BH 베이지안 분석을 통한 서울지점 강우자료의 확률적 변화시점 추정

A Probabilistic Estimation of Changing Points of Seoul Rainfall using BH Bayesian Analysis

황석환*, 김중훈, 유철상***, 정성원****, 김민석*****

Hwang, Seok Hwan, Kim, Joong Hoon, Yoo, Chulsang, Jung, Sung Won, Kim, Min Seok

요 지

본 논문에서는 각각의 시점에서의 변화확률을 산정하여 변화시점을 추정하는 Barry와 Hartigan(BH)의 베이지안 변화시점 추정 방법(Bayesian changing points estimation method)을 이용하여 측우기 관측자료계열(CWK)과 근대우량계 관측자료계열(MRG)간의 변화에 대한 상대확률적 절점의 발생여부를 분석하였다. 각 강우특성별로 상대확률적인 변화시점 분석을 통하여 CWK와 MRG 간의 동질성 분석을 실시하였다. 분석 결과, CWK의 정성적인(본질적인) 통계적 특성은 MRG와 큰 차이가 없어 보인다. 다만, 관측정밀도의 한계로 인한 정량적인 차이가 존재하는 것으로 판단되었다.

핵심용어 : 베이지안, 기후변화, 측우기

1. 서론

본 논문에서는 각각의 시점에서의 변화확률을 산정하여 변화시점을 추정하는 베이지안 변화시점 추정 방법을 이용하여 측우기 자료와 근대우량계 자료사이의 변화에 대한 상대확률적 절점의 발생여부를 분석하여 두 자료계열간의 동질성 분석을 실시하였다. 변화시점을 파악하기 위해 일반적으로 빈번히 사용되는 방법들은 변화시점의 특정위치를 추정하는 형식이다. 그러나 베이지안 과정은 연속시계열에서 확률분포(연속시계열의 각각의 위치에서 변화확률을 산정)를 산정한다. 연속시계열에서의 변화시점은 변화전후 상태에 따라 결정이 되는 상대적인 개념으로 확정론적 방법으로 변화시점을 결정하는 것은 매우 어렵다. 이러한 관점에서 상대적인 변화시점의 확률을 계산하여 변화시점을 결정하는 베이지안 과정은 변화시점을 보다 합리적으로 추정할 수 있는 방법이다. 측우기 자료와 근대 강우량 자료는 연속 시계열자료로서 장기간의 시간적 차이를 보이기 때문에 경년변동을 고려하기 힘든 평균이나 표준편차와 같은 정량적인 기본통계특성의 크기 비교만으로는 정확히 동질성을 판단하는 것이 매우 힘들다. 이렇듯 정량적인 통계치에 의한 동질성 검정방법이 연속적인 경년변화(Trend)를 고려하지 못하는 한계를 가지고 있지만, 본 분석의 목적은 상대적으로 관측기록에 불신을 받고 있는 측우기 관측 강우량자료에 대한 정량적 통계치의 비교를 통해 신뢰수준을 가늠해 보는데 있기 때문에 이러한 분석을 통하여 측우기 관측 강우량자료의 기초 통계량에 대한 신뢰도가 확보된다면 보다 다양한 통계분석이나 추계학적 분석을 수행하는 데 보다 도움이 될것이다. 기존의 통계적인 유의수준에서의 측우기 관측 강우량 자료계열과 근대강우량 관측 자료계열간의 동질성 분석은 해당 강우량 계열의 모집단 분포형에 대한 정확한 추정이 수반되어야만 가설검정에 대한 신뢰가 가능하고 동질성을 판단하는 유의수준의 선택에 있어 객관적인 기준을 부여하기 어려워 결과에 대한 절대적인 신뢰도를 부여하기 어렵다는 단점이 있다. 따라서 본 연구에서는 베이지안 분석을 이용한 변화시점 추정기법을 이용하여 측우기 관측 강우량 자료계열과 근대우량계 강우량 관측계열간의 강우특성별로 변화시점이 발생하는 가를 분석해 이를 토대로 두 자료계열의 동질성에 대한 판단 근거로 이용하고자 한다. 수자원분야에서 베이지안 방법이 적용된 예는, 초기(1970년대)에는 주로 수자원 정책 결정을 위해 필요한 불확실도를 가늠해 보는데 이용이 되었고, 이러한 연구로는 Vicens 등(1975)과 Wood와 Rodriguez-Iturbe(1975)의 결과가 있다. 그 후 1990년 이후

