

標本サイズと水文頻度解析

HYDROLOGICAL FREQUENCY ANALYSIS METHODS SUITABLE TO THE SAMPLE SIZE OF EXTREME EVENTS

宝 馨¹・小林 健一郎²

Kaoru TAKARA and Kenichiro KOBAYASHI

¹フェロー 工博 京都大学教授 防災研究所 社会防災研究部門 (〒611-0011 京都府宇治市五ヶ庄)

²正会員 Dr.-Ing. 京都大学助教 生存基盤科学研究ユニット (〒611-0011 京都府宇治市五ヶ庄)

This paper describes an important consideration on the size of hydrological extreme samples (annual maximum series) in hydrological frequency analysis. When the sample size is small (less than 30), the authors recommend a graphical method with the Weibull plotting formula, which provides quantile (T -year event) estimates in a safety side. Traditional parametric methods, which use various kinds of probability distributions, are recommended for samples with a size of not less than 30 to the return period T considered. Highlighting a new consideration for large samples with a size of $T > 100$, which is often taken as the return period for design flood estimation in major river basins in Japan, this paper suggests a non-parametric method that can be applied to many rain gauges already having long records $T > 100$ years. The method is distribution-free, objective and easy to use. The bootstrap method is effective to correct the bias of the estimates obtained by the non-parametric method and to give the accuracy of the estimates.

Key Words: sample size, graphical method, parametric method, non-parametric approach, quantile estimation

1. はじめに

従来、水文データは標本サイズ（観測年数）が大きくなり、水文頻度解析においては、確率紙を用いた図式推定法（最小二乗法）や種々の確率分布を用いた解析的方法（パラメトリックな方法）が用いられてきた。しかしながら、観測年数が100年を超える場所も多くなってきたので、このような大きなサイズを持つ標本（大標本）に対しては従来とは異なる方法を用いた方がよい可能性も出てきた。

ここでは、豪雨や洪水のような極値データの頻度解析において、標本サイズごとに解析方法を変えることを提案し、その意義を明らかにする。すなわち、まず、小標本（30年未満のデータセット）に対しては安全側の推定値を与える簡略推定法として、ワイブル公式を用いた確率紙による方法を推奨するとともに、中標本（30年以上再現期間以下）の場合の、一般に良く行われている水文頻度解析の方法を概説する。

想定する再現期間（たとえば100年）を超えるような大標本に対しては、確率分布を想定しない順序統計学的方法（ノンパラメトリックな方法）を提案する。この方

法においては、ブートストラップ（Bootstrap）法による確率水文量の推定誤差を補正することをも合わせて提案し、その精度を例示した。こうした考え方は、今後ますますデータの蓄積が進む実務現場における一つの客観的な分かり易い新しい方法として一考の価値があると言えよう。

2. 小標本の場合の水文頻度解析

ここで「小標本」とは、観測年数が30年に満たない場合とし、特別な取り扱いをすることとする。30年を区切りとするのは、次の二つの理由による。

- ・ 気象諸量の平年値は西暦で 10 年ごとにそれまでの 30 年の平均値をとっている。たとえば、西暦 2009 年にいるとすれば、1971 年から 2000 年までの 30 年間の平均値を当該気象量の平年値としており、それを日常生活における一つの目安としている。
- ・ 頻度解析において標本（一群の観測データ）に確率分布を当てはめる場合、30〜40 程度の標本サイズから再現確率統計量の安定性が増してくる。

したがって、30年未満の小標本の場合は特別な取り扱い

をする必要があると考えて良いという立場を取る。

これまでの国内外の長年にわたる研究によって、水文極値データ（年最大雨量，年最大河川流量）は、一般に右にひずむ分布をすることが知られている。したがって、観測年数 N が短い場合においても、極値確率紙（グンベル確率紙）あるいは対数正規確率紙にデータをプロットし、それに平分線（fair curve）を引いて所定の非超過確率（あるいはリターンピリオド）に対応する再現確率統計量（確率点，クオンタイル，quantile）を求める。このとき、 N 個のデータのうち小さいほうから i 番目のデータ（順序統計量）に非超過確率を与えるプロットイング・ポジション公式として、ワイブル（Weibull）公式 $i/(N+1)$ を用いると良い。なぜなら、この公式は、もともと一様分布を仮定したときに誘導されるものであり、 $N < 30$ というような小さい標本のそれぞれのデータは、等確率で発生したと見るのが、公平・客観的な立場であるからである。

