS
1	494.10	1628.68	85.52	0.10	26.82
2	532.20	1158.80	65.15	0.10	26.43
3	571.50	609.42	78.55	0.12	33.43
4	458.80	1975.77	81.61	0.06	36.82
5	465.30	1314.32	72.37	0.09	19.11
6	495.70	599.98	68.97	0.12	27.10
7	501.80	1533.25	66.42	0.08	18.57
8	1156.60	1240.69	75.01	0.09	28.01
9	587.30	1292.75	74.75	0.08	30.53
10	1345.70	3005.29	72.79	0.06	29.30

* Data source : www.wamis.go.kr

점 빈도분석에서의 과정을 그대로 이용하였다. 즉 산정된 7Q유량을 이용하여 년 최소 7Q유량을 산정한 후, 2변수 Weibull 확률분포식을 적용하였으며 확률분포식의 모수는 최우추정법(maximum likelihood estimation method)을 사용하여 추정하고, 최종적으로 분위수를 산정하여 각 빈도별, 소유역별로 각각 종속변수를 구성하여 회귀분석을 수행하였다.

Bayesian 회귀분석과 회귀분석 결과를 이용하여 신뢰구간을 산정하기에 앞서서 구축된 다중회귀모형의 안정성을 검정하였다. 먼저 OLS의 사용 가정이 되는 등분산성의 경우 White 테스트 결과를 Table 3에 나타내었으며 잔차플롯을 Fig. 3에 나타내었다. White 테스트 결과 95 % 신뢰구간에서 등분성의 귀무가설이 기각되지 않았으므로 95 % 신뢰구간에서 잔차의 분산이 등분산성을 가진다고 설명할 수 있어 OLS를 사용하는 가정에 크게 위배되지 않음을 알 수 있었다. 그러나 GLS를 사용하면 불확실성의 근원을 종류별로 분석할 수 있는 장점이 있어 향후에는 GLS를 사용하여 Bayesian 회귀분석을 수행한 결과를 상호 비교해 보는 것이 바람직할 것으로 판단된다. 또한 Fig. 3에는 각각의 잔차를 도시하였는데 대부분의 잔차의 도시가 치중되어 있지 않고 균등하게 퍼져있는 것을 확인할 수 있었다.

또한 Table 4에는 검정항목과 각 항목별 검정결과를 나타내었는데 결정계수 및 조정 결정계수는 아주 높은 편은 아니었지만, 합리적인 수치를 보이는 것을 알 수 있었으며, Durbin-Watson 검정치는 2.263으로서 10보다 작아 잔차간 자기상관성이 없음을 가정하기에 충분함을 알 수 있었다. 또한 VIF는 모든 설명변수에서 10보다 작게 나타나 다중공선성에도 영향을 받지 않는다고 할 수 있으며, Cook의 거리와 leverage 통계치도 각각 1과 1.2보다 작은 값으로 나타나 구축된 다중회귀모형이 안정적임을 알 수 있었다.

Table 3. The used Values for Explanatory Variables

Significance level (α)	Test statistic (F)	p-value	Result
0.05	4.9648	5.0503	Not reject
0.1	4.9648	3.4530	Reject

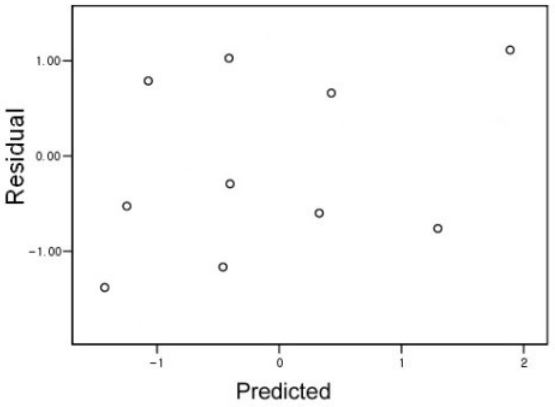

Fig. 3. Residual Plot with 10 Sub-basins

Table 4. Statistical Tests for Multiple Regression Model

Test Items		Test Values
1) Coeff. of determination		0.831
2) Adjusted coeff. of determination		0.649
3) Durbin-Watson statistic		2.263
4) VIF for explanatory variables	P	1.517
	A	6.398
	BFA	2.018
	D	5.599
	BS	2.491
5) Cook's distance	Sub-basin 1	0.555
	Sub-basin 2	0.388
	Sub-basin 3	0.222
	Sub-basin 4	0.010
	Sub-basin 5	0.065
	Sub-basin 6	0.246
	Sub-basin 7	0.038
	Sub-basin 8	0.010
	Sub-basin 9	0.122
	Sub-basin 10	0.385
6) Leverage statistic	Sub-basin 1	0.831
	Sub-basin 2	0.722
	Sub-basin 3	0.631
	Sub-basin 4	0.440
	Sub-basin 5	0.409
	Sub-basin 6	0.328
	Sub-basin 7	0.461
	Sub-basin 8	0.451
	Sub-basin 9	0.105
	Sub-basin 10	0.623

