

最小 2 乗法を用いた極値統計解析モデルの censored data に対する精度の検討

山 口 正 隆*・畑 田 佳 男**・大 福 学***・野 中 浩 一****

1. 緒 言

最小 2 乗法に基づく極値統計解析モデル(合田, 1990; Yamaguchi ら, 1997) は, 欠落を伴う censored data に対して, 面倒な数値計算を要する最尤法に基づく方法(山口ら, 1998) のように一般的な形では適用できないが, 順位の明らかな資料を欠く場合のように censored data のプロテイングポジションが定まるときには, 相当する順位を空位とする簡単な操作によって適用可能(合田, 1990)になるが, モデルの精度は必ずしも明らかでない。

そこで, 本研究では母集団分布を Gumbel 分布, Weibull 分布, FT-II 型分布に指定した場合に, 広範な入力条件のもとにモンテカルロシミュレーションで発生した censored data の極値統計解析に基づいて, 最小 2 乗法に基づく拡張型モデル(Yamaguchi ら, 1997) の精度を種々の観点から明らかにする。また, 欠落を伴う風速資料などの観測資料に拡張型モデルを適用して再現確率統計量とその分散(標準偏差)を推定し, 最尤法に基づく結果との比較検討を行う。

2. 最小 2 乗法に基づく極値統計解析モデル

拡張型モデルは, Gumbel 分布, 形状母数 $k \geq 0.5$ の任意個数(ここでは, $k=0.5 \sim 10.0$ の 27 個)の Weibull 分布あるいは $k \geq 2.5$ の任意個数(ここでは, $k=2.5 \sim 40.0$ の 20 個)の FT-II 型分布を候補分布として, 合田(1990)のプロテイングポジション公式を用いた非超過確率表示に基づく再現確率統計量推定法, 最大相関係数基準に基づく最適分布選択法および jackknife 法に基づく再現確率統計量の分散推定法よりなる。モデルの変更点は候補分布における形状母数の個数と範囲の拡大および分散推定のための jackknife 法の適用である。拡張型モデルは合田(1990)のモデルを基礎としているので, 事象の年平均発生数 $\lambda (=N_T/K, N_T: \text{資料総数}, K: \text{資料年数})$ を導入することによって, 年最大値資料および極大値資料の両者に適用できる汎用性をもつ。

Gumbel 分布, Weibull 分布および FT-II 型分布の非超過確率 $F(H)$ はそれぞれ次式で表される。

$$F(H) = \exp[-\exp\{-(H-B)/A\}];$$
$$-\infty < H < \infty \dots\dots\dots (1)$$

$$F(H) = 1 - \exp[-\{(H-B)/A\}^k];$$
$$B \leq H < \infty \dots\dots\dots (2)$$

$$F(H) = \exp[-\{1 + (H-B)/kA\}^{-k}];$$
$$B - kA \leq H < \infty \dots\dots\dots (3)$$