もう一つの理由は、そのほかのプロットイング・ポジション公式に比べて、大きい目の再現確率統計量を与えるからである（たとえば、高棹ら¹⁾）。短い観測年数のうちには、大きな（危険な）事象が含まれないことが容易に想定されるので、こうした小標本にはワイブル公式を適用して大きい目の（すなわち、安全側の）再現確率統計量を求めておいたほうが良いと言えるのである。

用いる確率紙としては、極値（グンベル）確率紙をまず推奨する。なぜならば、最大極値データを扱うのであり、それに対して理論的背景を持つ確率分布ならびに確率紙を用いるのが正当な態度である。ただし、小標本であるから、極値確率紙上でプロットする点が直線に並ぶことを期待しすぎはいけない。この場合、平分線を引くのが難しいが、目分量で引くと客観性に欠けるので、最小二乗法で求めるのが良い。プロットされたデータ点と平分線との乖離がある程度あってもやむをえない。

小標本を使って50年、100年という再現期間の確率水文量を求めようとする自体そもそも無理があるのであるから、この場合は、あくまでもラフな推定であると割り切って、一時近似的な推定に過ぎないことを解析者も解析結果を利用しようとする者も十分理解しておく必要がある。

極値確率紙が利用できないときは、データの対数をとって、対数正規確率紙にワイブル公式でプロットすると良い。対数正規分布はグンベル分布に近い形状を表現できるからである。あくまでもラフな推定であるから、極値確率紙（グンベル分布）にこだわりすぎる必要はない。

3. 中標本の場合の水文頻度解析

標本サイズが30～40以上になってくると、極値データ

に色々な確率分布を当てはめて頻度解析を行うことができるようになる。すなわち、次の①～⑤の手順を行うのが一般的である。

- ① 水文極値データの収集・吟味
（独立性・均質性などをチェック）
- ② 候補となる確率分布を複数選ぶ
（極値分布，対数正規分布，ガンマ分布などが候補となる）
- ③ データに確率分布を当てはめる（確率紙や解析の方法を用いる）
- ④ 当てはまり具合（適合度）を評価する
（目視，客観的適合度評価規準による）
- ⑤ 適合度の良い確率分布を選び、それによって再現確率統計量を求める

この場合、候補となる確率分布は多数あり、そのどれもが似たような良い適合度を示す。適合度の評価指標として、第一著者^{1)・2)}は標準最小二乗基準（SLSC）を提案した。これは日本全国の国レベル，地方レベルの水文頻度解析において広く使われている。

似たような適合度は示しても、確率水文量は少なからず異なる。また、データがさらに蓄積したときに確率水文量の値が大きく変わるようでは現場の設計・計画において大きな変更や見直しが要請されることにもなりかねない。したがって、適合度が良く、かつ、データが増えても確率水文量の変化が小さくてすむような確率分布を採用するのが望ましい。

こうした観点から、上記の⑤のステップでとどまっていた従来の水文頻度解析に、もう1ステップ追加することを提案した²⁾。すなわち、適合度に加えて、確率水文量の変動性（あるいは推定精度）を評価の基準に加えることを提案したのである。計算機集約型統計学（computer intensive statistics）として近年よく使われるようになったジャックナイフ法やブートストラップ法に代表されるリサンプリング（resampling）の手法を導入することを試みた。これによって、確率水文量の推定値のバイアス補正ならびに推定精度の定量化ができる。同じような適合度を示す確率分布が複数ある場合にはジャックナイフ推定誤差が小さい分布のほうを選ぶことを推奨する。

当然のことながら、母数の個数が少ない確率分布のほうは、リサンプリングによって生成されるデータ（標本）群に対して変動が小さくなる。母数の個数が多いとそれぞれの標本に対して適合度が良くなる分、逆に確率水文量の推定値の変動は大きくなるのである。

3母数の分布よりも2母数の分布のほうが、安定性が良い、あるいはロバスト(robust)であると言える。したがって、グンベル分布の適合度が優れていればそれを用いるのが良い。グンベル分布が適合しない場合には、3母数の分布を試み、以下の手順で適合度が良くかつ確率