4.3 적용결과의 비교 및 빈도곡선의 작성

안정성이 검증된 회귀모형을 이용하여 Bayesian 회귀분석을 수행하였다. Bayesian 회귀분석을 수행하기 위해서는 이론적 배경에서 언급한 바와 같이 먼저 $\hat{\beta}$, $(X^TX)^{-1}$, s^2을 자료로부터 산정한 후 이 값들을 이용하여 위의 Eq. (22)로부터 σ^2을 생성하고, 생성된 σ^2을 이용하여 위의 Eq. (23)으로부터 최종적으로 β 를 생성시킴으로서 회귀계수를 얻을 수 있으며 이로부터 회귀계수의 평균추정치와 원하는 유의수준에서의 신뢰구간을 산정할 수 있다. 본 연구에서는 위와 같은 절차를 수행하기 하기 위하여 10,000개의 σ^2을 생성함으로써 최종적으로 10,000개의 β 를 추정하였으며, 최초 100개의 추정치는 모형의 안정성을 위해서 제거하고 평균추정치와 신뢰구간을 산정하였다. 또한 Bayesian 회귀분석결과와의 비교를 위하여 위 Eqs. (4) and (5)를 이용하여 t 분포를 이용한 회귀계수의 하한값과 상한값을 추정하였다. Table 5에는 t 분포를 이용한 기존방법과 Bayesian 회귀분석결과의 95 % 신뢰구간에서의 추정치를 나타내었으며, 지면 상 10년, 50년, 100년 빈도에 해당하는 추정치만을 표시하였다.

본 연구의 목적은 회귀계수의 불확실성에 대한 분석이기 보다는 회귀계수를 이용하여 산정된 종속변수, 즉 추정된 빈도유량을 이용하여 불확실성 측면에서 Bayesian 회귀분석이 기존 방법과 어떤 차이가 있으며 어떤 점에서 유리한지를 알아보고자하는 것이라 할 수 있다. 그러므로 Table 5에서 나타난 Bayesian 회귀분석에 의한 회귀계수를 이용한 종속변수의 추정결과와 위의 Eq. (6)을 이용한 종속변수의 추정결과를 비교해야 한다. 이를 위하여 추정된 각각의 회귀계수를 이용하여 빈도유량을 추정하였으며 소유역 1과 4에 대해서만 간략히 결과를 정리하면 Tables 6 및 7과 같다.

Table 6에서 기존 방법과 Bayesian 회귀분석과의 평균값에서의 차이는 약 0.000 m^3/s에서 약 0.043 m^3/s의 범위 안에 존재하고, Table 7에서는 약 0.008 m^3/s와 약 0.078 m^3/s사이에서 존재하는 것을 알 수 있다. 즉 이로부터 기존 방법과 Bayesian 회귀분석의 평균값에서는 큰 차이가 없이 비슷한 결과로 산정되는 것을 알 수 있다. 그러므로 불확실성을 고려하지 않고 지역빈도분석을 수행하는 경우에는 복잡한 Bayesian 회귀분석을 사용하는 것보다 널리 사용되는 각 종 통계 패키지를 이용하여 일반적인 회귀분석을 수행하는 것이 노력의 감소 측면에서 우월함을 알 수 있다.

그러나 불확실성을 고려하는 경우에는 이와 다른 결과를 보인다. Table 6에서 기존 방법의 상한값과 하한값의 차이를 나타내는 (3)-(1)은 약 0.625 m^3/s에서 약 2.920 m^3/s의 범위 안에 존재 하지만 Bayesian 회귀분

Table 5. Results of Regression Coefficient

RP	RC	Con. (2.5 %)	Con. (Mean)	Con. (97.5 %)	Bayesian (2.5 %)	Bayesian (Mean)	Bayesian (97.5 %)
10	β_0	-10.019	-9.651	-9.283	-10.031	-9.662	-9.572
	β_1	1.519	1.619	1.720	1.520	1.621	1.703
	β_2	1.381	1.452	1.523	1.383	1.453	1.507
	β_3	0.779	0.814	0.850	0.780	0.815	0.841
	β_4	1.452	1.510	1.568	1.454	1.512	1.552
	β_5	0.032	0.063	0.094	0.032	0.063	0.093
50	β_0	-13.790	-13.275	-12.761	-13.807	-13.425	-13.158
	β_1	1.990	2.131	2.272	1.992	2.155	2.249
	β_2	1.786	1.885	1.984	1.788	1.906	1.964
	β_3	1.016	1.066	1.116	1.017	1.078	1.104
	β_4	1.930	2.011	2.092	1.932	2.034	2.071
	β_5	0.044	0.088	0.132	0.044	0.089	0.130
100	β_0	-15.360	-14.788	-14.217	-15.379	-14.954	-14.659
	β_1	2.173	2.329	2.486	2.175	2.355	2.460
	β_2	1.961	2.071	2.181	1.963	2.094	2.159
	β_3	1.112	1.168	1.223	1.114	1.181	1.211
	β_4	2.121	2.211	2.301	2.124	2.236	2.278
	β_5	0.057	0.105	0.154	0.057	0.107	0.152

