The Statistical Model for Predicting Flood Frequency

Noh, Jae Sik* / Lee, Kil Choon**

ABSTRACT/This study is to verify the applicability of statistical models in predicting flood frequency at the stage gaging stations of which the flow is under natural condition in the Han River basin. The results of the study show that the statistical flood frequency models were proven to be fairly reasonable to apply in practice, and also were compared with sampling variance to calibrate the statistical efficiency of the estimators of the T year floods Q (T) by two different flood frequency models. As a result, it was showed that for return periods greater than about T = 10 years the annual exceedance series estimators of Q (T) has smaller sampling variance than the annual maximum series estimators. It was showed that for the range of return periods the partial duration series estimators of Q (T) has smaller sampling variance than the annual maximum series estimate only if the POT model contains at least 2N (N:record length) items or more in order to estimate Q (T) more efficiently than the ANNMAX model.

1. Introduction

In the planning and design of water resources development project an estimation of river floods is a main task, and is not only important factor in evaluating economic consideration and safety but also basic information for the river arrangement and the optimal size of hydraulic structures. But design floods based on probability concept is very difficult to determine mathematically so that the determination of the design floods depends on the occurrence probability or frequency of floods.

Recently, an estimation of design floods by the frequency analysis in Korea rivers was done based on rainfall data of relatively long period records because the measured flood data are not sufficient, otherwise was accomplished by using fragmentary flood data accumulated in short period.

The primary objective of this study is to verify the applicability and efficiency of statistical flood frequency models based on the method of abstracting data from the discharge records at the stage

* Lecturer, Dept. of Civil Eng., Dankook University, Korea
** Professor, Dept. of Civil Eng., Dankook University, Korea

gaging stations within natural river basin. In this study, the possibility for developing of the regionalized regression model which can estimate the magnitude and frequency of river floods in the ungaged watersheds in Korea rivers is investigated by implementing regional frequency analysis.

2. Theoretical Background

2. 1 Outline of Models

In flood frequency analysis, the statistical analysis is required in order to estimate the flood magnitudes and is needed to relate the flood to its frequency, Estimates of the relationship of these flood magnitude (Q) to return period (T) are required in hydroeconomic analysis such as cost benefit analysis and flood insurance.

Flood frequency methods estimating the Q – T relationship from flood records are categorized into three types of statistical models as follows (NERC, 1975) :

(a) Partial Duration Series Model, POT Model (Peaks Over a Threshold Model)

(b) Annual Maximum Series Model, ANNMAX Model

(c) Time Series Model, TS Model

In the POT model a frequency distribution is fitted to the series of all floods which exceed a certain particular base or threshold value (q_0). Also, a frequency distribution in the ANNMAX model is fitted to the series of annual maximum floods. On the other hand, the TS model is fitted to the sequence of daily flows and also long synthetic sequences are generated from which a Q – T relationship is obtained (O' Donell, 1970 ; Quimpo, 1968).

Another method recently presented uses some results from crossing theory (Mejia, 1971). As indicated above, one criterion of theoretical comparison of these methods is the statistical efficiency with which either method estimates the parameters of the Q – T relationship in the same statistical condition.

2. 2 Probability Distribution Theory of Models

POT model for the partial duration series is specified as follows :

(1) The number k' of flood peaks exceeding a threshold q_0 every year is random.

This number has a Poisson distribution with mean $k\ (=\lambda q_0)$.

$$PR\,(K' = k') = \exp(-k)\frac{k^{k'}}{k'!}\ ,\ k' = 0, 1, \cdots\cdots \qquad (1)$$

(2) The flood peaks Qi exceeding threshold q_0 are independent and have an exponential distribution with parameter q_0 and β.

$$PR\,(Qi > q \mid Qi \geq q_0) = e^{-(q - q_0)/\beta} \qquad (2)$$

For the ANNMAX model the annual maximum flood may be less than q_0, the distribution of the annual maximum for the equation (1) and (2) is then given by the following set of equations.