ここに, A : 尺度母数, B : 位置母数, である。

3. 極値統計解析モデルの精度の検討

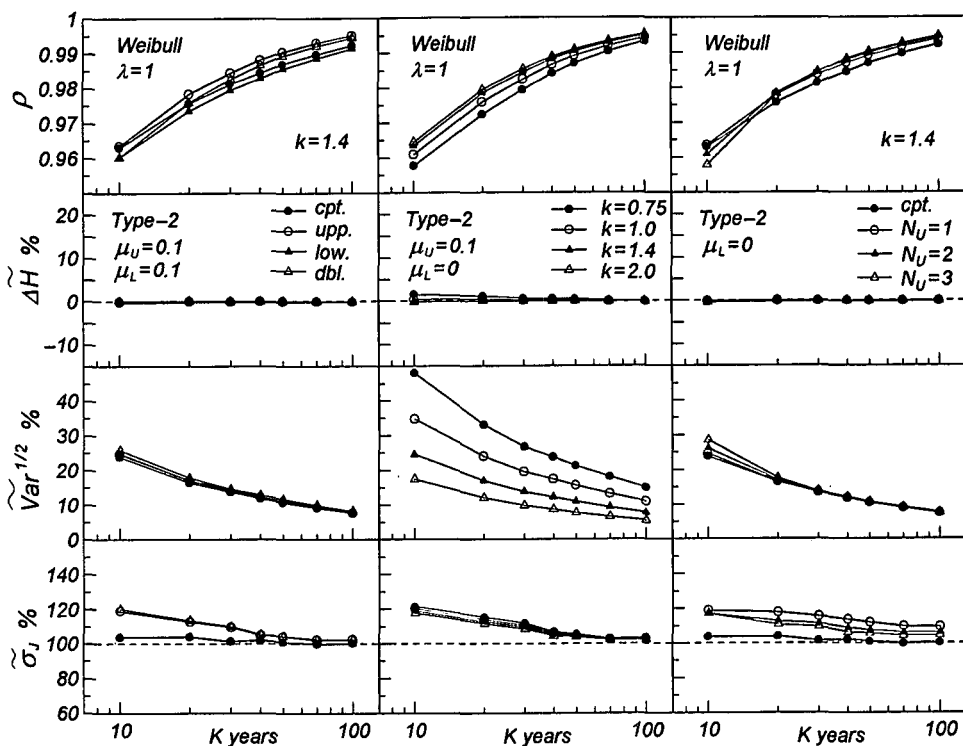
(1) シミュレーションの方法

censoring は, censoring 値が固定され標本数がランダムである第 1 種上限・下限・上下限 censoring と, その逆の第 2 種上限・下限・上下限 censoring の 6 種類(ここで検討する第 1～第 3 位資料が censoring を受ける場合は第 2 種上限 censoring に相当)に分けられるので, Gumbel 分布 ($A=1.39, B=4.5$), Weibull 分布 ($A=4.0, B=1.0$), FT-II 型分布 ($A=1.0, B=5.0$) を対象として, censoring 条件, 形状母数 k および年平均発生数 λ を適宜変えてそれぞれの censored data をモンテカルロシミュレーションにより発生する。年最大値資料および極大値資料は式(1)～式(3)の $F(H)$ に一様乱数を入力することによって作成されるが, 極大値資料の場合には, $F(H)$ のかわりに $1 + (1/\lambda) \log F(H)$ を用いる。そして, 最小 2 乗法による母数推定と適合度指標である相関係数 ρ , 再現確率統計量 H_R , jackknife 法による分散 σ_f^2 の計算を各標本ごとに行い, 5000 回の繰返し計算により諸量の平均値 ($\bar{\rho}, \bar{H}_R, \bar{\sigma}_f^2$), 再現確率統計量の bias ($\Delta H = \bar{H}_R - H_U; H_U: \text{再現確率統計量の真値}$) および再現確率統計量の分散 Var を求める。ただし, 簡単のため以下では「 $\bar{}$ 」を削除するとともに, 再現期間 100 年に対する検討を行う。また, 解析を既知母分布(母分布と候補分布の形状母数が一致)と未知母分布(最大相関係数規準により最適分布を選択)の場合に行う。

censored data の発生手順はつぎの通りである。

① 母集団の確率分布 $F(H)$ と母数 (A, B, K), 年平均発生数 λ および censoring 条件を指定する。censor-

* 正 会 員 工 博 愛媛大学教授 工学部環境建設工学科
** 正 会 員 工 修 愛媛大学助手 工学部環境建設工学科
*** 愛媛大学技官 工学部環境建設工学科
**** 学生会員 工 修 愛媛大学大学院理工学研究科



図一 第2種 censoring を受けた年最大値資料に対する誤差統計量 (既知母分布)

ing 条件には, 第1種 censoring の場合 censoring の上・下限値 H_U および H_L に相当する非超過確率値 $F(H_U)$ および $F(H_L)$, 第2種 censoring の場合上・下限の資料削除率 μ_U および μ_L を与える. 上限 censoring 値 H_U および下限 censoring 値 H_L は式 (1)~式 (3) から求める.

② 一樣乱数を式 (1)~式 (3) の inverse form に与えて, 大きさ $N_T = \lambda K$ ($K=10, 20, 30, 40, 50, 70, 100$ の7ケース, $\lambda=1, 4, 8, 16$, 最大16000) の標本 $H(i)$ を抽出し, 大きいものから小さいものにならべかえる. 第1種 censoring の場合には, N_T 個の資料から $H(N_U) \leq H_U < H(N_U-1)$ および $H(N_L+1) < H_L \leq H(N_L)$ に相当する資料順位 N_U および N_L を決めるとともに, H_L 未満の資料を削除し, N_L 個の資料とする. また, 第2種 censoring の場合には上限 censoring 比 μ_U に相当する資料の順位 N_U および下限 censoring 比 μ_L に相当する資料順位 N_L を決めるとともに, 順位 N_L 未満の資料を削除し, N_L 個の資料とする. いずれの censoring の場合にも, 計算には順位 $N_U+1 \sim N_L$ の資料 (個数 $N_L - N_U$) を使用するが, 第1種 censoring の場合には, H_U および H_L が N_T によらず固定され, N_U および N_L がシミュレーションごとにランダム値をとるのに対して, 第2種 censoring の場合には, N_U および N_L が N_T ごとに固定され, H_U および H_L が1回のシミュレーションごとにランダムに

変化する.