* RP : Return period, RC : Regression coefficient

Table 6. Results of Low Flow Characteristics (sub-basin 1: Andong Dam sub-basin) (m^3/s)

Return Period	Con. (2.5 %) (1)	Con. (Mean) (2)	Con. (97.5 %) (3)	Bayesian (2.5 %) (4)	Bayesian (Mean) (5)	Bayesian (97.5 %) (6)	(3) - (1)	(6) - (4)	(5) - (2)
10	0.042	0.353	2.962	0.066	0.396	1.623	2.920	1.557	0.043
20	0.013	0.158	1.887	0.021	0.172	1.016	1.874	0.995	0.014
30	0.007	0.097	1.451	0.010	0.102	0.770	1.444	0.760	0.005
40	0.004	0.069	1.190	0.006	0.069	0.622	1.186	0.616	0.000
50	0.003	0.053	1.027	0.004	0.050	0.530	1.024	0.526	-0.003
60	0.002	0.042	0.897	0.003	0.038	0.457	0.895	0.454	-0.004
70	0.002	0.035	0.792	0.002	0.030	0.398	0.790	0.396	-0.005
80	0.001	0.030	0.736	0.002	0.024	0.366	0.735	0.364	-0.006
90	0.001	0.026	0.690	0.002	0.020	0.340	0.689	0.338	-0.006
100	0.000	0.023	0.625	0.001	0.016	0.303	0.625	0.302	-0.007

석의 경우에는 상한값과 하한값의 차이((6)-(4))는 약 0.302 m^3/s에서 약 1.557 m^3/s를 보이고 Table 7에서는 기존 방법의 경우 약 0.207 m^3/s에서 약 1.377 m^3/s를, Bayesian 회귀분석에서는 약 0.101 m^3/s에서 약 0.647 m^3/s를 보이는 것을 알 수 있다. 그러므로 불확실성 측면을 고려해야 하는 경우에는 Bayesian 회귀분석에 의해 산정된 상한값과 하한값의 차이가 기존 방법보다 훨씬 작은 값으로 산정됨을 알 수 있었으므로, 수자원의 관리 측면에서 불확실성이 강조되는 시점에서 기존 방법보다 Bayesian 회귀분석을 사용하여 지역 빈도분석을 수행하는 것이 기존 방법보다 훨씬 불확실성을 감소시켜 나타낼 수 있음을 알 수 있었다. 이와 같은 결과는 기존 방법에 있어서 불확실성을 산정하기 위한 여러 가지 가정이 최종적인 결과를 과대추정하게 함으

Table 7. Results of Low Flow Characteristics (sub-basin 4: Imha Dam sub-basin) (m^3/s)

Return Period	Con. (2.5 %) (1)	Con. (Mean) (2)	Con. (97.5 %) (3)	Bayesian (2.5 %) (4)	Bayesian (Mean) (5)	Bayesian (97.5 %) (6)	(3) - (1)	(6) - (4)	(5) - (2)
10	0.021	0.170	1.398	0.037	0.248	0.684	1.377	0.647	0.078
20	0.006	0.068	0.798	0.010	0.108	0.390	0.792	0.380	0.039
30	0.003	0.040	0.579	0.005	0.059	0.283	0.576	0.278	0.020
40	0.002	0.027	0.456	0.003	0.047	0.223	0.454	0.220	0.020
50	0.001	0.020	0.375	0.002	0.036	0.184	0.374	0.182	0.016
60	0.001	0.015	0.320	0.001	0.028	0.156	0.319	0.155	0.013
70	0.001	0.013	0.276	0.001	0.023	0.135	0.276	0.134	0.010
80	0.000	0.011	0.252	0.001	0.020	0.123	0.252	0.122	0.010
90	0.000	0.009	0.231	0.001	0.018	0.113	0.231	0.112	0.009
100	0.000	0.008	0.207	0.001	0.015	0.101	0.207	0.101	0.008

로써 발생하는 것으로 간주할 수 있으며, Bayesian 회귀분석의 경우에는 불확실성을 산정하기 위한 어떠한 가정도 사용하지 않기 때문에 보다 감소된 불확실성을 산정하게 됨을 알 수 있다.