水文学の変動が少ないモデルを選ぶ。

- ① 水文極値データの収集・吟味（独立性・均質性などをチェック）
- ② 候補となる確率分布を複数選ぶ
（対数正規分布，極値分布，ガンマ分布などが候補となる）
- ③ データに確率分布を当てはめる（確率紙や解析的方法を用いる）
- ④ 当てはまり具合（適合度）を評価する
（目視，客観的適合度評価規準による）
- ⑤ 適合度の良い確率分布を複数個選ぶ．1つしかない場合はその確率分布を用いて確率水文量を求める。
- ⑥ 同程度の適合度を示す複数の確率分布に対して，ジャックナイフ法を適用し確率水文量の推定精度（ジャックナイフ推定誤差）を求め，これの小さいモデルを選ぶ。

この方法は，すでに我が国のほとんどの一級河川，二級河川等の実務において採用されており，その基本的な解析ソフト「水文統計ユーティリティ」が国土技術研究センターのウェブ（<http://www.kasen-keikaku.jp/suimon/>）から参照して利用できる。

4. 大標本時代の水文頻度解析

近代的な水文観測が始められたのは明治時代であり，当時は19世紀の終盤であった．いまや21世紀であり，1世紀（100年）以上のデータが蓄積されてきたことになる．これからもデータの蓄積は進んでいく。

わが国の主要河川では水文頻度解析によって100年確率水文量を求めることが多い．そして，100年以上のデータを備える流域が出てき始めたのである．このような時代になっても，上述した中標本に対する解析手法を使い続けなければならないのであろうか．このような場合には，解析的方法（確率分布を用いるパラメトリックな方法）は必要なく，ノンパラメトリックな方法を用いたら良いのではないか．本章では，これについて考察する。

4.1 パラメトリックな従来型の手法による結果

まず，参考のために従来型の水文頻度解析の結果を示す．滋賀県北東部に位置する姉川流域（標本サイズ $N=108$ ），天野川流域（ $N=107$ ），芹川流域（ $N=107$ ），彦根（ $N=100$ ），余呉川流域（ $N=106$ ），愛知県・豊川流域（石田地点 $N=104$ ）の6カ所の年最大雨量データを極値（グンベル）確率紙にカナン公式でプロットする。

図1は姉川流域年最大2日雨量の例である³⁾．データ点が直線状に並んでいないので，グンベル分布の適合が極めて悪いことは，一見して明らかである．このプロットに最小二乗法を適用すると，推定されたグンベル分布は，図1の回帰直線のようになり，非超過確率 0.990

（再現期間100年に相当）とこの回帰直線の交点から x 軸の雨量を読み取ると 360 mm となる．これが，図解法（ただし，カナン公式を用いた）による推定値（小標本の場合と同様のやり方）である。

同じ手法を，天野川流域，芹川流域，彦根，余呉川流域，豊川流域に対して適用すると，順に 330, 430, 335, 260, 302 mm という100年確率水文量が得られた。

3. で述べた，中標本の場合の方法を姉川流域の年最大2日雨量に適用してみたところ，グンベル分布で 313 mm という確率水文量が得られた．一般化極値（GEV）分布を用いると357 mm，対数Pearson III型分布を用いると 373 mm となった。

グンベル確率紙上でプロット点を直線でつなぎ，100年確率水文量を求めると，図-1に示すように，440 mm となる．同様に，天野川流域，芹川流域，彦根，余呉川流域，豊川流域に対して適用すると，順にそれぞれ 380, 480, 380, 285, 285 mm という100年確率水文量になった．これらの結果は，グンベル分布とGEV分布によるパラメトリックな解析の場合と併せて，表-1に整理されている。

なお，この計算については，前述の国土技術研究センターの水文統計ユーティリティを用いた．表-1からわかるように，適合度指標 $SLSC > 0.03$ となる場合が多く，グンベル分布はデータへの適合度が全体的に悪いと言える⁴⁾．またグンベル分布，GEV分布の両者とも芹川流域，彦根に対しての適合度が特に悪い（芹川に対して $SLSC=0.149, 0.039$ ，彦根に対して $SLSC=0.181, 0.044$ ）。