* 한국건설기술연구원 수자원연구실 연구원
Researcher, Korea Institute of Construction Technology, GyeongGi-Do 411-712, Korea
(e-mail: sukany@kict.re.kr)

** 고려대학교 공과대학 건축·사회환경공학부 교수
Prof., School of Architecture, Civil & Environmental Eng., Korea University, Seoul, 136-713, Korea
(e-mail: jaykim@korea.ac.kr)

*** 고려대학교 공과대학 건축·사회환경공학부 교수
Prof., School of Architecture, Civil & Environmental Eng., Korea University, Seoul, 136-713, Korea
(e-mail: envchul@korea.ac.kr)

**** 한국건설기술연구원 수자원연구실 책임연구원
Chief Researcher, Korea Institute of Construction Technology, GyeongGi-Do 411-712, Korea
(e-mail: swjung@kict.re.kr)

*****고려대학교 공과대학 건축·사회환경공학부 석사과정
Master Student, School of Architecture, Civil, and Environmental Eng., Korea University, Seoul, 136-713, Korea
(e-mail: stynerz@naver.com)

컴퓨터의 비약적인 발전으로 인하여 방대한 양의 자료처리가 가능해짐으로써 홍수빈도 분석과 같은 분야에 적용되기 시작하였다. 이러한 연구로는 Madsen과 Rosbjerg(1997), Kuczera(1999), O'Connell(2002), Reis Jr. 등(2005), Reis Jr. 와 Stedinger(2005), Seidou 등(2006)이 있다. 국내의 경우는 최근에 김상욱과 이길성이 저수량 점 빈도해석(2008a, 2008b, 2008c)을 수행한 바 있고 수위-유량관계곡선의 불확실성을 분석(2008c)한 바 있다. 이러한 연구들은 주로 점빈도해석이나 지역빈도해석에 있어 자료에 기반한 사전분포를 구축하고 모수와 모수의 불확실성을 추정하기 위해 베이지안 기법을 적용하였다. Perreault 등(2000)은 복잡한 통계모형에 대한 베이지안 추론을 위해 많이 사용하는 깁스(Gibbs) 표본추출 방법을 이용하여 대형 수력발전 시스템에 대해 연수력발전량의 변화시점과 변화의 강도를 추정하기 위한 분석을 실시한 바 있다. 그러나 각각의 시점에서의 변화확률을 산정하여 변화시점을 추정하기위한 상대확률적 베이지안 변화시점 추정방법인 Barry와 Hartigan(BH, 1992)의 베이지안 방법은 수자원 분야에서 적용된바 없다.

2. 월별 동질성 분석을 위한 통계치의 설정

강우는 강우량과 같은 양적 특성과 발생횟수와 같은 빈도 특성 그리고 지속시간에 따른 강도 특성을 복합적으로 가지고 있기 때문에 강우의 변화를 파악하기 위해서는 양적 통계치는 물론 빈도와 강도를 적절히 표현할 수 있는 통계치를 동시에 비교해야만 강우특성의 변화여부에 대한 적절한 판단이 가능하다. *Monthly rainfall*은 각 월의 월강우량이고 $D_{\max}$ *ratio*는 월강우량에 대한 해당월 일최대 강우량의 평균적인 비율의 변화를 분석하기 위한 통계치이다. $N_{rainy\ days}$는 각 월별 평균적인 강우일수를 산정한 통계치이고 $I_{rainy\ days}$는 각 월별 강우일수에 대한 월강우량의 비로 강우일에 대한 평균강우강도이다. *Monthly rainfall*을 강우량의 정량적인 변화를 파악하기 위한 기본 통계치이고 $D_{\max}$ *ratio*는 일최대 강우량의 정량적 변화를 분석하여 강우의 규모변화를 파악할 수 있기 때문에 선택하였다. $N_{rainy\ days}$는 강우일수의 변화를 정량적으로 나타내므로 이를 분석하면 강우발생 경향의 변화 여부를 쉽게 파악할 수 있고 $I_{rainy\ days}$는 강우강도의 변화를 정량적으로 분석 가능하기 때문에 사용하였다.

