

여러 차수의 확률 가중 모멘트를 이용한 적정 분포함수의 매개변수 추정 Using Various Order Probability Weighted Moments for the Parameter Estimation of Appropriate Distribution Functions

이길성*, 김지영**
Kil Seong Lee, Ji Young Kim

요 지

댐과 같은 구조물의 설계시 큰 강우량에 대한 분포함수의 적합성을 높일 필요가 있다. 이에 대해 Wang (1997a and b)은 큰 설계량에 대한 적합성을 높이기 위해 LH 모멘트와 고차 PWM(higher Probability Weighted Moments)방법을 제안하였다. 따라서 본 연구에서는 우리나라의 각 지역별로 대표적인 4개 지점의 일 강우량 자료를 사용하여 제안된 고차 PWM 방법의 적용성을 살펴보았다. 그 과정으로 가장 낮은 차수인 일반적인 PWM 방법과 더 높은 차수의 PWM 방법을 이용하여, GEV(Generalized Extreme Value) 분포와 Gumbel 분포에 대한 매개변수를 추정한 후 이 추정치를 확률지에 실측치와 함께 도시하여 결과를 비교하였다. 그리고 PPCC(Probability Plot Correlation Coefficient) 적합도 검정결과를 통해 추정된 매개변수의 적합성을 확인하였다.

핵심용어 : PWM 방법, 매개변수 추정, GEV 분포, Gumbel 분포, PPCC 적합도 검정

1. 서론

강우 빈도해석에서 지속적으로 요구되는 것은 적정 분포함수의 선정과 분포함수의 매개변수 추정이며, 매개변수를 추정하는 방법으로 모멘트법(method of moments), 최우도법(method of maximum likelihood)을 비롯하여, 확률 가중 모멘트(probability weighted moments, 이하 PWM) 방법과 L 모멘트법 등이 많이 사용되고 있다. 그러나 이러한 방법들을 통해 매개변수를 추정할 때 비교적 큰 강우자료에 대해 정확히 계산되지 못하면 긴 재현기간의 설계량 계산시 심각한 오차가 발생할 수 있다. 따라서, 댐과 같은 구조물의 설계시 큰 강우량에 대한 분포함수의 적합성을 높일 필요가 있다. 이에 대해 Wang(1997a and b)은 큰 설계량에 대한 적합성을 높이기 위해 LH 모멘트(higher L moments)와 고차 PWM(higher PWM) 방법을 제안하였다. 따라서 본 연구에서는 우리나라의 각 지역별로 대표적인 4개 지점의 연 최대 강우량 자료를 사용하여 제안된 방법의 적용성을 살펴보려 한다. 그 과정으로 가장 낮은 차수인 일반적인 PWM 방법과 더 높은 차수의 PWM 방법을 이용하여, 김정덕(1996)등에 의해 우리나라 연 최대 강우량 자료에 가장 적절하다고 연구된바 있는 GEV(Generalized Extreme Value) 분포와 그 외 Gumbel 분포에 대한 매개변수를 추정한 후 이 추정치를 확률지에 실측치와 함께 도시하여 결과를 비교하며, LH 모멘트 방법에 의한 결과와도 비교한다. 그리고 PPCC(Probability Plot Correlation Coefficient) 적합도 검정 결과를 통해 두 가지 분포에 대하여 적합성을 확인한다.

2. 이론적 배경

2.1 PWM

PWM은 다음과 같이 정의된다.

$$\beta_r = \int_0^1 x(F)^{r-1} dF \quad (1)$$

* 정회원 · 서울대학교 지구환경시스템공학부 교수 · E-mail : kilselee@snu.ac.kr

** 정회원 · 서울대학교 지구환경시스템공학부 석사과정 · E-mail : zziyoi@snu.ac.kr

여기서, r 은 1, 2, 3, ... 으로서 PWM의 차수가 되며, F 는 변수 x 의 CDF(Cumulative Distribution Function)이다. 식(1)에 정의된 PWM은 차수를 높임으로써 큰 x 에 대해 가중치를 줄 수 있다. 확률가중치 F^r 은 r 이 커질 때 더 높은 비율로 x 와 함께 증가하기 때문에 큰 x 에 더 많은 가중치를 주게 된다. 즉, 고차 PWM의 값은 변수의 큰 범위에서 더욱 많은 영향을 받는다는 것이다. 그리고 오름차순으로 정렬된 표본, $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 이 주어졌을 때 β_r 의 불편 추정치(unbiased estimator)는 다음과 같이 주어진다.