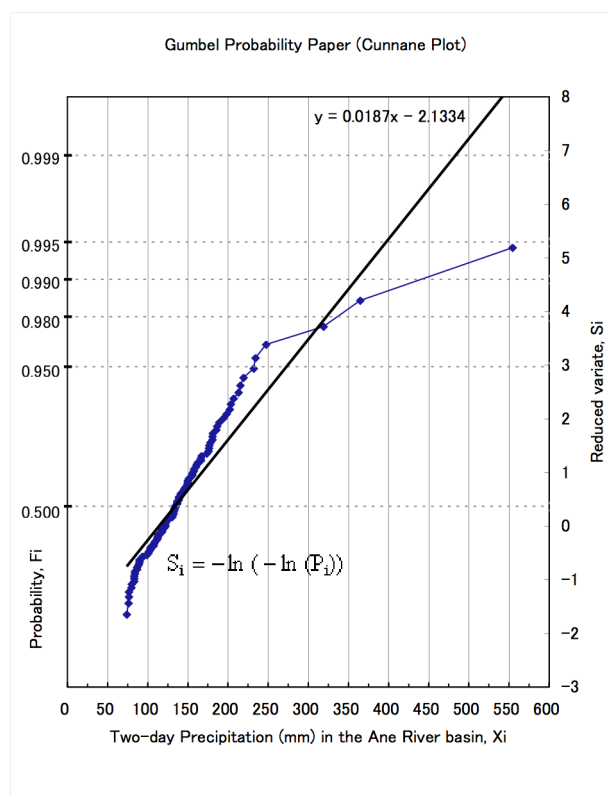


図-1 姉川流域年最大2日雨量の極値確率紙へのプロット

表-1 従来型のパラメトリックな手法による100年確率日雨量(mm) [姉川のみ2日雨量]

確率分布	ゲンベル分布 (極値確率紙)		ゲンベル分布		GEV分布	
推定法	最小二乗法	経験分布法	L積率法	SLSC	L積率法	SLSC
姉川流域	360	440	290	0.033	298	0.026
天野川流域	330	380	250	0.032	281	0.021
芹川流域	430	480	313	0.149	401	0.039
彦根	335	380	231	0.181	292	0.044
余呉川流域	260	285	195	0.035	177	0.020
豊川流域	302	285	302	0.021	300	0.021

表-2 ノンパラメトリックな手法による100年確率日雨量(mm) [姉川のみ2日雨量]

	統計年数 N (年)	(I) ワイブル公式	(II) カナン公式	(II') ブートストラップ法
姉川流域	108	537	463	426
天野川流域	107	517	413	383
芹川流域	107	735	541	494
彦根	100	560	436	381
余呉川流域	106	413	315	288
豊川流域	104	286	284	280

4.2 提案する方法 (ノンパラメトリック手法)

ここでは、対象とする確率年(T 年)以上の観測年数をもつ観測所の極値データを用いて、ノンパラメトリック法で T 年確率統計量を求める方法を提案する。計算機集約型統計手法 (リサンプリング法の一種であるブートストラップ法) でその精度を検証する。具体的な手順は、以下のようである。

Step 1: 大標本極値データ ($N>T$) を確率紙にプロットする。分布に依存しないため、単純な普通目盛 (横軸は水文量, 縦軸は0から1の非超過確率で普通目盛とする) を考える。プロットング公式は、近年よく使われるカナン公式 $(i-0.4)/(N+0.2)$ を採用する。比較のためワイブル公式 $i/(N+1)$ も併用する。

Step 2: 各プロット点を順順に直線でつないでいく。すなわち、これがいわゆる経験分布である。

Step 3: 非超過確率 $1-1/T$ (すなわち、リターンペリオド T 年) に相当する水文量の値を線形内挿手法で求める。

Step 4: ブートストラップ法により確率水文量の偏りを補正し、その推定精度を調べる。すなわち、 N 個あるデータの内から繰り返しを許して N 個のデータを抽出することによりブートストラップ標本を生成する。これに **Step 1** から **Step 3** の手順を多数回繰り返して確率水文量のブートストラップ推定量を得る。

4.3 結果と考察 (ノンパラメトリック手法)