최종적으로 산정된 각 재현기간별 유량을 이용하여 10개 소유역에 대한 빈도곡선을 도시할 수 있으며, 빈도곡선의 작성 시에 저수량을 표시하는 데 합리적으로 알려진 다음과 같은 Gringorten (1963)의 공식을 사용하였으며, n은 사용된 자료의 개수를 의미한다. 확률도시공식을 선정하는 데 있어서 여러 가지 공식을 이용하여 실측값의 위치를 산정해 보고 그 중 가장 우수한 공식을 사용해야 하지만, 본 연구의 주제는 실측값과 한 개의 빈도곡선이 얼마나 잘 일치하느냐 보다는 빈도곡선사이의 불확실성을 주제로 다루고 있으므로 Weibull 공식과 Gringorten 공식만을 이용하여 우수성을 판단하여 Gringorten 공식을 선정하였다.

$$\xi(i) = \frac{i - 0.44}{n + 0.12} \tag{23}$$

Figs. 4, 5, 6, 7에는 각각 1, 4, 8, 10번 소유역에 대한 빈도곡선을 나타내었으며, 그림에서 검은색 실선은 Bayesian 회귀분석결과로 아래 실선부터 2.5 %, 평균값, 97.5 %의 빈도별 유량을 나타내며 붉은 색 점선은 기존 방법에 의한 추정결과로서 마찬가지로 아래부터 2.5 %, 평균값, 97.5 %값을 나타낸다. 각 그림으로부터 기존방법에 의한 추정결과와 Bayesian 회귀분석에 의한 추정결과를 함께 표시함으로써 불확실성 측면에서 Bayesian 방법이 기존 방법보다 불확실성을 감소시켜 나타낼 수 있음을 나타내었다.

Figs. 4, 5, 6, 7에서 기존 방법과 Bayesian 방법에 의한 평균에서의 빈도곡선과 실측값과의 차이를 도해적 비교를 통하여 비교할 수 있다. 향후 여러 지점에 대한 유사 연구를 수행하고 산정된 빈도곡선을 curve fitting하여 곡선식을 제시할 수 있을 것으로 판단되며, 곡선식이 작성될 수 있다면 도해적 비교가 아닌 교차검증을 통하여 정량적인 비교를 수행할 수 있으리라 생각된다. 도해적 방법으로 실측값과 두 방법 간의 적용결과를 비교하면, 대부분의 유역에서 10년 빈도를 기준으로 10년 미만의 재현기간에서는 기존 방법이 Bayesian 방법보다 조금 더 실측치와 가깝게 저수유량을 추정하는 것을 알 수 있으며, 10년 이상에서는 Bayesian 방법이 좀 더 실측치와 가까운 유량을 산정하는 것을 알 수 있다.

5. 미계측유역의 선정 및 저수량의 예측

필요한 재현기간에서의 유량을 알기 위해서는 빈도분석을 수행해야 하며, 빈도분석을 수행하기 위해서는 일정기간 이상의 자료가 확보되어야 빈도분석 결과가 신뢰도를 가지게 된다. 그러나 대부분의 수자원 계획과 이에 따른 수공구조물의 설치, 단지 계획 등은 자료가 없는 지점에서 수행되므로 이러한 경우 자료의 결핍을 충당하기 위하여 지역 빈도분석을 수행하고 지역 빈도분석 결과로 얻어진 결과를 이용하여 필요 지점에서의 빈도유량을 추정하여 사용하게 된다. 지역 빈도분석 방법으로 회귀분석을 사용하는 경우에는 최종 결과로 회귀식을 구성하는 회귀계수의 추정치를 얻게 되며, 원하는 지점의 빈도유량을 얻기 위해서는 추정된 회귀계수와 원하는 지점에 해당되는 설명변수 값을 이용함으로

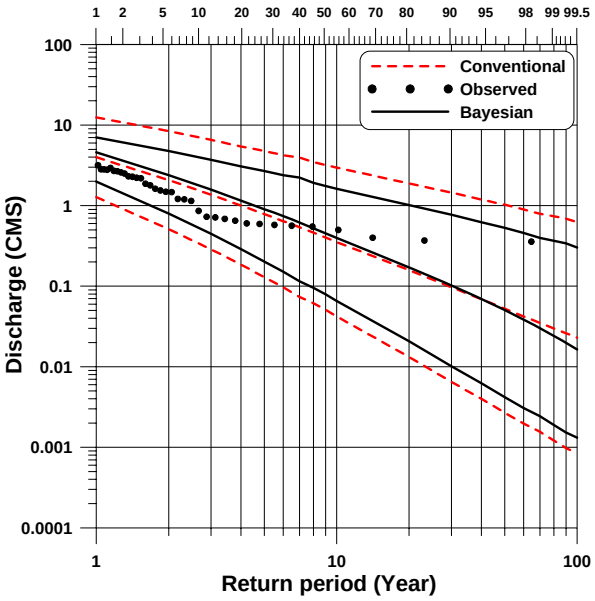

Fig. 4. Regional Frequency Curve at Sub-basin 1 (Andong Dam sub-basin)