$$b_r = \frac{1}{n} \sum_{i=1}^n \binom{i-1}{r} x_i / \binom{n-1}{r} = \frac{1}{n} \sum_{i=1}^n \frac{(i-1)(i-2)\dots(i-r)}{(n-1)(n-2)\dots(n-r)} x_{(i)} \quad (2)$$

2.2 GEV 분포 및 Gumbel 분포 함수로의 적합

GEV 분포에서는 미지의 매개변수 ξ , α , k 가 있으므로 이 매개변수를 구하기 위해서는 세 개의 식이 필요하다. 따라서 PWM의 차수를 $r = \eta, \eta+1, \eta+2$ 로 두고 GEV 분포함수에 적용하면 $k \neq 0$ 일 때 다음과 같이 표현된다(Wang, 1997b).

$$(\eta+1)\beta_\eta = \xi + \frac{\alpha}{k} \left[1 - \frac{\Gamma(1+k)}{(\eta+1)^k} \right] \quad (3)$$

$$(\eta+2)\beta_{\eta+1} - (\eta+1)\beta_\eta = \alpha \frac{\Gamma(1+k)}{k} [(\eta+1)^{-k} - (\eta+2)^{-k}] \quad (4)$$

$$\frac{(\eta+2)\beta_{\eta+1} - (\eta+1)\beta_\eta}{(\eta+3)\beta_{\eta+2} - (\eta+1)\beta_\eta} = \frac{(\eta+1)^{-k} - (\eta+2)^{-k}}{(\eta+1)^{-k} - (\eta+3)^{-k}} \quad (5)$$

또한, $k=0$ 일 때도 마찬가지로 다음과 같이 나타낼 수 있으며 이는 Gumbel 분포에 대한 식이다.

$$(\eta+1)\beta_\eta = \xi + \alpha[\varepsilon + \ln(\eta+1)] \quad (6)$$

$$(\eta+2)\beta_{\eta+1} - (\eta+1)\beta_\eta = \alpha[\ln(\eta+2) - \ln(\eta+1)] \quad (7)$$

여기서, $\Gamma()$ 는 Gamma 함수이고, $\varepsilon \approx 0.5772$ 로서 Euler 상수이다. 그리고 $r = \{0, 1, 2\}, \{1, 2, 3\}, \dots$ 식으로 차수를 높여감으로써 여러 차수의 PWM을 적용할 수 있다. $k \neq 0$ 인 GEV 분포에 대한 매개변수 추정시 식 (5)에서 k 의 값을 양해적으로 구할 수 없기 때문에 반복법을 통해 수치적으로 구해야 한다. 이 과정을 다음과 같은 근사식으로 대체할 수 있다.

$$k = e_1 c + e_2 c^2 \quad (8)$$

여기서, e_1, e_2 는 상수이며, c 는 다음과 같다.

$$c = \frac{(\eta+2)\beta_{\eta+1} - (\eta+1)\beta_\eta}{(\eta+3)\beta_{\eta+2} - (\eta+1)\beta_\eta} \frac{\ln(\eta+2) - \ln(\eta+1)}{\ln(\eta+3) - \ln(\eta+1)} \quad (9)$$

식 (9)에 β_r 대신 식 (2)의 추정치 b_r 을 적용하여 식 (8)로부터 \hat{k} 의 값을 구한 후 식 (3)과 (4)로부터 나머지 매개변수를 다음과 같이 추정할 수 있다. 이때 β_r 은 역시 b_r 을 사용한다.