3. 기본이론

본 논문에서는 Barry 와 Hartigan (1993)이 제시한 PPM(product partition model)에 근거한 베이지안 변화시점 분석방법을 이용하였다. 베이지안 방법과 대응하는 대표적인 방법들로는 CBS(circular binary segmentation, Olshen and Venkatraman, 2004)와 BP(breakpoints, Bai and Perron, 2003) 기법 등이 있다. Bai 와 Perron의 BP 방법은 다양한 조각들(segments)로 구성된 최적의 분리구간(partitions)을 결정하기 위해 동적 프로그래밍 알고리즘을 이용한다. Olshen 과 Venkatraman의 CBS 방법은 BS(binary segmentation) 방법의 개선된 형태로 BS가 단일 변화시점 검정에 기반하고 있어 큰 변화구간 중심에 묻혀버린 작은 변화구간을 찾는데는 문제가 발생할 수 있다. 이러한 문제를 보완한 것이 CBS 방법이다. 그러나 BP와 달리 CBS는 권장된 변화시점 수에 대한 최적의 변화시점의 위치를 찾는데 어려움이 있다. CBS와 마찬가지로, Barry 와 Hartigan의 PPM은 관측치들이, 독립적인 위치 i에서의 평균이 μ_i이고 분산이 σ^2일 때, $N(\mu_i,\ \sigma^2)$으로 독립적이고 각각이 독립적인 위치 i에서의 변화시점의 확률을 p로 가정하였다. 이 알고리즘은 분리구간(partition) $\rho=(U_1, U_2, \ldots, U_n)$을 사용한다. 여기서 $U_i=1$은 $i+1$위치가 변화시점임을 의미한다; n보다 작은 모든 $i(i<n)$에 대하여 U_i는 0으로, $U_n \equiv 1$로 초기화한다. $i+1$위치에서 변화시점의 조건부 확률에 대한 전이확률 p는 Barry 와 Hartigan에 제시된 간편 비율법(simplified ratio)으로 구할수 있다.

$$\frac{p_i}{1-p_i}=\frac{P(U_i=1|\mathbf{X},\ U_j, j\neq i)}{P(U_i=0|\mathbf{X},\ U_j, j\neq i)} = \frac{\left[\int_0^{\gamma} p^b(1-p)^{n-b-1}dp\right]\left[\int_0^{\lambda}\frac{w^{b/2}}{(W_1+B_1w)^{(n-1)/2}}dw\right]}{\left[\int_0^{\gamma} p^{b-1}(1-p)^{n-b}dp\right]\left[\int_0^{\lambda}\frac{w^{(b-1)/2}}{(W_0+B_0w)^{(n-1)/2}}dw\right]} \quad (1)$$

여기서 W_0, B_0, W_1과 B_1은 각각 $U_i=0$과 $U_i=1$일 때 얻어지는 블록내(within-block)와 블록간(between-block)의 제곱합이다. 그리고 $\mathbf{X}$는 자료계열이다. 조절(tuning) 매개변수 γ와 λ는 $[0,1]$에서 선택하게 되고 각각의 반복과정을 통하여 사후평균은 현재의 분리구간 상에서 조건부로 갱신된다. 그러나 다음의 두 적분항 때문에 긴 연속자료에 대한 BH MCMC(Markov Chain Monte Carlo) 알고리즘의 직접 적용시 발산하거나 0으로 수렴하여 수치적으로 불안정할 수 있다.

$$\left[\int_0^{\lambda}\frac{w^{b/2}}{(W_1+B_1w)^{(n-1)/2}}dw\right], \qquad \left[\int_0^{\lambda}\frac{w^{(b-1)/2}}{(W_0+B_0w)^{(n-1)/2}}dw\right] \quad (2)$$

다행히, 이 적분항은 불완전 베타 적분법(incomplete beta integrals)에 의해 간략화될 수 있다. 분리구간의 특정 위치(주어진 자료와 현재 분리구간)에서의 변화시점에 대한 상대확률(odds)은 다음과 같이 다시 표현할 수 있다.