$PR(Q_{max} < q_0) = \exp(-k)$ (3)

$PR(Q_{max} < q) = \exp(-\exp(-q - u)/\alpha))$ (4)

$= \exp[-\exp(-y)],\ q_0 \geq q_0$

where, $\alpha = \beta$ (5)

$u = q_0 + \beta(\ln k + y(T))$ (6)

$y(T) = -\ln[-\ln(1 - 1/T)]$ (7)

In the equation (7), y(T) is the $(1/T \times 100\%)$ point of the standardized or reduced Gumbel distribution.

2. 3 The Estimators $\hat{Q}(T)$ of Models

The T year flood Q(T) can be expressed in terms of either (u, α), or (q_0, β, k). The series of annual maximum can be used to derive estimates of (u, α), Q(T) can be replaced by $Q(T)_{AM}$ indicating an estimators obtained from an annual maximum series. On the other hand, the partial duration series can be used to estimate $(q_0\beta, k)$, Q(T) can be replaced by $Q(T)_{PD}$ indicating an estimators from a partial duration series.

2. 3. 1 The Estimators $\hat{Q}(T)_{AM}$ and its Sampling Variance

From the equation (4) and (6), the magnitude $Q(T)_{AM}$ with return period T of the annual maximum distribution can be written as :

$Q(T)AM = u + \alpha y(T)$ (8)

where, u : location parameter

α : scale parameter

y(T) : Gumbel reduced variate

In flood frequency analysis with the extreme value type－Ⅰ used as the probability model of equation (8) estimates of the parameter (u, α) are derived from the annual maximum series. A general method of parameter estimation can be explicitly defined in both forms $F = F(x)$ and $x = x(F)$, $F = F(x)$ is given in equation (4) and $x = x(F)$ can be obtained as (Greenwood, 1979) :

$x = u - \alpha \ln(-\ln F)$ (9)

In this case, two probability weighted moments are needed as :

$B_k = E[x F^k],\ k = 0, 1$ (10)

Also, equation (10) can be given as (Phien, 1987) :

$B_k = \{u + \alpha[\ln(1 + k) + \gamma]\} / (1 + k)$ (11)

Thus :

$B_0 = u + \gamma\alpha = E(x)$

$B_1 = [u + \alpha(\ln 2 + \gamma)] / 2$ (12)

Therefore, the PWM estimates of α and u can be expressed as :

$\hat{\alpha} = (2b_1 - b_0) / \ln 2$ (13)

$\hat{u} = b_0 - \gamma\hat{\alpha} = b_0 - 0.5772\hat{\alpha}$ (14)

where b_0 and b_1 are respectively the estimates of B_0 and B_1.

$$b_0 = \frac{1}{N}\sum_{i=1}^{N} q_i \tag{15}$$

$$b_1 = \frac{1}{N}\sum_{i=1}^{1} (i-1)\, q_i \,/\, (N-1) \tag{16}$$

γ = Euler's constant (= 0.5772)

where i is the rank of q_i in the sequence $q_1, q_2 \cdots\cdots q_n$ arranged in ascending order of magnitudes.

Therefore, from equation (13) and (14) the minimum variance unbiased (MVU) estimators of $Q(T)_{AM}$ is obtained as :

$$\hat{Q}(T)_{AM} = \hat{u} + \hat{\alpha}\, y(T) \tag{17}$$

and its sampling variance is (Kimball, 1949) :

$$Var(\hat{Q}(T)_{AM}) = \frac{\alpha^2}{N}\{1 + \frac{6}{\pi^2}(1-\gamma+y(T))^2\} \tag{18}$$

or

$$Var(\hat{Q}(T)_{AM}) = \frac{\alpha^2}{N}\{1 + 0.61\,(0.4228 + y(T))\}$$

$$= \frac{\alpha^2}{N}\{1.11 + 0.52y(T) + 0.61y^2(T)\} \tag{19}$$

In the equation (19), the expression in { } brackets depends only on T and may be denoted by AM(T). Referring to equation (5), i.e., $\alpha = \beta$, equation (19) may be rewritten :