$$\hat{\alpha} = \frac{(\eta+2)b_{\eta+1} - (\eta+1)b_\eta}{(\eta+1)^{-\hat{k}} - (\eta+2)^{-\hat{k}}} \frac{\hat{k}}{\Gamma(1+\hat{k})} \quad (10)$$

$$\hat{\xi} = (\eta+1)b_\eta - \frac{\hat{\alpha}}{\hat{k}} \left[1 - \frac{\Gamma(1+\hat{k})}{(\eta+1)^{\hat{k}}} \right] \quad (11)$$

$k=0$ 인 Gumbel 분포의 경우에는 식 (6)과 (7)로부터 다음과 같이 매개변수를 추정할 수 있다.

$$\hat{\alpha} = \frac{(\eta+2)b_{\eta+1} - (\eta+1)b_\eta}{\ln(\eta+2) - \ln(\eta+1)} \quad (12)$$

$$\hat{\xi} = (\eta+1)b_\eta - \hat{\alpha}[\varepsilon + \ln(\eta+1)] \quad (13)$$

3. 대상 선정 및 적용

3.1 자료의 선정

우리나라의 강우 자료는 오랜 기간 동안 여러 관측소에서 관측하여 그 기록을 보유하고 있으며, 자료의 수준 또한 양질이어서 연구 대상으로 선정하였다. 우리나라의 각 도별로 대표적인 지점 중 관측기간이 50년 이상인 지역을 선정하였으며 그 대상은 표 1과 같다. 4개의 지점을 선정하여 일 강우량 자료를 사용하였으며 대부분 90년 내외의 긴 자료이다.

3.2 기본 통계분석

일련의 수문자료계열에서 자료의 일반적인 균형분포의 상태에서 훨씬 위나 또는 아래로 격리되는 outlier의 유발로 인해 부적절한 통계학적 매개변수의 유도로서 설계강우량의 제시에 불확실성을 초래하게 된다. 따라서 본 분석에서는 Grubbs-Beck(1972)에 의한 검정법(G-B test)을 사용하여 outlier의 유무를 검정하였다. 검정결과, 지점별로 high outlier 및 low outlier가 있음이 확인되었고 이 값들은 분석대상에서 제외하였으며, 그 결과는 표 2와 같다.

표 1. 연구 대상 지점

지 점	기록기간	기록년수
강 룡	1911-2002	92년
대 구	1907-2002	96년
인 천	1904-2002	99년
전 주	1918-2002	85년

표 2. 각 지점에서의 outlier (단위는 mm/day)

지 점	x_H	x_L	Outlier
강 룡	535.17	31.12	870.5
대 구	277.79	27.08	-
인 천	455.76	27.73	23.2
전 주	344.15	30.96	-

3.3 대상지점의 매개변수 추정 및 적합도 검정

앞서 제시된 식 (8)~(13)을 바탕으로 GEV 분포와 Gumbel 분포에 대한 매개변수를 추정하였으며 추정된 매개변수의 적합도를 검정하기 위하여 PPCC 방법을 사용하였다. 그리고 plotting position은 다음 식 (14)와 같은 Cunnane의 방법을 사용하였는데 본 연구에서 unbiased estimator를 사용하였기에 이 방법을 적용하였다.

$$q_i = \frac{i-0.4}{n+0.2} \quad (14)$$

여기서, i 는 오름차순 정렬시 각 자료의 순서이고, n 은 자료의 크기이다. 결과는 표 3과 같은데 그 값이 0~1 사이에 존재하며 1에 가까울수록 perfect correlation함을 나타낸다. 적합도 검정결과를 살펴보면 대부분 0.99 내외의 값을 나타내며 적합도가 우수하게 나타났다. GEV 분포의 경우 차수를 달리함으로써 더 좋은 적합도가 나타남을 알 수 있다. 그러나 고차 PWM 방법의 특징상 차수를 달리하면 오른쪽 꼬리 부분 즉, 큰 설계량에 대해 좋은 적합결과가 나타날 수 있으나 반면 반대쪽 꼬리 부분의 적합성이 그만큼 불량해지기 때문에 차수를 무조건 높인다고 좋아지는 것은 아니다. 이번 경우에는 대상지점 모두 차수 $r=2, 3, 4$ 일 때 가장 높은 적합도가 나타났다. 이는 분포 함수의 좌·우 꼬리 부분에 대해 가장 균형있게 잘 맞는 차수로 여겨진다.