$$\frac{p_i}{1-p_i} = \frac{P(U_i=1|\mathbf{X},\, U_j, j\neq i)}{P(U_i=0|\mathbf{X},\, U_j, j\neq i)}$$
$$= \left(\frac{W_0}{W_1}\right)^{\frac{n-b-2}{2}} \cdot \left(\frac{B_0}{B_1}\right)^{\frac{b+1}{2}} \cdot \sqrt{\frac{W_1}{B_1}} \qquad (3)$$
$$\cdot \frac{\int_0^{\frac{B_1\lambda/W_1}{1+B_1\lambda/W_1}} p^{(b+2)/2}(1-p)^{(n-b-3)/2}dp}{\int_0^{\frac{B_0\lambda/W_0}{1+B_0\lambda/W_0}} p^{(b+1)/2}(1-p)^{(n-b-2)/2}dp}$$
$$\cdot \frac{\int_0^{\gamma} p^{b}(1-p)^{n-b-1}dp}{\int_0^{\gamma} p^{b-1}(1-p)^{n-b}dp}$$

이 수식은 자료계열의 길이에 상관없이 BH 과정이 적용될 수 있도록 수치적으로 안정한 항들로 구성되어 있다. 이를 이용하여 BH의 MCMC 수행을 통해 변화시점과 평균(μ_{ij})의 사후분포를 추정한다.

4.분석결과

4.1 월별 변화시점 분석결과

6월의 경우는 그림1과 같다. *Monthly rainfall*의 경우 변화에 대한 사후확률이 높은 시점이 여럿존재하나 변화시점전후로 지속성이 없어(변동시점) 사후평균이나 사후확률에서 명확한 변화시점을 찾기는 힘들다. $D_{\max}$ *ratio*는 전체적으로 변화가 없음을 알 수 있다. $N_{rainy\ days}$의 경우는 4월이나 5월과 마찬가지로 M00의 경우 1908년을 전후로 뚜렷한 변화를 보이고 있다. 그리고 M20을 기준으로 볼 때 6월의 $N_{rainy\ days}$(사후평균의 경우)는 예전이나 근대이후 크게 차이를 보이지 않으나 약간 증가한 것을 볼 수 있다. $I_{rainy\ days}$는 명확한 변화시점을 찾기 어렵다. 7월의 경우는 그림2와 같다. *Monthly rainfall*의 경우 M00과 M20 모두 1908년을 전후로 뚜렷한 변화시점을 보이고 있지 않다. 이는, 7월의 경우, 상대적으로 강우량이나 강우빈도가 커서 2mm이하의 강우는 월강우량의 차이에 거의 영향을 미치지 못함을 의미한다. 더불어 강우량의 양적차이도 예전이나 근대이후 별 차이를 보이지 않고 있다. 단, 1960년을 기준으로 사후확률이 크게 나타나는 점은 주목할 만한 결과이다. $D_{\max}$ *ratio*는 변화양상을 보이지 않고 있다. $N_{rainy\ days}$는 M00의 경우는 1908년을 전후로 변화양상을 보이나 M20의 경우는 변화양상을 보이고 있지 않다. 따라서 7월의 경우 예전과 근대이후 $N_{rainy\ days}$에 변화가 있다고 보기는 힘들다. $I_{rainy\ days}$는 뚜렷한 변화시점을 찾기 힘들다. 8월의 경우는 그림3와 같다. *Monthly rainfall*의 경우 7월과 마찬가지로 M00과 M20 모두 1908년을 전후로 뚜렷한 변화시점을 보이고 있지 않다. 8월의 경우도 상대적으로 강우량이나 빈도가 커서 2mm이하의 강우는 월강우량의 차이에 거의 영향을 미치지 못함을 의미한다. $D_{\max}$ *ratio*도 변화시점을 보이지 않는다. $N_{rainy\ days}$는 M00의 경우 1900년 초반을 전후로 변화양상을 보이나 M20의 경우는 변화시점을 보이지 않는다. 따라서 8월의 $N_{rainy\ days}$의 변화는 과거와 근대이후에 차이가 있다고 보기 힘들고 M00에서의 차이는 관측의 정밀도 차이에서 기인한 것으로 판단된다. $I_{rainy\ days}$도 1908년을 전후로 뚜렷한 변화시점을 보이지 않는다. 단, M00의 사후평균이 근대이후 평균적으로 약간 감소한 경향을 보이나 이는 측우기 기록에서 2mm 이하 강우의 누락으로 인한 강우일수의 감소에 따라 측우기 기록의 $I_{rainy\ days}$가 상대적으로 크게 산정된 결과로 해석할 수 있다. 9월의 경우는 그림4과 같다. *Monthly rainfall*의 경우 M00과 M20 모두 변화시점을 보이지 않는다. $D_{\max}$ *ratio*도 변화시점을 보이지 않는다. $N_{rainy\ days}$의 경우 M00의 경우는 1908년을 전후로 변화양상을 보이나 M20의 경우는 변화양상을 보이지 않는다. 이 또한 측우기 관측 정밀도에 의한 차이로 설명이 가능하고 9월의 경우 $N_{rainy\ days}$의 변화는 특별히 포착되지 않는다. $I_{rainy\ days}$는 M00의 경우 1900년 전후로 변화양상(감소)을 보이나 이는 앞서 언급한 바와 같이 관측정밀도 차이에 의한 월강우일수의 차이에서 기인한 결과로 볼 수 있다. 이는 7월과 8월의 결과에서도 9월만큼 뚜렷하지는 않으나 확인할 수 있다. M20의 경우는 변화시점을 보이지 않고 있다.

