Doi:10.11840/j.issn.1001-6392.2019.06.007

基于广义 Pareto 分布的波高极值序列频率分析

陈子燊[1,2]，任杰[3]

（1. 中山大学 新华学院，广东 广州 510520；2. 中山大学 地理科学与规划学院，广东 广州 510275；
3. 中山大学 海洋学院近岸海洋科学与技术研究中心，广东 广州 510275）

摘 要：应用广义 Pareto 分布（GPD）分析超阈值波高序列的设计值。以位于美国北卡罗来纳州的 FRF 历时 32 年连续测量的逐日波高序列为例，检验了不同波高阈值样本的泊松分布，采用多种方法综合确定最佳阈值的拟合优度指标。对最优广义 Pareto 分布和 GEV 分布及 P-Ⅲ分布推算的波高重现水平做了对比分析。得到以下结论：（1）波高的 GPD 属于短尾型分布；（2）拟合优度指标表明构建的波高 GPD 模型普遍优于 GEV 和 P-Ⅲ型；（3）GPD 的参数估计方法对设计波高的计算结果有较大影响。

关键词：广义 Pareto 分布；广义极值分布；P-Ⅲ型分布；波高阈值；拟合优度指标

中图分类号：P731.22 **文献标识码**：A **文章编号**：1001-6932(2019)06-0656-06

Frequency analysis for extreme sequence of wave heights based on generalized Pareto distribution

CHEN Zi-shen[1,2], REN Jie[3]

(1. Xinhua College of Sun Yat-sen University, Guangzhou 510520, China; 2. School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China; 3. Institute of Estuarine and Coastal Research, Marine School of Sun Yat-Sen University, Guangzhou 510275, China)

Abstract: Generalized Pareto distribution (GPD) was used to analyze the design value of the peak over threshold wave height sequence. Taking the daily wave height sequence 32 years continuously measured by Field Research Facility (FRF) in north Carolina of the USA as an example, the Poisson distribution under different wave height thresholds samples was tested, and the fitting optimization index of the best threshold was used to comprehensively determine the optimal generalized Pareto distribution. A comparative analysis was made of the return period wave heights among the optimal generalized Pareto distribution, GEV distribution and P-Ⅲ distribution. The following conclusions are obtained: (1) the GPD of wave heights belongs to the short-tailed distribution; (2) the goodness of fit test indicates that the wave height GPD model constructed is generally superior to GEV and P-Ⅲ models; (3) the parameter estimation method of GPD has a great influence on the calculation results of designed wave heights.

Keywords: generalized Pareto distribution; generalized extreme value distribution; P-Ⅲ distribution; wave height threshold; goodness of fit index

进入 21 世纪，全球气候变化发生极端事件频率和强度有所增大。风暴浪是一典型的灾害事件，是极值分布的重要研究对象。陈子燊等（2017）采用非对称 Archimedean Copula 函数与 Kendall 分布函数分析极端波况下的波高、周期和风速 3 变量年最大值的联合概率分布与风险率及其设计分位数，为海岸海洋工程设计和风险评估提供了参考依据。然而，如何利用有限的海洋水文观测资料，获取尽可能多的极端海浪信息，提高推算波高重现水平的精度是海岸海洋工程规划设计和灾害风险评估非常

收稿日期：2018-09-23；修订日期：2019-07-16
基金项目：国家自然科学基金（41371498）。
作者简介：陈子燊（1952-），教授，研究方向：极端水文气象事件的概率分布。电子邮箱：eesczs@mail.sysu.edu.cn。

重要的研究内容。超限量频率分析是极值统计建模理论的重要组成部分（Coles，2001），国外已有不少探索与研究（Van et al，1985；Rosbjerg et al，1992；Wang，1991；Madsen et al，1997；Hosking et al，1987）。目前国内对极值序列的超限频率研究主要应用在社会学、保险精算学和洪水序列分析领域（王剑峰 等，2010；颜亦琪 等，2010；方彬 等，2005；王善序，1999；岑泰林 等，2018），还少见对波高序列的超限频率研究。分析中如何选取阈值、超限量数和超限量分布的拟合优度检验等关键问题还需要更多的探索与实践（戴昌军 等，2006）。