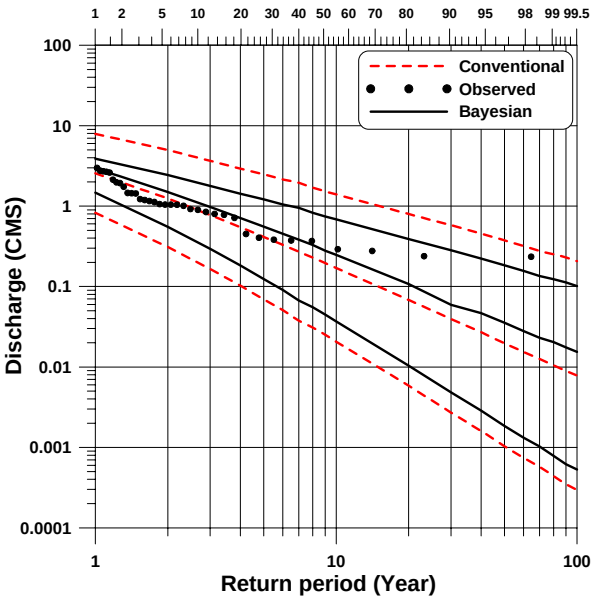

Fig. 5. Regional Frequency Curve at Sub-basin 4 (Imha Dam sub-basin)

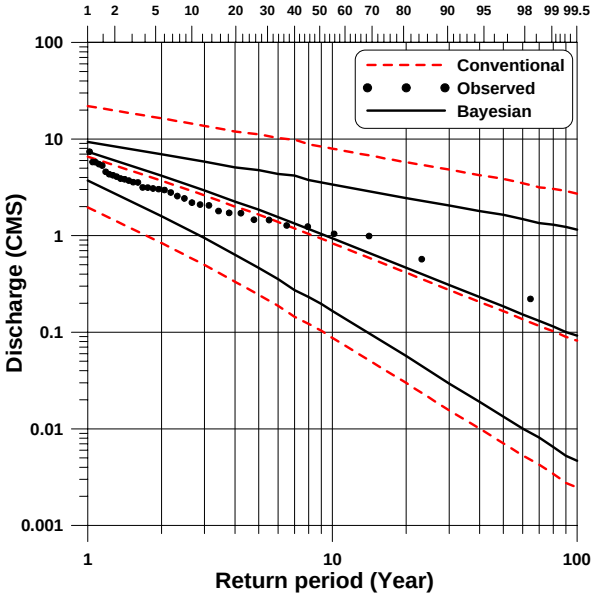

Fig. 6. Regional Frequency Curve at Sub-basin 8

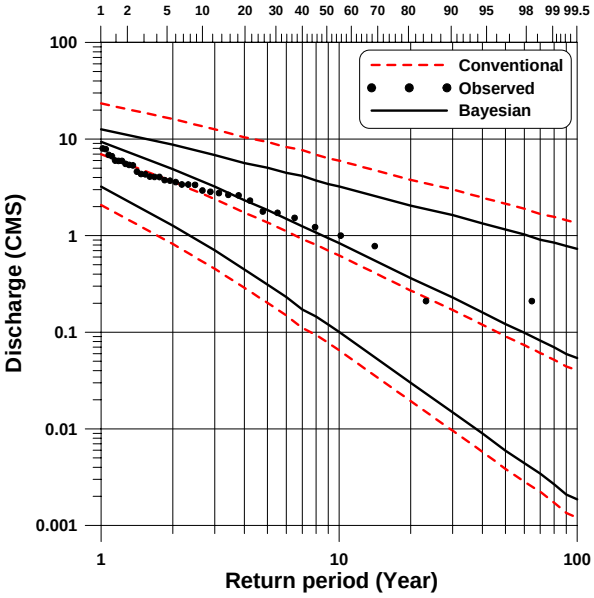

Fig. 7. Regional Frequency Curve at Sub-basin 10

써 빈도유량을 추정할 수 있다.

회귀분석 결과의 신뢰구간을 t 분포를 이용하는 경우에는 Eq. (24)와 같이 미래 예측값에 대한 신뢰구간 산정식을 이용하여 신뢰구간의 상한값과 하한값을 추정할 수 있으며, 또한 Bayesian 회귀분석의 경우에는 예측값을 추정하기 위하여 Eq. (25)와 같은 Bayesian 예측을 수행할 수 있다.

$$a^{T}\hat{\beta} \pm t_{(\alpha/2,\ n-p-1)}\sqrt{MSE \cdot [1+a(X^{T}X)^{-1}a^{T}]} \quad (24)$$

$$\pi(\bar{y}|y) = \iint \pi(\bar{y}|\beta, \sigma^2)\,\pi(\beta, \sigma^2|y)\,d\beta\,d\sigma^2 \Big\} \quad (25)$$

Eq. (25)는 사후분포로부터 추정하고자 하는 회귀계수와 분산을 생성할 수 있는 경우에는 예측을 위한 회귀모형의 잔차가 평균 0과 분산 σ^2을 가지는 정규분포를 따른다는 가정 하에 다음과 같은 과정을 통하여 추정될 수 있으며 본 연구에서는 Figs. 8과 9에 나타낸 8-1유역과 10-1유역을 미계측 유역으로 상정하여 Bayesian 예측을 수행하였다. 단, 8-1 유역과 10-1 유역은 8번 유역과 10번 유역의 일부분으로써 유역의 그