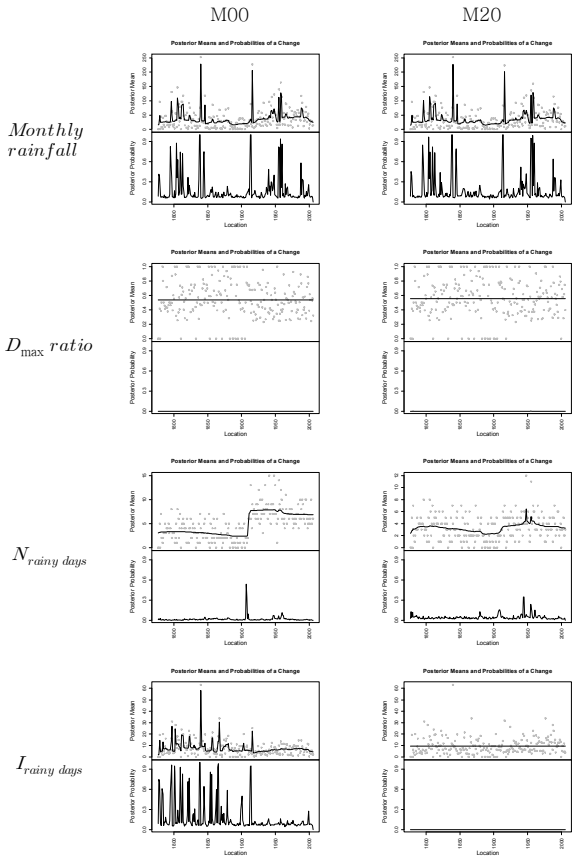

Fig. 1. Estimated Posterior Means and Posterior Probabilities for JUN

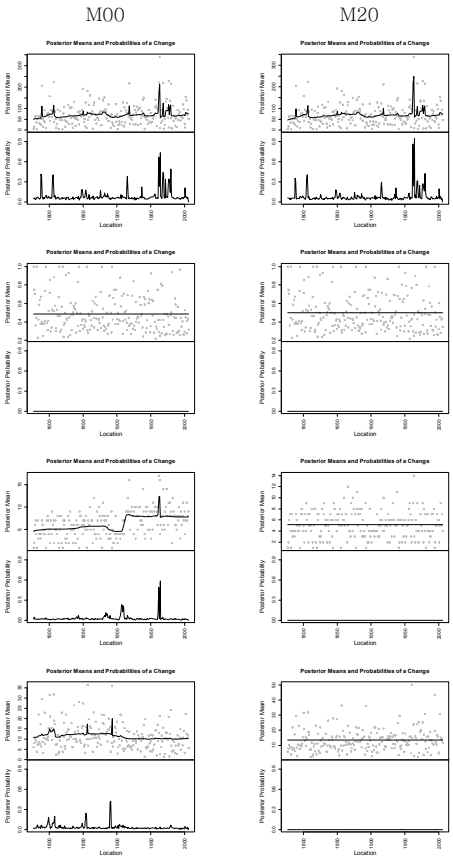

Fig. 2. Estimated Posterior Means and Posterior Probabilities for JUL

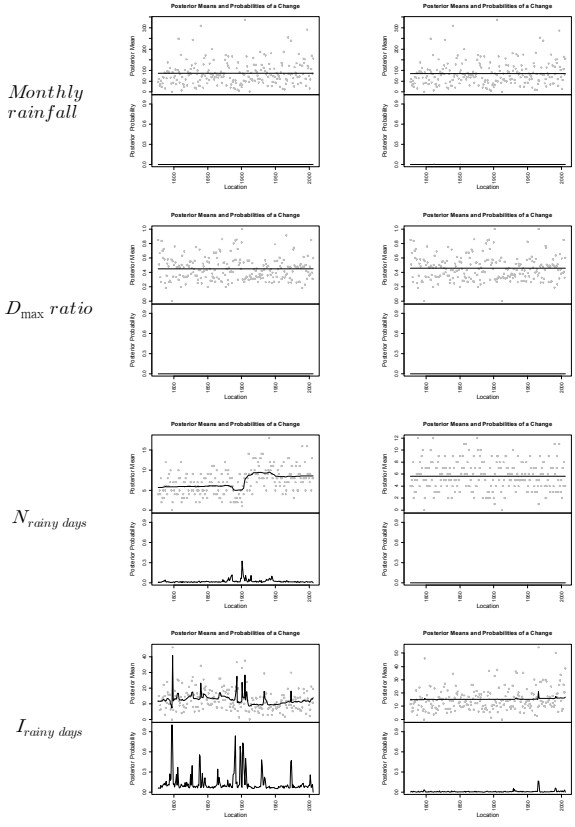

Fig. 3. Estimated Posterior Means and Posterior Probabilities for AUG

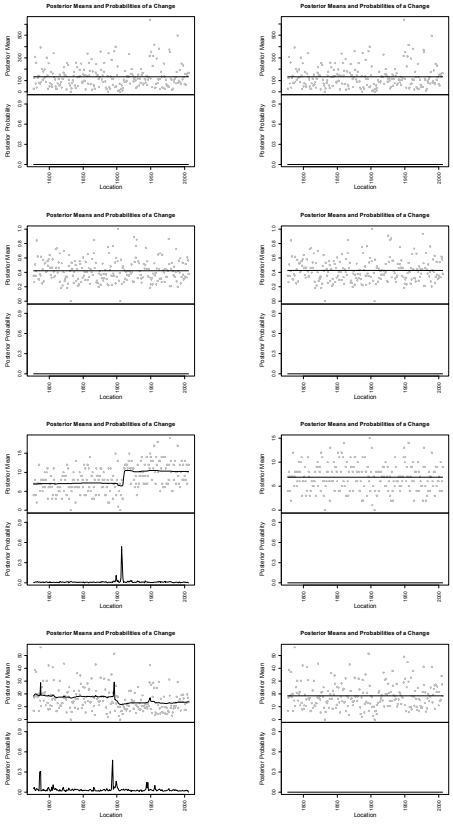

Fig. 4. Estimated Posterior Means and Posterior Probabilities for SEP

5.결론

본 논문에서는 각각의 시점에서의 변화확률을 산정하여 변화시점을 추정하는 베이지안 변화시점 추정 방법(Bayesian changing points estimation method)을 이용하여 측우기 관측자료계열(CWK, 1777년-1907년)과 근대우량계 관측자료계열(MRG, 1908년-2006년) 사이의 변화에 대한 상대확률적 절점의 발생여부도 분석하였다. 강설이 포함된 동절기를 제외한 4월에서 10월 까지 월별로 변동성 분석을 실시한 결과는 다음과 같다. *Monthly rainfall*은, M00과 M20 모두, 모든 달에서 두자료계열의 경계(1907년과 1908년)를 전후로 지속성 있는 뚜렷한 변화시점은 나타나지 않았다. $D_{\max}$ *ratio*는, 6월, 7월, 8월, 9월에서 M00과 M20 모두 두자료계열의 경계(1907년과 1908년)를 전후로 지속성 있는 뚜렷한 변화시점이 나타나지 않았다. $N_{rainy\ days}$는, M00의 경우 모든 달의 두자료계열의 경계에서 지속성 있는 뚜렷한 변화시점이 나타났다. 특히 6월은 7월, 8월, 9월에 비해 보다 명확한 변화양상을 보였다. M20의 경우 6월은 두자료계열의 경계에서 약한 변화양상을 보였다. 특히 6월의 경우는 상대적으로 변화시점을 전후로 지속성이 있었고 1900년을 전후로 사후평균이 감소한 경향을 보였다. 그 외 7월에서 9월까지는 두자료계열의 경계에서 변화시점이 나타나지 않았다. $I_{rainy\ days}$는, M00과 M20 모두 6월은 두자료계열의 경계에서 변화시점이 나타나지 않았다. M00의 7월, 8월, 9월은 두자료계열의 경계에서 약한 변화양상이 나타나고 있고 변화시점 전후로 지속성도 있었다. 특히 9월은 상대적으로 뚜렷한 변화시점을 보였다. 그러나 M20의 7월, 8월, 9월, 두자료계열의 경계에서 변화시점이 나타나지 않았다. 이러한 결과로부터 CWK의 정성적인(본질적인) 통계적 특성은 MRG와 큰 차이가 없어 보인다. 단, 관측정밀도 한계와 관측방식의 차이에서 기인한 강우특성의 정량적 차이가 일부 존재한다고 판단된다.