$$Var(\hat{Q}(T)_{AM}) = \frac{\hat{\beta}^2}{N} AM(T) \tag{20}$$

2. 3. 2 The Estimators $\hat{Q}(T)_{PD}$ and its Sampling Variance

In order to estimate Q(T) from a partial duration series parameters (q_0, β, k) of model must be determined. The derivation of the parameter estimates depends on the method used to abstract the peak flows the historical records. In the first method the variation between years and between seasons in the number of peaks exceeding the threshold q_0 is ignored and a constant number of exceedances (k) is assumed to occur each year. Therefore the threshold q_0 is a fixed quantity. Otherwise, in the second and distinct method the variation between years in the number of peaks exceeding the threshold q_0 is formally considered, but the variation between seasons within the year ignored. The number of peaks in a year is considered to be a random variable with mean k. Because this is general and simple method of abstracting data from the historical records, it is usually used to flood frequency analysis.

The likelihood may be written :

$$L(q_1, q_2, \cdots\cdots q_M\, ;\, q_0, \beta) = \prod_{i=1}^{M} \frac{1}{\beta} e^{-(q_M - q_0)/\beta}$$

$$= \frac{1}{\beta^M} e^{-M(\bar{q} - q_0)/\beta} \tag{21}$$

where, $\bar{q} = \frac{1}{M} \prod_{i=1}^{M} q_i$ (22)

The maximum likelihood estimates for q_0 and β are :

$$\hat{q}_0 = q_{min} \quad (23)$$

$$\hat{\beta} = \bar{q} - q_{min} \quad (24)$$

where, q_{min} : smallest observed flood in the partial duration series (q_0)

$\bar{q}$: mean value flood in the partial duration series.

But showing that the $\hat{q}_0$ and $\hat{\beta}$ of equation (23) and (24) are biased. The unbiased estimators are :

$$\hat{q}_0 = \frac{M q_{min} - \bar{q}}{M-1} = q_{min} - \frac{\hat{\beta}}{M} \quad (25)$$

$$\hat{\beta} = \frac{M}{M-1}(\bar{q} - q_{min}) \quad (26)$$

Therefore, the estimated T - year flood $\hat{Q}(T)$ is :

$$\hat{Q}(T)_{PD} = \hat{q}_0 + \hat{\beta}(\ln k + y(T)) \quad (27)$$

The sample variance of $\hat{Q}(T)_{PD}$ is :

$$Var(\hat{Q}T)_{PD} = \frac{\frac{\hat{\beta}^2}{N}\left[\frac{(1-\ln k - y(T))^2}{Nk-1} + (\ln k + y(T))^2\right]}{k} \quad (28)$$

In the equation (28), $\hat{\beta}^2 / N$ depends not only on T and k but also on N. Letting $PD(T, k, N) = \{(1 - \ln k - y(T))^2 / (Nk - 1) + (\ln k + y(T))^2\} / k$ equation (28) may be expressed:

$$Var(\hat{Q}(T))_{PD} = \frac{\hat{\beta}^2}{N} PD(T, k, N) \quad (29)$$

3. Application and Analysis of the Models

3. 1 Selection of Sample Area and Station

Sample area and stations are stage gaging stations selected by considering whether the flow is natural condition in the Han River basin and quality of discharge data is good. In the application of the model, representative stations selected for the analysis are 6 stage gaging stations in Dam upstream area that have the same record length of the 20 years and the stage - discharge relation curves drawn up according to the items of river bed change.

Fig. 1 shows the location of the stage gaging stations.

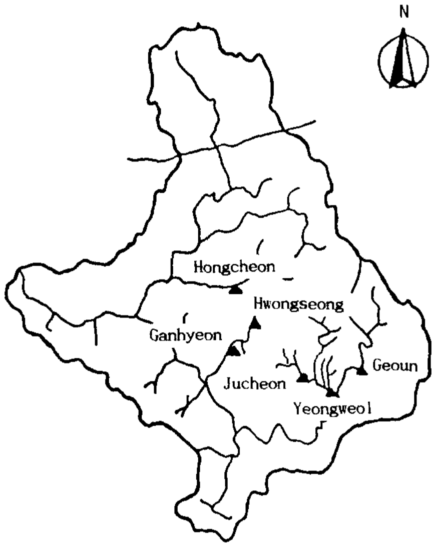

Fig. 1 Location of the selected stage gaging stations

3. 2 Data Collection and Arrangement

The basic data is daily mean stage with record length from 1969 to 1988 years at 6 stage gaging stations within watersheds. These stage data were collected from the hydrologic annual report in Korea and hydrologic investigation report in Ministry of Construction. Also, stage – discharge relation curves were collceted from the existing research report of various kinds.