(2) 結果の考察

図一は Weibull 分布に基づく第2種 censoring を受けた年最大値資料に対する誤差統計量 (既知母分布の場合), すなわち相関係数 ρ , 再現期間100年に対して再現確率統計量の真値 H_U で無次元化した bias $\Delta \hat{H}$ と標準偏差 $\widehat{Var}^{1/2}$, および $Var^{1/2}$ で無次元化した標準偏差推定値 $\hat{\sigma}_j$ の%値と資料年数 K との関係を censoring の種類 (上限・下限 censoring 比 $\mu_U=0.1$, $\mu_L=0.1$), 形状母数 k および第1位~第3位までの資料削除数 ($N_U=1, 2, 3$) をパラメータとして示す. 図中の記号 cpt., upp., low., dbl. はそれぞれ censoring を考慮しない資料 (uncensored data) および下限, 上限, 上下限 censored data を意味する. 本モデルは, 既知母分布を仮定する場合, Yamaguchi ら (1997) が示したように, あるいは図でみるように, 年最大値資料および極大値資料に対する下限 censoring の条件下では bias および再現確率統計量の分散推定値に対する誤差をほとんど生じない. 以下では上限・上下限 censored data に対する検討結果を主に述べる.

相関係数は, 標本ごとの変動性が高い上位値資料の削除により上限・上下限 censored data の場合に大きい, 分散平方根 $Var^{1/2}$ は censoring の有無にあまり依存し

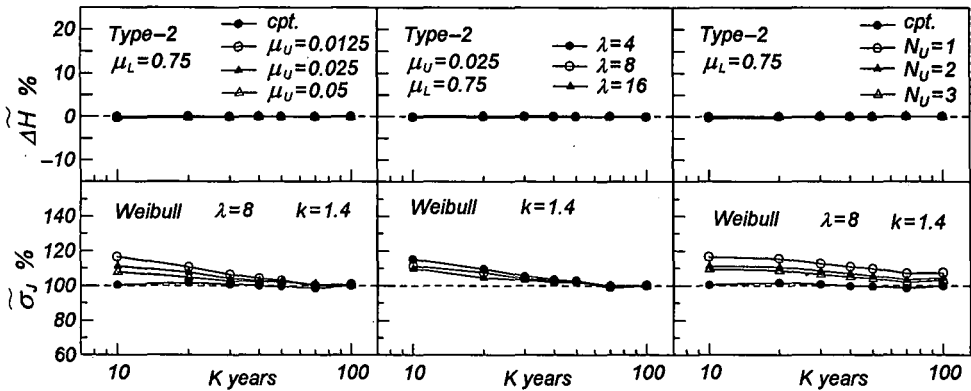


図-2 第2種 censoring を受けた極大値資料に対する誤差統計量 (既知母分布)

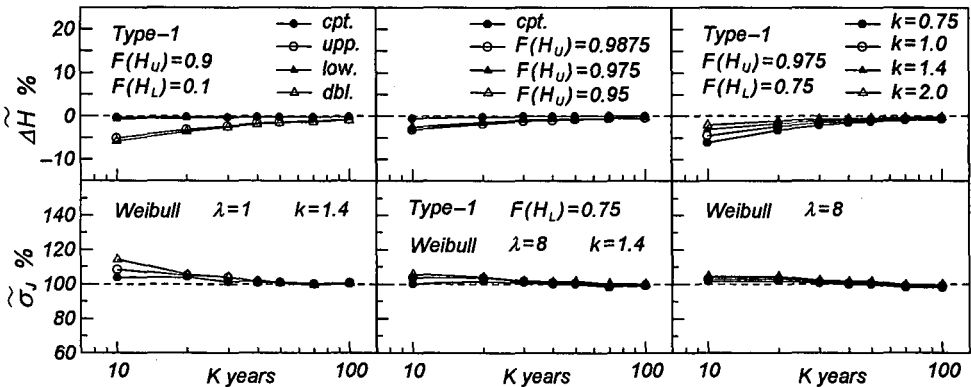


図-3 第1種 censoring を受けた極大値資料に対する誤差統計量 (既知母分布)