표 3. 각 분포에 대한 적합도 검정 결과

지 점	PWM 차수	GEV	Gumbel	지 점	PWM 차수	GEV	Gumbel
강 룡	$r=0, 1, 2$	0.9837	0.9879	인 천	$r=0, 1, 2$	0.9755	0.9774
	$r=1, 2, 3$	0.9869			$r=1, 2, 3$	0.9788	
	$r=2, 3, 4$	0.9879			$r=2, 3, 4$	0.9819	
	$r=3, 4, 5$	0.9872			$r=3, 4, 5$	0.9811	
	$r=4, 5, 6$	0.9848			$r=4, 5, 6$	0.9757	
대 구	$r=0, 1, 2$	0.9918	0.9933	전 주	$r=0, 1, 2$	0.9879	0.9911
	$r=1, 2, 3$	0.9928			$r=1, 2, 3$	0.9912	
	$r=2, 3, 4$	0.9943			$r=2, 3, 4$	0.9916	
	$r=3, 4, 5$	0.9928			$r=3, 4, 5$	0.9902	
	$r=4, 5, 6$	0.9932			$r=4, 5, 6$	0.9866	

3.4 확률도시를 통한 실측치와의 비교

확률 분포 곡선은 계산된 quantile값을 reduced Gumbel variate에 대해 도시하였다. 이와 같이 reduced Gumbel variate에 대해 도시하게 되면 GEV 분포는 2차함수와 같은 곡선으로, Gumbel 분포의 경우는 1차함수와 같은 직선으로 나타나게 된다. 이와 같이 각 지점별로 도시한 결과는 그림 1~4와 같으며, 이는 실측치와 차수에 따른 추정치를 같이 나타내었기 때문에 비교·분석이 용이하다.