The collected daily mean stage data were arranged as the forms of partial duration series and annual maximum series for each stations which corresponds to the purpose of this study. A arranged stage data for stations were converted into discharge data using stage – discharge relation curves. A converted discharge were rearranged as the forms of partial duration series and annual maximum series in order to apply to the statistical model for predicting flood frequency.

3. 3 Determination of Magnitude and Frequency of Floods

3. 3. 1 Parameter Estimation of the Models

First of all, the scale and location parameters of the models require to estimate floods Q (T) with specified recurrence intervals. Therefore, the estimates of parameter were obtained using the equation (13) (14) and (25) (26).

The results of these estimates are presented in Table 1.

Table 1. Parameter estimates of the model for each station

Parameter		Station					
		Hong-cheon	Gan-hyeon	Hwong-seong	Ju-cheon	Yeong-weol	Ge-oun
ANNMAX	α	634.93	534.17	365.40	572.23	1984.38	1924.80
model	u	522.00	617.46	468.07	331.87	1596.97	1556.10
POT	β	518.82	505.52	396.76	582.27	2078.01	1523.51
model	q_0	666.14	671.79	399.31	428.46	2412.74	1703.22

3. 3. 2 Estimators $\hat{Q}(T)$ Determination of the Models

In order to estimate Q(T) with specified recurrence intervals the results of Table 1 and equation (17)(27) were used.

The results of estimators are presented in Table 2 and Fig. 2 shows comparatively the flood frequency curve for each station.

Table 2. Estimators $\hat{Q}(T)$ of the models

(unit : CMS)

st.	model	exceedance probability (P), %						
		50	20	10	4	2	1	0.5
		return period (T), yrs.						
		2	5	10	25	50	100	200
Ho.	ANNMAX	754.7	1474.5	1950.9	2553.2	2999.6	3446.0	3886.1
	POT	856.3	1444.4	1833.7	2325.9	2690.6	3055.4	3415.0
Ga.	ANNMAX	813.2	1418.8	1819.6	2326.3	2701.8	3077.4	3447.7
	POT	857.1	1430.1	1809.4	2289.0	2644.4	2999.8	3350.2
Hw.	ANNMAX	602.0	1016.2	1290.3	1637.0	1893.9	2150.8	2404.0
	POT	544.7	994.5	1292.2	1668.6	1947.5	2226.5	2501.5
Ju.	ANNMAX	541.6	1190.3	1619.6	2162.5	2564.8	2967.1	3363.7
	POT	641.9	1301.9	1738.8	2291.2	2700.5	3109.9	3513.3
Ye.	ANNMAX	2324.3	4573.7	6062.6	7945.2	9340.2	10735.4	12110.8
	POT	3174.3	5529.9	7089.1	9060.5	10521.3	11982.4	13422.7
Ge.	ANNMAX	2261.5	4443.5	5887.7	7713.7	9066.9	10420.2	11754.3
	POT	2261.6	3988.6	5131.7	6577.1	7648.1	8719.3	9775.2