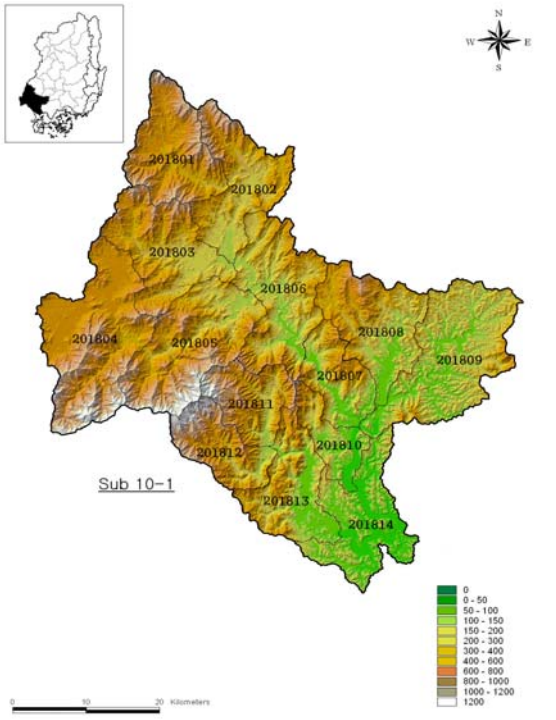

Fig. 8. Selected Catchment for Prediction (8-1)

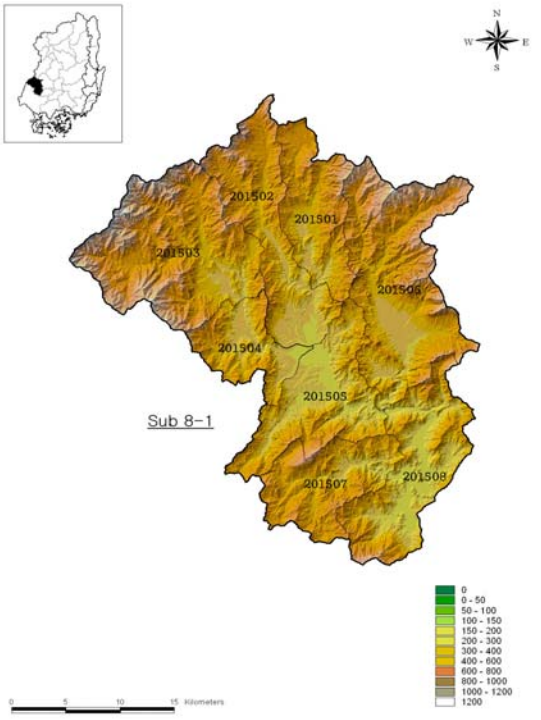

Fig. 9. Selected Catchment for Prediction (10-1)

Table 8. Explanatory Variables for Prediction (Log-transformed value)

Sub-basin	P (mm)	A (km^2)	PBF (%)	D	BS
8-1	2.76	2.97	1.90	-1.17	1.54
10-1	2.83	3.36	1.88	-1.46	1.55

* Data source: www.wamis.go.kr

Table 9. Comparison with Conventional and Bayesian Prediction (m^3/s)

Sub -basin	Con. (2.5%)	Con. (Mean)	Con. (97.5%)	Diff.	Bay. (2.5%)	Bay. (Mean)	Bay. (97.5%)	Diff.
8-1	0.043	0.418	1.753	1.710	0.056	0.423	0.765	0.709
10-1	0.015	0.218	1.098	1.083	0.020	0.221	0.479	0.459

림은 건교부와 한국수자원공사(2006)에서 수행한 낙동강 유역의 유역조사 사업에서 차용하였다.

1) 먼저 알고 있는 사후분포로부터 β, σ^2을 생성한다.

2) 잔차가 평균 0과 분산 σ^2을 가지므로, OLS로부터 예측종속변수는 평균 $\beta_0+\beta_1 x_1+...\beta_n x_n$과 분산 σ^2을 가지는 정규분포를 따르므로 $N(\beta_0+\beta_1 x_1+...\beta_n x_n, \sigma^2)$으로부터 생성할 수 있다.

본 연구에서는 Bayesian 회귀분석과정에서 생성된 10,000개의 β와 σ^2을 이용하여 정규분포로부터 다시 이에 따른 예측값을 10,000개를 생성하였다. 이 때 사용된 유역 8-1과 10-1의 설명변수는 Table 8과 같고, 두 개의 미계측 유역의 정보를 이용하여 산정한 기존 방법에 따른 예측값과 Bayesian 예측에 따른 결과는 Table 9에 나타내었다.

Table 9의 결과는 위에서 언급한 기존 방법의 경우와 Bayesian 회귀분석 방법의 특성과 유사한 결과를 나타내고 있는 것을 알 수 있다. 즉 평균값에서는 두 방법의 차이가 없이 거의 비슷한 결과를 산정하지만 불확실성을 나타내는 신뢰구간에 있어서는 Bayesian 회귀분석의 결과가 기존 방법보다 불확실성을 감소시켜 나타냄을 알 수 있다.