ない。この特性は第3位までの censored data の場合にもみられる。一方、形状母数が小さい値をとるほど、分布は正側に裾を長く引くことから、相関係数は小さく、 $Var^{1/2}$ は大きくなる。bias は censoring の有無および形状母数によらずほとんど生じない。 $Var^{1/2}$ の推定誤差 ($\sigma_j - 1$) は上限・上下限 censored data の場合最大 20% に達するが、資料年数の増加とともに減少する。この挙動は形状母数によらない。また、第3位までの censored data に対する $Var^{1/2}$ の推定誤差が第1位のみ censored data の場合より小さいという奇妙な挙動がみられる。

図-2 は形状母数 $k=1.4$ の Weibull 分布に基づく第2種 censoring を受けた極大値資料の場合に、誤差統計量 $\Delta \hat{H}$ と σ_j に及ぼす上限 censoring 比 μ_u 、年平均発生数 λ および第3位までの資料の censoring ($N_u=1, 2, 3$) の影響を示す。極大値資料の場合にも、bias は上限 censoring 比 μ_u 、年平均発生数 λ および上位値削除数 N_u によらずほとんど生じないが、 $Var^{1/2}$ の推定誤差は上限 censoring 比あるいは上位値削除数が小さいほど大きい。すなわち、 $Var^{1/2}$ の推定誤差は上限 censoring の導入

によって不連続的に増加し、上限 censoring 比あるいは上位値削除数が増加するほど減少するという奇妙な挙動を示す。また、資料年数 K や年平均発生数 λ の増加に伴う資料数 N の増大は $Var^{1/2}$ の推定誤差の減少をもたらす。しかし、第3位までの資料の censoring を受けた極大値資料のみならず年最大値資料に対しても、資料年数の増加に伴う $Var^{1/2}$ の推定誤差の減少はあまりみられない。

図-3 は Weibull 分布に基づく第1種 censoring を受けた年最大値資料および極大値資料の場合に、誤差統計量に及ぼす censoring の種類、上限 censoring 値 H_u および形状母数 k の影響を示したものである。第1種上限・上下限 censored data の場合、負の bias を生じるが、 $Var^{1/2}$ の推定誤差は小さい。これらの絶対値は上限 censoring 値にほとんど依存しない。形状母数の影響をみると、形状母数が小さいほど負の bias が大きい。また、 $Var^{1/2}$ の推定誤差は無視しうる。また、bias および $Var^{1/2}$ の推定誤差は資料年数、つまり資料数の増加とともにほぼ消失する。

以上のように、既知母分布の場合、第2種上限 censoring を受けた年最大値資料および極大値資料では、bias

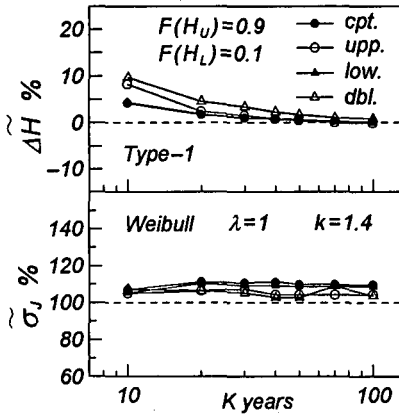


図-4 第1種 censoring を受けた年最大値資料に対する誤差統計量 (未知母分布)

はほとんど生じないが、 $Var^{1/2}$ の推定誤差は形状母数によらず最大 20% 程度であって、上限 censoring 比が小さいほど、あるいは上位値削除数が少ないほど大きい。一方、第1種上限 censored data では、bias は負の値をとるが、 $Var^{1/2}$ の推定誤差は小さいし、資料年数や年平均発生数の増加に伴う資料数の増大により、bias は減少する。つまり、第1種上限 censored data と第2種上限 censored data の誤差挙動は対照的である。この特性は Gumbel 分布および FT-II 型分布の場合にも同様にみられる。