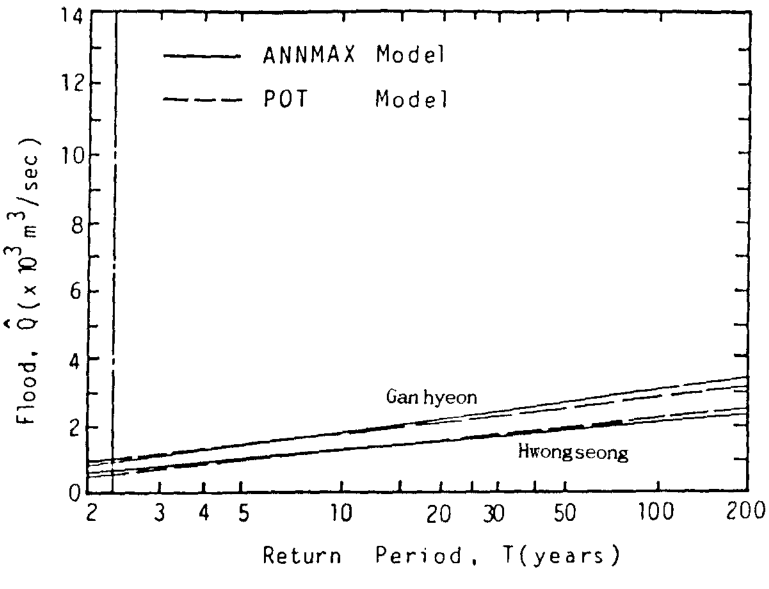

Fig. 2 Flood frequency curves for each station

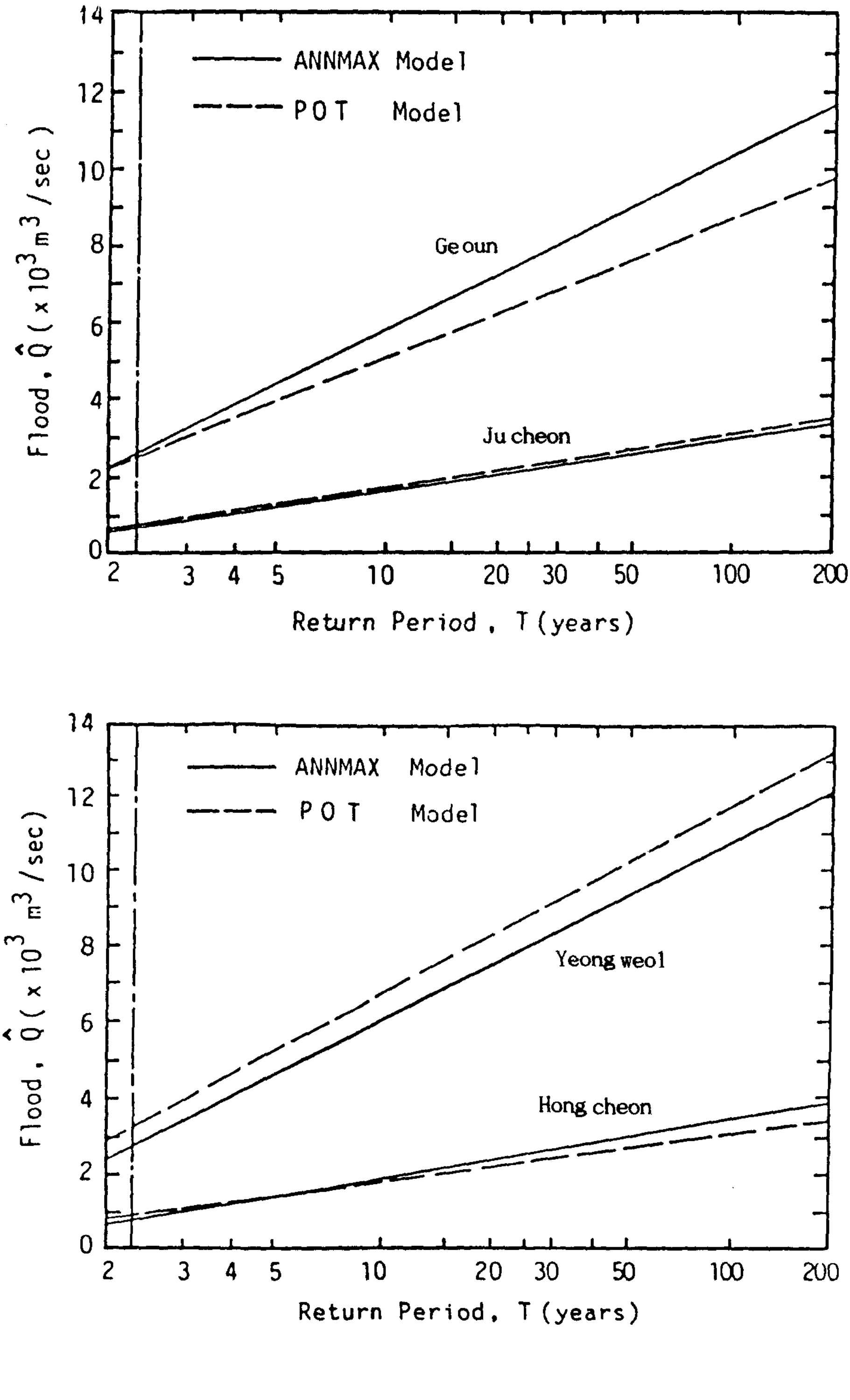

Fig. 2 (continued)

Table 3. Estimators ($\hat{Q}(2.33)$) of mean annual floods for each station

stat.	model	$\hat{Q}(2.33)$ (CMS)	return period (T), yrs.						
			2	5	10	25	50	100	200
Hong-cheon	ANNMAX	889.4	0.85	1.66	2.19	2.87	3.37	3.87	4.37
	POT	966.3	0.89	1.49	1.90	2.41	2.78	3.16	3.53
Gan-hyeon	ANNMAX	926.5	0.88	1.53	1.96	2.51	2.92	3.32	3.72
	POT	964.3	0.89	1.48	1.88	2.37	2.74	3.11	3.47
Hwong-seong	ANNMAX	679.5	0.89	1.50	1.90	2.40	2.79	3.17	3.54
	POT	628.9	0.87	1.58	2.05	2.65	3.10	3.54	3.98
Ju-cheon	ANNMAX	663.0	0.82	1.80	2.44	3.26	3.87	4.48	5.07
	POT	765.4	0.84	1.70	2.27	2.99	4.06	3.53	4.59
Yeong-weol	ANNMAX	2745.1	0.85	1.67	2.21	2.89	3.40	3.91	4.41
	POT	3615.1	0.88	1.53	1.96	2.51	2.91	3.31	3.71
Ge-oun	ANNMAX	2669.8	0.85	1.66	2.21	2.89	3.40	3.90	4.40
	POT	2584.7	0.87	1.54	1.99	2.54	2.96	3.37	3.78
Median	ANNMAX		0.85	1.66	2.19	2.87	3.37	3.87	4.37
	POT		0.87	1.54	1.96	2.51	2.91	3.31	3.71

3. 3. 3 Estimation of Mean Annual Floods ($\hat{Q}(2.33)$)

Mean annual floods Q(2.33) in index-flood method is obtained from floods corresponding the return period 2.33 years after drawn on extreme value probability paper.

in this study the return period T-year flood for each station from the equation (17) and (27) are as follows :

$$\hat{Q}(2.33)_{AM} = \hat{u} + \hat{\alpha} y(2.33)$$
$$\hat{Q}(2.33)_{PD} = \hat{q}_0 + \hat{\beta}(\ln k + y(2.33)) \qquad (30)$$