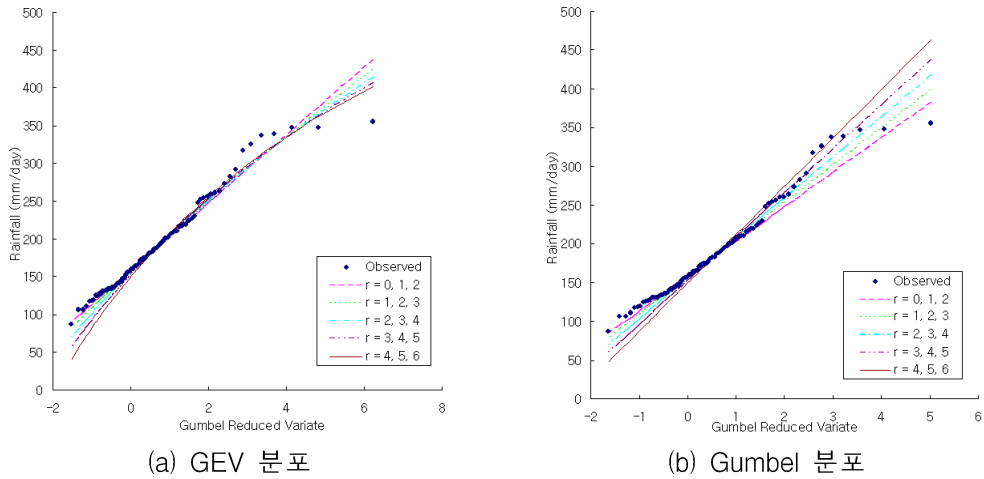


그림 1. 강릉 지점에 대한 확률 분포 곡선

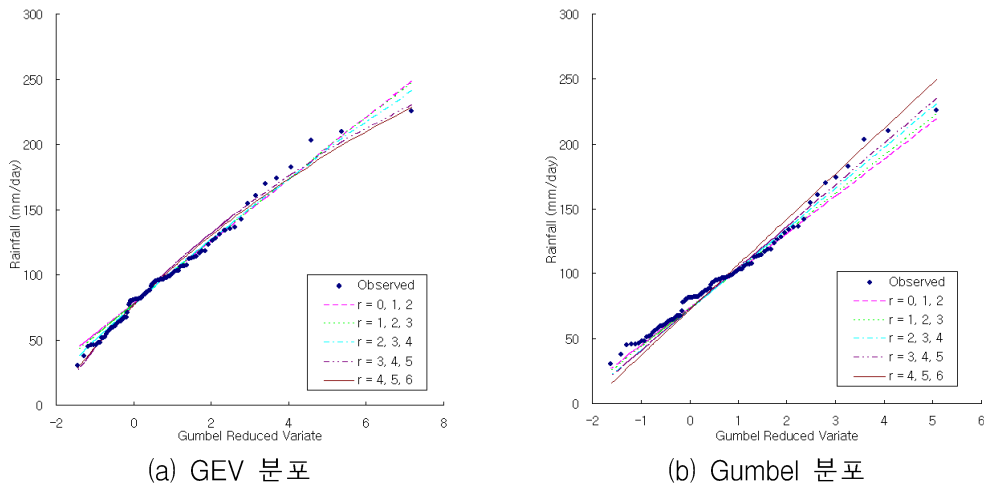


그림 2. 대구 지점에 대한 확률 분포 곡선

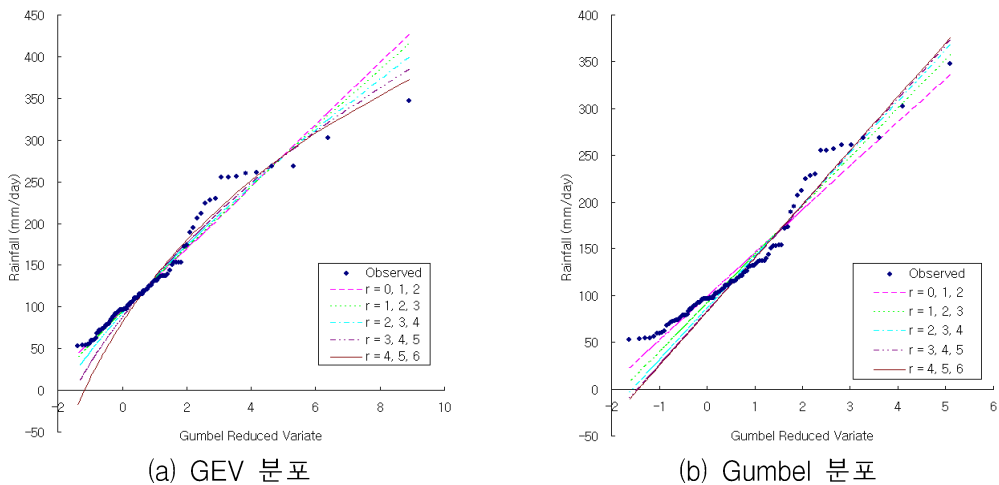


그림 3. 인천 지점에 대한 확률 분포 곡선

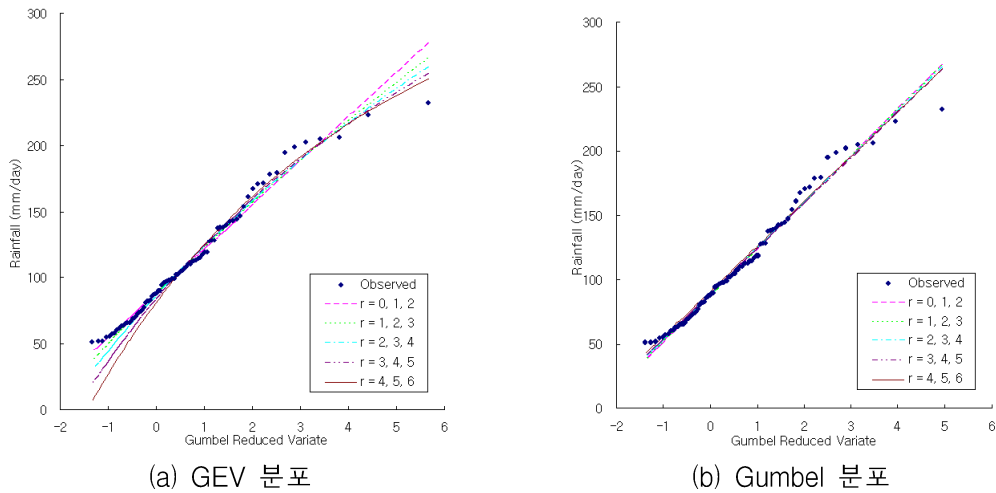


그림 4. 전주 지점에 대한 확률 분포 곡선

위의 그래프를 통해 GEV 분포와 Gumbel 분포에서 모두 차수를 달리 함으로써 적합성이 달라짐을 확인할 수 있다. 그러나 앞서 언급했듯이 큰 강우량에 대해 적합성이 향상되는 만큼 작은 강우량에 대해서는 오히려 불량해졌으며, 차수를 높일수록 큰 강우량에 대한 적합성이 계속 좋아지는 것은 아님을 확인할 수 있다. 따라서 적절한 차수를 찾아내면 최대의 효과를 낼 수 있을 것이다. 대부분의 지점에서 목표했던 결과를 잘 보여주었지만 Gumbel 분포의 경우는 GEV 분포의 경우에 비하여 차수를 높여도 큰 설계량에 대한 적합도가 나아짐을 보기 어렵다.

4. 결론 및 향후 연구과제

댐과 같은 구조물의 설계시 큰 설계량에 대한 적합성을 높일 필요가 있을 것이다. 고차 PWM 방법은 큰 설계량에 대한 적합성을 높이기 위해 제안되었다. 따라서 본 연구에서는 4개 지점의 연 최대 강우량 자료를 사용하여 PWM의 차수를 달리하며 적용성을 살펴보았다. 그 결과, 차수가 높아질수록 큰 설계량에 대한 적합도가 커짐을 알 수 있었다. 그러나 차수가 높아질수록 큰 강우량에 대한 적합도가 계속 커지는 것은 아니며 그만큼 작은 강우량에 대한 적합도가 불량해지기 때문에 이 방법이 모든 경우에 대해 우수하다고 할 수 없으며 단지 큰 설계량에 대한 적합성을 개선하는데 이용되어 질 수 있을 것이다. 