つぎに、母分布が未知の場合の検討結果を述べる。図-4 は Weibull 分布を母分布とする第1種 censoring を受けた年最大値資料に対する誤差統計量を例示したものである。解析では、Gumbel 分布と形状母数 $k=0.5\sim 10.0$ の Weibull 分布の計 28 種類を候補分布として用いている。これは現地観測資料へのモデルの適用においても同じである。bias はいずれの censoring 条件でも正の値をとり、censored data の場合により大きい。資料年数の増加とともに減少する。一方、 $Var^{1/2}$ の推定誤差は下限 censored data や uncensored data の場合より上限・上下限 censored data の場合にやや小さく、また資料年数が増加してもそれほど減少しない。未知母分布を仮定する場合、誤差統計量は censoring 条件に対して必ずしも一貫した挙動を示さないことから、未知母分布に対するシミュレーション方法および解析方法に再考の余地があると推測される。

4. 現地観測資料への適用

(1) 高潮偏差資料

鹿児島における年最大高潮偏差資料は図-5 に示すように、1945～1996 年の 52 年間のうち 1972 年以降毎年得られているだけである。また、100 cm 以上の高潮偏差は

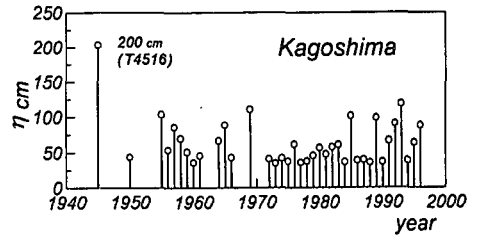


図-5 鹿児島における年最大高潮偏差の時系列

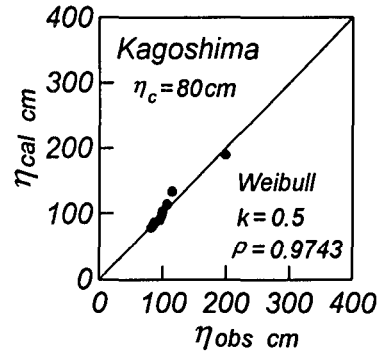


図-6 鹿児島における年最大高潮偏差資料に対する Q-Q プロット図

表-1 鹿児島における年最大高潮偏差資料に対する極値統計解析結果

data (AM)	N	k	ρ	η ₁₀₀	σ _{η100}
η _c = 50 cm	19	0.6	0.9747	183	52
η _c = 60 cm	14	0.5	0.9733	181	55
η _c = 80 cm	10	0.5	0.9743	178	56
MLM		F 4.34		182	19

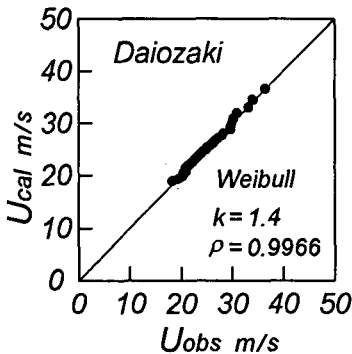
K = 52 years η₁₀₀, σ_{η100} : cm F : FT-II

52 年間で 4 回生起しているが、そのうち 3 回は 1972 年以前であり、しかも第1位高潮偏差 200 cm は 1945 年の台風 4516 号 (枕崎台風) により発生している。解析は下限 censoring 値を η_c = 50, 60, 80 cm の 3 種類に設定し、第1種下限 censoring 問題として行う。図-6 は η_c = 80 cm とした場合に、資料値 η_{obs} を横軸、各資料値に割当てた非超過確率値に対して最適分布 (形状母数 $k=0.5$ の Weibull 分布) から得られる計算値 η_{cal} を縦軸とする Q-Q プロット図である。相関係数の値は大きくないが、上位値資料に対する最適分布の適合性は悪くない。表-1 は極値統計解析結果の一覧であり、最尤法に基づく推定値も与えられる。100 年確率高潮偏差 η₁₀₀ は 178～183 cm、その標準偏差 σ_{η100} は 52～56 cm である。すなわち、第1位高潮偏差が突出した値をとるので、確率高潮偏差の標準偏差がかなり大きく、したがって推定値の信頼度は高

表一 大王崎における年最大風速資料に対する極値統計解析結果

data (AM)	K	k	ρ	U_{100}	$\sigma_{U_{100}}$
censored	34	1.4	0.9966	42.7	3.6
uncensored	33	1.7	0.9954	38.9	2.1
MLM	34	1.512		42.7	2.9