極値確率紙ではなく、4.2で示した普通目盛り確率紙を用いる方法を適用する。確率分布を用いないので確

率分布関数に含まれる母数 (parameters) の推定が必要でないため、ノンパラメトリックな方法である。

ただし、各順序統計量 ($x_1, x_2, \dots, x_N$) に非超過確率を与える必要がある。ここでは、ワイブル公式、カナン公式によりプロットして得られる T 年確率雨量を求める。 $T=100$ 年を超えるような大きさの標本 ($N>100$) は、水文統計では「大標本」と称して良いで

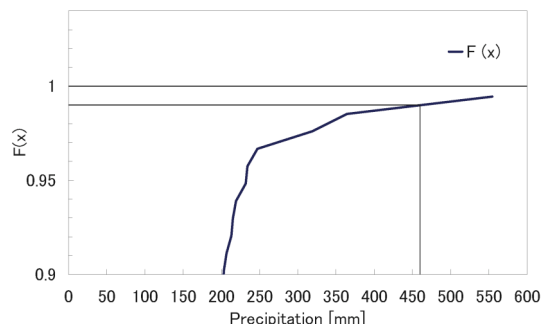
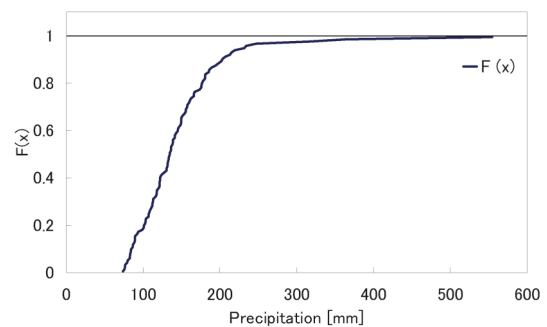


図-2 普通目盛にプロットした経験分布(上:全データ, 下:非超過確率0.90以上の部分を拡大表示)

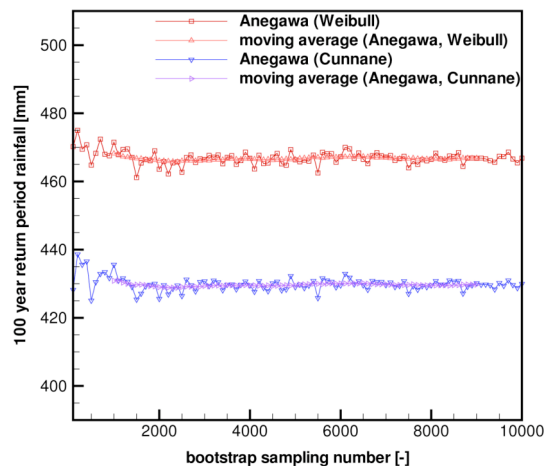


図-3 ブートストラップ標本の数とそれに応じて得られる100年確率雨量の変動(赤線(上):ワイブル公式を使った場合, 青線(下):カナン公式を使った場合;両者とも移動平均をとって平滑化した値も併記してある)。

あろう。

普通目盛でプロット(カナン公式)すると, 図-2 のようになる。図-2の上のグラフが, 標本に含まれる全データに対する確率分布関数に相当する。このままでは見にくいので, 非超過確率の大きい部分だけ拡大したのが, 図-2の下グラフである。この下のグラフから読み取れるように, この方法によれば, 100年確率雨量は463 mm となった。同様にしてワイブル公式を用いると, 100年確率雨量は537 mm となった。

これは, ここで扱った年最大雨量系列に対して, 従来のパラメトリックな方法では, 確率分布の適合度が悪く, 図-1に示すような, 最大値が飛び抜けて大きいような標本(データセット)に対して再現確率統計量(確率水文量)を過小推定することによる。

他の4カ所についても同様の方法を適用して得られた結果を整理すると, 表-2の(I), (II)のようになる。これと表-1の結果とを比べると分かるように, 提案する方法の方が確率雨量を多く見積もることになる。

バイアス補正のためにブートストラップ法^{2), 3)}を適用する。これは, 元の標本(サイズ N) から繰り返しを許して N 個のデータを取り出した標本(ブートストラップ標本)について, 対象とする統計量(ここでは100年確率水文量)を求めるという作業を多数のブートストラップ標本に対して行い, それに基づいて, 元の標本から得られる統計量を補正するとともに, 補正された統計量の精度(ばらつき)を明らかにする手法である。