6. 결 론

저수량(low flow)의 분석은 저수(貯水)용량의 설계, 수자원 공급 계획, 오염원의 배치 문제, 취수의 허가 및 유지유량의 설정 등에 있어서 매우 중요한 요소이며, 다양한 저수량 분석의 지표 중에서 빈도분석에 의한 저수량은 가장 많이 사용되어 지는 지표라 할 수 있다. 그러나 자료의 길이가 부족하거나 결핍된 경우에는 점 빈도분석의 수행이 불가능하므로 지역 빈도분석을 수행하여 원하는 지점에서의 저수량을 추정해야 한다. 또한 수자원관리의 측면에서 추정된 저수량은 확정적인 값만이 필요한 것이 아니라 불확실성을 나타내는 하한값과 상한값이 함께 표시될 필요가 있다.

위와 같은 필요성을 바탕으로 본 연구에서는 지역 빈도분석을 수행함에 있어서 가장 많이 사용되는 회귀모형을 구축함에 있어 어떠한 가정도 필요없이 불확실성을 표현할 수 있는 Bayesian 방법을 회귀모형에 적용하였다. 또한 Bayesian 회귀모형의 결과를 비교 및 평가하기 위하여 t 분포를 이용하여 추정결과의 신뢰구간을 산정하는 기존 방법을 함께 적용하여 그 결과를 비교하였다.

기존 방법과 Bayesian 회귀분석에 의한 결과를 비교한 결과, 확정적인 값만은 사용하는 경우에는 Bayesian 회귀분석과 기존 방법이 큰 차이가 없어 적용과정이 복잡한 Bayesian 회귀분석은 큰 우월성을 보이지 못했다. 그러나 불확실성을 나타내는 상한값과 하한값의 차이에 있어서는 Bayesian 회귀분석이 기존 신뢰구간 산정방법 보다 불확실성을 훨씬 감소시켜 나타냄을 알 수 있었다. 그러므로 수자원관리의 측면에서 불확실성을 고려하는 것이 점점 더 중요해지는 추세를 감안할 때, Bayesian 회귀분석을 이용하여 지역 빈도분석을 수행하는 것이 불확실성을 표현하는 데 있어서 우월함을 입증하였다. 또한 향 후 연구과제로서 OLS를 이용한 Bayesian 회귀분석을 GLS를 이용하여 수행함으로써 불확실성의 정확도와 불확실성의 종류에 따른 분석을 제안할 수 있다.

이와 같은 연구는 수자원의 관리 측면에서 중요한 저수량의 관리에 있어서 확률적인 개념을 도입할 수 있을 것으로 예측되며, 이에 따라 특정한 한 가지의 계획의 실행보다는 여러 가지 요소를 감안한 융통성 있는 다양한 수자원 관리 기법을 계획하는 데 응용될 수 있으리라 판단된다.

감사의 글

본 연구는 21세기 프런티어 연구개발 사업인 수자원의 지속적 확보기술개발 사업단(과제번호 1-7-3)의 서울대학교 공학연구소를 통한 연구비 지원(30 %)과 서울대학교 BK21 안전하고 지속가능한 사회기반건설사업단의 연구비 지원(70 %)에 의해 수행되었습니다. 연구비 지원에 심심한 감사의 뜻을 표합니다.

참 고 문 헌

건교부, 한국수자원공사 (2006). **유역조사 보고서(낙동강유역).**

김상욱 (2007). **Low flow frequency analysis using Bayesian approach.** 박사학위논문, 서울대학교.

김상욱 (2008). “Bayesian MCMC를 이용한 저수량 점빈도분석: Ⅱ. 적용과 비교분석.” **한국수자원학회논문집**, 한국수자원학회, 제41권, 제1호, pp. 49-63.

Ashkar, F. and Quarda, T.B.M.J. (1998). "Approximate confidence intervals for quantiles of gamma and generalized gamma distributions." *Journal of Hydrologic Engineering*, Vol. 3, No. 1, pp. 43-51.

Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example*. John Wiley & Sons, N.Y.

Chowdhury, J.U., and Stedinger, J.R. (1991). "Confidence interval for design flood with estimated skew coefficient." *Journal of Hydraulic Engineering*, Vol. 117, No. 7, pp. 811-931.

Cohn, T.A., Lane, W.L., and Stedinger, J.R. (2001). "Confidence intervals for expected moments algorithm flood quantile estimates." *Water Resources Research*, Vol. 37, No. 6, pp. 1695-1706.

Coles, S.G., and Powell, E.A. (1996). "Bayesian methods in extreme value modeling: A review and new developments." *International Statistical Review*, Vol. 64, No. 1, pp. 119-136.

Durrans, S.R. and Tomic, S. (1996). "Regionalization of low-flow frequency estimates: An Alabama case study." *Water Resources Bulletin*, Vol. 32, No. 1, pp. 23-37.

Gringorten, I.I. (1963). "A plotting rule for extreme probability paper." *Journal of Geophysics Research*, Vol. 68, No. 3, pp. 813-814.