The results estimators of mean annual floods $\hat{Q}(2.33)$ are shown in Table 3.

4. Discussion and Consideration

4. 1 Applicability Verification of the Models

In order to verify the applicability of the models proposed in this study the sampling variance of estimators $\hat{Q}(T)$ by the models is obtained using equation (19) and (28), and then converted into standard error of estimators.

The results are shown in Table 4.

Table 4. Standard error of estimators ($\hat{Q}(T)$) for each station

(unit : %)

stat.	model	return period (T), yrs.						
		2	5	10	25	50	100	200
Hong-cheon	ANNMAX	22	17	17	17	17	17	17
	POT	5	12	14	16	17	18	18
Gan-hyeon	ANNMAX	17	15	15	15	15	15	15
	POT	5	12	14	16	17	18	18
Hwong-seong	ANNMAX	16	15	15	15	15	15	15
	POT	6	13	16	17	18	19	19
Ju-cheon	ANNMAX	28	19	18	18	18	17	17
	POT	8	15	17	18	19	20	20
Yeong-weol	ANNMAX	22	18	17	17	17	17	17
	POT	6	13	15	17	17	18	19
Ge-oun	ANNMAX	22	17	17	17	17	17	17
	POT	6	13	15	17	18	18	19

From the results of Table 4, it was seen that for large return period the ANNMAX model estimators has smaller standard error than the POT model estimators. Also, for the same values of return period the ANNMAX model has large values at $T < 10$ years and the POT model has generally large values at $T > 10$ years. Therefore, it was concluded that the ANNMAX model is useful for estimating the $\hat{Q}(T)$ which have higher return periods and for lower return periods the POT model is available.