試算を繰り返し, ブートストラップ標本の数 B を $B=100$ から10000まで100個きざみに変えてみたところ, 2000以上とすると結果が安定することがわかった。図-

3は, 姉川流域の年最大2日雨量に対して, 試算した結果を示している。横軸がブートストラップ標本の数で, 縦軸がブートストラップ法によって得られる100年確率雨量である。図中, 赤線(上)がワイブル公式を使った場合, 青線(下)がカナン公式を使った場合で, 両者とも移動平均をとって平滑化した値も併記してある。

結局 $B=2000$ として, ブートストラップ法で推定し直した姉川流域の100年確率2日雨量はカナン公式で426 mm となった。すなわち, ここで提案した手法によれば, 100年確率2日雨量は426 mmである。これは, 図-1, 表-1の極値確率紙における経験分布法による440 mm を少し下回る結果である。なお, 参考のため試みたワイブル公式の場合は470 mm となった。これは, このプロット・ポジション公式がもつ特徴からして, 過大評価であるとみなすことができる。

カナン公式を使って同様に天野川流域, 芹川流域, 彦根, 余呉川流域, 豊川流域における100年確率日雨量をブートストラップ法で求めた結果は, それぞれ, 383, 494, 381, 288, 280 mm となった(表-2の(II'))。

豊川流域のデータを極値確率紙にプロットすると, 図-4のようになる。このデータセットの特徴は, 第1位の雨量と第2位の雨量の値が近いことであり, 図-1で見たような姉川ほか滋賀県北東部の雨量のように第1位が大きく飛び離れているものとは異なる。すなわち, 表-2の豊川流域の行を見てわかるように, (I), (II), (II')の推定値の差が小さくなる。第1位と第2位との狭い区間において100年確率の値が決まるからである。

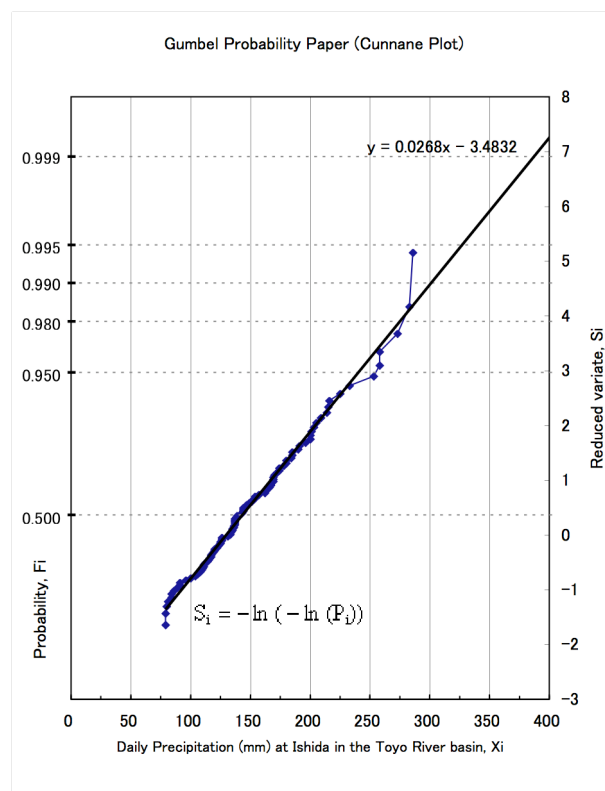


図-4 豊川流域年最大日雨量の極値確率紙へのプロット

図-4からわかるように、パラメトリックな従来法であれば、極値（グンベル）分布に当てはめた解は、図中の回帰直線であり、これによると100年確率（すなわち非超過確率0.990の線との交点）の日雨量値は302 mmということになる（表-1のグンベル分布の最小二乗法の欄）。なお、極値確率紙上でのプロットがこの程度の高い直線性を示すと、グンベル分布の他の推定法（L積率法）でも302 mm、GEV分布でも300 mmとほぼ近い推定値を与えることになる。

ただし、ノンパラメトリック法では、この豊川のデータセットでは第1位と第2位とが近い値であるので、プロット点が縦に立った形でプロットされるので、これらの推定値よりも15 mm以上下方修正されることになる。一方、図-1で見た姉川のように第1位と第2位とが離れていると、パラメトリック法による解析解は、ノンパラメトリック法によって上方修正されることになる。