또한 PPCC 검정을 통하여 추정된 매개변수의 적합성을 파악하였으며 이를 통해 전체 설계량에 대해 가장 잘 맞는 적절한 차수를 구할 수 있음을 알았다. 이에 대해 향후 차수를 결정하는데 있어 좀 더 효율적이고 합리적인 방법에 관한 연구가 수행되어야 할 것이다. 또한 GEV나 Gumbel 분포 외에 다른 분포형에 대해서도 그 분포형의 PWM만 유도 되면 여러 차수의 PWM을 적용할 수 있으므로, 그 적용성을 살펴볼 필요가 있겠다. 그리고 큰 설계량에 대한 적합성을 높이는데 그치지 말고 전체 설계량에 대한 적합성을 개선할 수 있는 방법에 대한 연구가 수행되어야 함은 당연할 것이다.

감사의 글

본 연구는 21세기 프론티어연구개발사업인 수자원의 지속적 확보기술개발사업단의 연구비지원(과제번호: 6-2-1)에 의해 수행되었습니다.

참고 문헌

1. 김경덕, 허준행, 조원철 (1996). "연 최대 강우자료의 적정 확률 분포형 선정에 관한 연구", **대한토목학회 논문집**, 대한토목학회, 제16권, 제4호, pp.335-344.
2. Grubbs, F., Beck, G. (1972). "Extension of sample sizes and percentage points for significance tests of outlying observations.", *Technometrics*, Vol.14, No.4, pp.847-854.
3. Wang, Q. J. (1997a). "LH moments for statistical analysis of extreme events.", *Water Resources Research*, Vol.33, No.12, pp.2841-2848.
4. Wang, Q. J. (1997b). "Using higher probability weighted moments for flood frequency analysis.", *Journal of Hydrology*, Vol.194, pp.95-106.