Flood frequency models in this study are considered that ANNMAX model can be adopted as Standard Project Flood (SPF) for the design of the hydraulic structures, while the POT model seems to predict the design frequency of Probable Maximum Flood (PMF).

4. 2 Comparison and Discussion of Efficiency of Estimators $\hat{Q}(T)$ by the Models

From the equation (20) and (29), the efficiency of estimators ($\hat{Q}(T)$) by the models is as follows :

$$E = Var(Q(T)_{AM}) / Var(Q(T)_{PD}) = AM(T) / PD(T, k, N) \tag{31}$$

Fig. 3 shows the relationship between the AM(T) and PD(T, k, N).

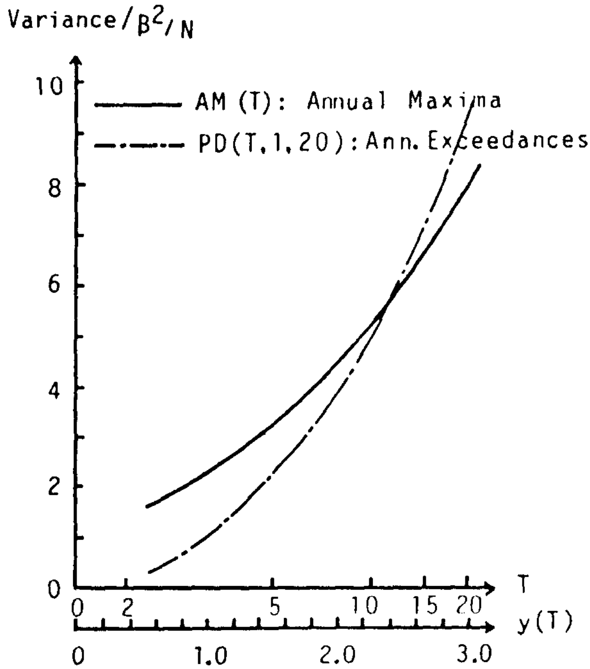

Fig. 3 Sampling variance of the estimators $\hat{Q}(T)$ by the models

As shown in Fig. 3, the PD(T, k, N) curve is shown for k = 1 and N = 20 which corresponds to the annual exceedance series. The curve does not change greatly with N and hence the values of k required to make E equal to unity do not depend greatly on N. Also, the figure shows that PD(T, 1, 20) is less than AM(T) for $T < 10$ years and that AM(T) is less than PD(T, 1, 20) for $T > 10$ years and hence $E < 1$ on condition that T is greater than 10 years. Therefore the T year flood Q(T) is estimated more efficiently by the ANNMAX model than by the POT model in case of $T > 10$ years, this corresponds to the applicability verification of the models.

On the other hand, because for higher return periods the value of k for which E is unity tends to 2, it is found that for the same return period the POT model requires 2N items (M = 2N) or more in order to estimate the flood of return period more efficiently than the ANNMAX model.

Consequently, it is considered that the POT model is usefully used for estimating relatively accurate and reasonable floods of return period as an alternative means of ANNMAX model, in view of the existing state of things in Korea where only a short flood record is available.

5. Conclusions

This study is to compare the applicability and efficiency of statitical models for predicting flood frequency at the stage gaging stations selected by considering whether the flow is natural condition in the Han River basin that have same record length of historical data and the stage - discharge relation curves.

The results of this study are as follows:

1) As a results of investigation of applicability of statitical models for predicting flood frequency, the ANNMAX model is useful for estimating the floods of higher return periods and the POT model is available for lower return periods.

2) From the comparison of efficiency of flood estimators by the flood frequency models, it is considered that the POT model is usefully used for predicting the magnitude and frequency of floods at each station where existing flood data of short record length are available.