本論文は、筆者らが提案してきた考え方^{4)・5)}を整理して再編集したものである。最近になって、石原・仲江川⁶⁾が、この方法を気象庁の全国51地点の日降水量データ（1901年～2006年の106年の系列、那覇のみ戦中データの欠損により101年）に適用した結果を報告している。その報告によると、ノンパラメトリック法で得られた100年確率日雨量推定値（A）と従来型のパラメトリックな方法で得られた推定値（B）とを比較したところ、両者の比（A/B）は51地点平均で1.03であった。すなわち、ノンパラメトリック法が、従来法よりも3%程度大きい確率水文量を与える結果を示している。106年のうちの最大値が、ノンパラメトリック法で得られる推定値よりもかなり大きい地点がいくつもあり（名古屋、高知、宮崎、津など）、これらがその結果に影響しているともいえる。ノンパラメトリック法の推定値の方が小さくなる地点もあることに留意しなければならない。いずれにせよ、さらに検討を重ねる必要があると考えている。

5. 結 論

従来、求めようとする再現確率統計量（確率水文量）のリターンピリオドに対して、標本サイズが相対的に小さいことが多かった。よって、確率分布を用いるパラメトリックな手法により、極値データの母集団を推定する方法がとられてきた。

しかしながら、データ数が極端に少ない小標本の場合には、推定精度が悪いことが水文統計学的に知られている。中標本の場合においては、いくつもの確率分布を検討しなければならないし、その作業における実務者の悩みは今なお大きいものがある。

本研究では、標本サイズの大きさによって、手法を変えるべきであるということを提案した。上述したよ

うに、小標本には確率紙を用いた安全側（ワイブル公式を用いる）の簡易推定、中標本に対しては従来のパラメトリックな方法、そして、リターンピリオドを越えるような観測年数を持つ標本に対しては、ノンパラメトリックな方法を推奨した。

特に、本論文で提案したように、リターンピリオドを超えるようなサンプルサイズを持つ標本に対してノンパラメトリックな方法（経験分布を用いる方法）は有用であろうと考える。ただし、上で見たように、ノンパラメトリックな方法（経験分布を用いる方法）を標本に対して1回だけ適用すると、偏りのある100年確率統計量の評価を与えるので、ブートストラップ法によって偏り補正した方が良い。

今後の課題として、これを現場に適用しようとする、たとえば、次のような問題が考えられる。

- (1) データ年数がリターンピリオドを超えたときに計画・設計値（確率水文量）が変更になることになるがそれでいいのだろうか。
- (2) 上流では50年確率、下流では100年確率という河川がたくさんあるが、上流と下流とで手法や計画の整合性が問題視されないか。

といったことである。これらについては、現場の実務者とともに検討を加える必要があろう。

謝 辞

本研究で用いた水文データは、滋賀県河港課及び国土交通省（豊川流域の極値データ）より提供いただいた。ここに記して謝意を表する次第である。

参考文献

- 1) 高棹琢馬・宝 馨・清水 章：琵琶湖流域水文データの基礎的分析，京都大学防災研究所年報，第29号B-2，pp. 157-171，1986。
- 2) 宝 馨・高棹琢馬：水文頻度解析における確率分布モデルの評価規準，土木学会論文集，第393号/II-9，pp. 151-160，1988。
- 3) 宝 馨：水文頻度解析の進歩と将来展望，水文・水資源学会誌，第11巻，第7号，pp. 740-756，1998。
- 4) 宝 馨：大標本時代の水文頻度解析手法—リターンピリオドを超えるようなサイズの標本に対する極値データ解析—，京都大学防災研究所年報，第49号B，pp. 7-12，2006。
- 5) 宝 馨・小林健一郎：水文極値の大標本に対するノンパラメトリックな頻度解析，日本自然災害学会年次学術講演会，九州大学，II-4-3，2008。
- 6) 石原幸司・仲江川敏之：全国51地点におけるノンパラメトリック手法を用いた確率降水量の算出，水文・水資源学会誌，第21巻，第6号，pp. 459-463，2008。

(2008.9.30受付)