6.참고문헌

김상욱, 이길성 (2008). "Bayesian MCMC를 이용한 저수량 점 빈도해석: Ⅰ. 이론적 배경과 사전분포의 구축." **한국수자원학회논문집**, 제41권, 제1호, pp. 35-47.

김상욱, 이길성 (2008). "Bayesian MCMC를 이용한 저수량 점 빈도해석: Ⅱ. 적용과 비교분석." **한국수자원학회논문집**, 제41권, 제1호, pp. 49-63.

김상욱, 이길성 (2008). "Bayesian 다중회귀분석을 이용한 저수량(Low flow) 지역 빈도분석." **한국수자원학회논문집**, 제41권, 제3호, pp. 325-340.

김상욱, 이길성 (2008). "베이지안 회귀분석을 이용한 수위-유량관계곡선의 불확실성 분석." **한국수자원학회논문집**, 제41권, 제9호, pp. 943-958.

Barry, D., and Hartigan, J.A. (1992). "PRODUCT PARTITION MODELS FOR CHANGE POINT PROBLEMS." *The annals of Statistics*, Vol. 20, No. 1, pp. 260-279.

Barry, D., and Hartigan, J.A. (1993). "A Bayesian Analysis for Change Point Problems." *Journal of the American Statistical Association*, Vol. 88, No. 421, pp. 309-319.

Erdman, C., and Emerson, J.W. (2007). "An R Package for Performing a Bayesian Analysis of Change Point Problem." *Journal of Statistical Software*, Vol. 23, Issue 3, pp. 1-13.

Erdman C., and Emerson J.W. (2007). "bcp: A Package for Performing a Bayesian Analysis of Change Point Problems." R package version 1.8.4, URL http://CRAN.R-project.org/.

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