3) The results of this study is expected to provide hydrologic information for predicting the magnitude and frequency of river floods. Furthermore, in order to estimate for the design flood which is needed for the design of the hydraulic structures in the ungaged sites, it is considered to be possible to develop the regionalized regression model by the regional frequency analysis.

Also, the supplement of the hydrologic data and the expansion of the stations, and the collection of the stage – discharge relation curves should be accumulated.

References

1. Ministry of Construction (1979), Hydrologic Data in Korea(Rainfall, Water Level:1962 – 1978).
2. Ministry of Construction (1989), Hydrologic Data in Korea(Rainfall,Water Level:1979 – 1988).
3. Ministry of Construction (1969 – 1988), An Annual Report for Hydrology Data.
4. Ministry of Construction (1974), An Report of River Investigation in Korea.
5. Ko, Jae Ung (1977), Regional Flood Frequency Studies in Korean Rivers, Proceedings of Korean Society of Civil Engineers, Vol. 25, No. 4, pp. 95 – 102.
6. Yang, Tong Yool and Ko, Jae Ung (1981), The Derivation of the Frequency Formulae from the Basin Characteristics, Journal of Korean Association of Hydrological Science, Vol. 14, No. 3, pp.37 – 46.
7. Lee, Won Hwan (1974), A Study on the Statistical Characteristics of Point Rainfall in Korea, Proceedings of Korean Society of Civil Engineers, Vol. 22, No. 1, pp. 63 – 82.
8. Lee, Won Hwan (1989), Hydrology, Munundang.
9. Yoon, Yong Nam (1989), Engineering Hydrology, Cheongmungak.
10. Korean Institute of Construction Technology(1989), Development of Synthetic Unit Hydrograph for Estimation of Design Flood, Final Report No. 88 – WR – 111.
11. Ashkar, F. and Rousselle, J. (1987), Partial Duration Series Modeling under the Assumption of a Poissonion Flood Count, J. Hydrol,. 90:135 – 144.
12. Benson, N.A. (1960), Characteristics of Frequency Curves Based on a Theoretical 1000 Years Record, USGS. Water Supply Paper, 1543 – A:51073.
13. Buishand, T.A. (1989), The Partial Duration Series Method with Fixed Number of Peaks, J. Hydrol., 109:1 – 9.
14. Buishand, T. A. (1989), Statistics of Extremes in Climatology, Stat. Neerl.

15. Cunnane, C. (1979), A Note on the Poisson Assumption in Partial Duration Series Models, WRR., 15(2):489-494.
16. Greenwood, J.A. et al. (1979), Probability Weighted Moments:Definition and Relation to Parameters of Several Distribution Expressable in Inverse Form, WRR., 15(5):1049-1054.
17. Kimball, B.F.(1949), An Approximation to the Sampling Variances of an Estimated Maximum Value of Given Frequency Based on the Fit of Doubly Exponential Distribution of Maximum Values, Ann. Math. Statist., 20:110-113.
18. Mejia, J.M. (1971), On the Generation of Multivariate Sequences Exhibiting the Hurst Phenomenon and Some Flood Frequency Analysis, Thesis, Colorado. State University.
19. National Environmental Research Council (1975), Flood Studies Report, Vol.1:Hydrological Studies, London.
20. O'Donnell, T.,etal.(1970), The Analysis and Generation of Hydrological Time Series, Lecture Course, Dep. Civil Eng., Imperial College of Science and Technology, London.
21. Phien, H.N.(1987), A Review of Methods of Parameter Estimation for the Extreme Value Type-I Distribution, J. Hydrol., 90:251-269.
22. Quimpo, R.G. (1968), Stochastic Analysis of Daily Flows, Proc. ASCE., J. Hydraul. Div., 94 (HY1):43-57.
23. Raghunath, H.M. (1985), Hydrology(Principles, Analysis and Design), John Wiley & Sons.
24. Weissman, I.(1978), Estimation of Parameters and Large Quantiles Based on the k Largest Observations, J. Am. Stat. Assoc., 73:812-815.