$N=33 U_{100}, \sigma_{U_{100}}: \text{m/s}$



図一 大王崎における年最大風速資料に対するQ-Qプロット図

くない。一方、最尤法に基づく結果によれば、最適分布は $k=4.34$ の FT-II 型分布（この場合、Weibull 分布の解は得られない）であるが、確率高潮偏差は最小2乗法に基づく結果とほぼ一致し、標準偏差は約 1/3 となつて小さい。

(2) 風速資料

これは太平洋岸志摩半島先端に位置する大王崎灯台（風速計の海拔高度 37.2 m）で取得された 1953～1986 年の 34 年間の年最大風速資料である。このうち、観測期間中の最大風速（第 1 位風速）は、周辺の風観測資料の調査から明らかなように、台風 5915 号（伊勢湾台風）によって生起していると推定されるが、欠測となっているので、ここでは第 1 位資料を欠いた第 2 種上限 censoring 問題として解析を行う。表一は極値統計解析結果の一覧であり、第 1 位資料を無視した解析結果および最尤法に基づく解析結果（山口ら，1998）も与えられる。また図一は第 1 位資料を欠測とした場合の Q-Q プロット図を示す。最適分布は $k=1.4$ の Weibull 分布であり、この場合の相関係数はほぼ 1 に近く、適合度が非常に高い。100 年確率風速 U_{100} とその標準偏差 $\sigma_{U_{100}}$ は $42.7 \pm 3.6 \text{ m/s}$ と推定される。100 年確率風速は最尤法に基づく結果と一致するが、標準偏差は約 25% 大きい。また、第 1 位資料を無視した解析では、確率風速の標準偏差が相対的に小さいが、確率風速も過小評価される。

(3) 波高資料

これは宮崎県油津港の水深 50 m 地点において

表一 油津における年最大波高資料に対する極値統計解析結果

data (AM)	K	k	ρ	H_{50}	$\sigma_{H_{50}}$
censored	16	1.0	0.9725	10.81	1.78
uncensored	15	1.0	0.9725	10.91	1.81
MLM	16	G		10.71	1.16

$N=15 H_{50}, \sigma_{H_{50}}: \text{m}$ G: Gumbel

1975～1990 年の 16 年間にわたり得られた年最大波高資料であるが、1984 年における波高資料は取得されていない。観測期間中の最大波高は台風 8013 号に伴う 10.34 m である。一方、1984 年の年最大波高は周辺海域における観測記録から年最大波高の最小値（1985 年の 4.57 m）より低いと推定されるので、最下位資料を欠く第 2 種下限 censoring 問題として解析を行う。表一は極値統計解析結果の一覧であり、最下位資料を無視した解析結果および最尤法に基づく解析結果（山口ら，1998）も与えられる。50 確率波高 H_{50} と標準偏差 $\sigma_{H_{50}}$ は $10.81 \pm 1.78 \text{ m}$ であり、第 1 位波高が突出した値をとるため、標準偏差そのものの値がかなり大きい。また、最下位資料を無視した推定値は、censored data として扱った場合と同程度の値をとる。最尤法に基づく 50 年確率波高の推定値（山口ら，1998）は上述の推定値より若干低い程度であるが、標準偏差は 0.6 m 以上小さい。

5. 結 語

最小2乗法に基づく極値統計解析モデルは uncensored data や下限 censored data に対して高精度の再現確率統計量推定値を生じるが、上限・上下限 censored data に対しては、censoring の種類に依存して再現確率統計量あるいはその標準偏差推定値のいずれかを過大・過小評価する可能性をもつこと、および本モデルは欠落を伴う観測資料に対して最尤法と同程度の再現確率統計量の推定値を簡単な計算によって与えるが、標準偏差に対してより大きい推定値を生じるので、推定値の有効性の点で最尤法に基づくモデルより劣ることなどが見出された。

参 考 文 献

- 合田良實(1990): 港湾構造物の耐波設計一波浪工学への序説一、鹿島出版会、333 p.
- 山口正隆・畑田佳男・大福 学・前川隆海(1998): censoring を考慮した極値統計解析モデルおよび現地観測資料への適用、海岸工学論文集、第 45 巻、pp. 216-220.
- Yamaguchi, M. and Y. Hatada (1997): An extremal analysis system and its application to the estimation of extremes of meteorological and oceanographic elements around the coasts of Japan, Proc. WAVES97, Vol. 2, pp. 932-946.