Madsen, H., and Rojsberg, 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.

O'Connel, D.R.H., Ostenaa, D.A., Levish, D.R., and Klinger, R.E. (2002). "Bayesian flood frequency analysis with paleohydrologic bound data." *Water Resources Research*, Vol. 38, Issue 5, 1058.

Perreault, L., Bernier, J., Bobèe, B., and Parent, E. (2000). "Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited." *Journal of Hydrology*, Vol. 235, pp. 221-241.

Perreault, L., Bernier, J., Bobèe, B., and Parent, E. (2000). "Bayesian change-point analysis in hydrometeorological time series. Part 2. Comparison of change-point models and forecasting." *Journal of Hydrology*, Vol. 235, pp. 242-263.

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 long 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." *Wate Resources Research*, Vol. 42, W11408.

Vicens, G.J., Rodriguez-Itrube, I., and Schaake Jr., J.C. (1975). "A Bayesian Framework for the Use of Regional Information in Hydrology." *Water Reources Research*, Vol. 11, No. 3, pp. 405-414.

Wood, E.F., and Rodriguez-Itrube, 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-Itrube, I. (1975b). "Bayesian Approach to Analyze Uncertainty Among Flood Frequency Models." *Water Resources Research*, Vol. 11, No. 6, pp. 839-843.