Hosking, J.R.M, and Wallis, J.R. (1997). *Regional Frequency Analysis*. Cambridge University Press, New York.

Kaufman L., and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster*

Analysis, John Wiley & Sons, N.Y.

Kavetski, D., Kuczera, G., and Fanks, S.W. (2006). "Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory." *Water Resources Research*, Vol. 42, W03407.

Kelly, K.S. and Krizysztofowicz, R. (1994). "Probability distributions for flood warning systems." *Water Resources Research*, Vol. 30, No. 4, pp. 1145-1152.

Kingston, G.B., Lambert, M.F., and Maier, H.R. (2005). "Bayesian training of artificial neural networks used for water resources modeling." *Water Resources Research*, Vol. 41, W12409.

Krzysztofowicz, R. (1983a). "Why should forecaster and a decision maker use Bayes theorem." *Water Resources Research*, Vol. 19, No. 2, pp. 327-336.

Krzysztofowicz, R. (1983b). "A Bayesian Markov model of the flood forecast process." *Water Resources Research*, Vol. 19, No. 6, pp. 1455-1465.

Kuczera, G. (1999). "Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference." *Water Resources Research*, Vol. 35, No. 5, pp. 1551-1557.

Kuczera, G., and Parent E. (1998). "Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm." *Journal of Hydrology*, Vol. 211, pp. 69-85.

Lee, K.S., and Kim, S.U. (2007). "Identification of uncertainty in low flow frequency analysis using Bayesian MCMC method." *Hydrological Processes*, In press(on-line published).

Madsen, H., and Rosbjerg, H.D. (1997). "Generalized least squares and empirical Bayes estimation in regional partial duration series index flood modeling." *Water Resources Research*, Vol. 33, No. 4, pp. 771-781.

Martz, H.F. and Waller, R.A. (1982). *Bayesian Reliability Analysis*. John Wiley & Sons, N.Y.

O'Connell, D.R.H., Ostenaa, D.A., Levish, D.R., Klinger, and R.E. (2002). "Bayesian flood frequency analysis with paleohydrologic bound data." *Water Resources Research*, Vol. 38, No. 5, pp. 1-14.

Reis Jr., D.S., and Stedinger, J.R. (2005). "Bayesian MCMC flood frequency analysis with historical information." *Journal of Hydrology*, Vol. 313, pp. 97-116.

Reis Jr., D.S., Stedinger, J.R., and Martins, E.S. (2005). "Bayesian generalized least squares regression with application to log Pearson type III regional skew estimation." *Water Resources Research*, Vol. 41, W10419.

Seidou, O., Ouarda, T.B.M.J., Barbet, M., Bruneau, P., and Bobee, B. (2006). "A parametric Bayesian combination of local and regional information in flood frequency analysis." *Water Resources Research*, Vol. 42, W11408.

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC methods in Quantitative Genetics*. Springer-Verlag, New York.

Stedinger, J.R. (1983). "Confidence intervals for design events." *Journal of Hydraulic Engineering*, Vol. 109, No. 1, pp. 13-27.

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou, E. (1993). "Frequency Analysis of Extreme Events." in *Handbook of Hydrology*, Maidment, D.(eds). McGraw-Hill, New York, Chapter 18.

Thiemann, M., Trosset, M., Gupta, H.V., and Sorooshian, S. (2001). "Bayesian recursive parameter estimation for hydrologic models." *Water Resources Research*, Vol. 37, No. 10, pp. 2521-2535.

Vicens, G.J., Rodriguez-Iturbe, I., and Schaake Jr, J.C. (1975). "A Bayesian framework for the use of regional information in hydrology." *Water Resources Research*, Vol. 11, No. 3, pp. 405-414.

Vrugt, J.A., Gupta, H.V., Bouten, W., and Sorooshian, S. (2003). "Shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters." *Water Resources Research*, Vol. 39, No. 8, SWC 1-16.

Wang, Q.J. (2001). "A Bayesian joint probability approach for flood record augmentation." *Water Resources Research*, Vol. 37, No. 6, pp. 1707-1712.

Whitley, R.J. and Hromadka II, T.V. (1999). "Approximate confidence intervals for design floods for a single site using a neural network." *Journal of Hydrology*, Vol. 153, pp. 265-290.

Wood, E.F., and Rodriguez-Iturbe, I. (1975a). "Bayesian inference and decision making for

extreme hydrologic events." *Water Resources Research*, Vol. 11, No. 4, pp. 533-542.

Wood, E.F., and Rodriguez-Iturbe, I. (1975b). A Bayesian approach to analyze uncertainty among flood frequency models. *Water Resources Research*, Vol. 11, No. 6, pp. 839-843.

Zhang, B., and Govindaraju, R.S. (2000). "Prediction of watershed runoff using Bayesian concepts and modular neural networks." *Water Resources Research*, Vol. 3, pp. 753-762.

(논문번호:07-113/접수:2007.10.23/심사완료:2008.02.14)