本文在概述广义 Pareto 分布（简记为 GPD）模型的基础上，重点介绍 GPD 模型阈值选择方法，最后以逐日波高序列的实例对阈值选择与超阈值数检验的拟合优度检验等加以详细分析，并将 GPD 与 GEV 分布及 P-Ⅲ型分布推算成果加以对比。

1　极值分布模型

单变量极值分布模型可分为两种：其一为分组区块最大值模型（Block Maximum Group of Models，简称 BM）。其原理是对所得到的数据进行分块，常用年最大值方法（Annual maximal series，AMS）选取年最大值样本作为建立模型的观测数据。BM 模型的前提条件是样本独立同分布（IID）。这种方法不尽合理之处在于观测数据中常常会出现一个大值之后还跟有其他大值的“集串”现象，如一年内多次出现的风暴事件，故按年最大值抽样后常常出现多个极端波高远大于年份出现的最大值的情况。按年最大值抽样显然会造成波高数据的浪费。另一种为超限量峰量模型（Peaks Over Threshold Models，POT），其以超特定阈值为样本抽样前提，可获取更多的波高极值数据来建立统计模型。POT 模型需满足超限量发生的时间服从泊松分布，超限彼此相互独立，服从 GPD（Generalized Pareto Distribution）分布等条件。

1.1　广义 Pareto 分布

设序列 $\{x_n\}$ 的分布函数为 $F(x)$，定义 $F_u(y)$ 为随机变量 X 超过阈值 u 的条件分布函数：

$$F_u(y)=\mathrm{P}(X-u\leqslant x|X>u)=\frac{F(u+y)-F(u)}{1-F(u)}=\frac{F(x)-F(u)}{1-F(u)}\Rightarrow F(x)=F_u(y)(1-F(u))+F(u) \tag{1}$$

研究表明（Scholz et al，1987），当 u 阈值足够高时，条件超量分布函数 $F_u(y)$ 收敛于广义 Pareto 分布，累积分布函数为：

$$F_u(y)\approx G(x,\ \xi,\ \sigma,\ u)=\begin{cases}1-(1-\xi\dfrac{x-u}{\sigma})^{1/\xi} & \xi\neq 0\\ 1-e^{-(x-u)/\sigma} & \xi=0\end{cases} \tag{2}$$

当 $\xi=0$ 时，GPD 对应于指数分布；当 $\xi<0$ 时为常 Pareto 分布，$x\in[u,\ \infty)$；当 $\xi>0$，为 Pareto Ⅱ型分布（短尾），上限为 $u+\sigma/\xi$。有关研究证明了超阈值（$X-u$）数服从泊松分布（Leadbetter，1991）。

GPD 模型 T 年一遇的分位数 x_T 为：

$$x_T=\begin{cases}u+\dfrac{\sigma}{\xi}[1-(\lambda T)^{-\xi}] & \xi\neq 0\\ u+\sigma\ln(\lambda T) & \xi=0\end{cases} \tag{3}$$

1.2　GPD 的阈值

阈值 u 的合理确定是正确估计 GPD 模型参数 ξ 和 σ 的重要前提。如果阈值 u 选取的过高，会导致超限数据量太少，使估计出来的参数方差很大；如果阈值 u 选取的过低，则不能保证超限量分布的收敛性，使估计产生大的偏差。本文采用以下三种方法相结合确定阈值 u：

一是采用 GPD 模型的 Anderson - Darling 拟合优度检验方法。Anderson-Darling 检验原理：单样本的 Anderson-Darling 检验是利用样本分布函数（CDF）和经验分布函数（ECDF）之间的Anderson-Darling 的统计量 A^2：$A^2=n\int_{-\infty}^{+\infty}[F_n(x)-F_0(x)]^2[F_0(x)(1-F_0(x))]^{-1}dF_0(x)$来检验样本是否属于此特定分布，即判断原假设 H_0 是否成立。K 个样本 Anderson-Darling 拟合优度检验方法是对单样本检验方法的扩展与通用化(Scholz et al,1987)。以 F_1，F_2，…，F_n 表示各样本的 CDF，原假设 H_0：$F_1=F_2=\cdots=F_n$。样本 x_i 的个数和 CDF 分别用 n_i 和 F_{in_i} 表示，$\mathrm{N}=\sum n_i$ 为所有样本的个数，$H_N(x)$ 为所有 N 个样本的 CDF。K 个样本的 Anderson-Darling 统计量 A_{kN}^2 为：

$$A_{kN}^2 = \sum_{i=1}^{k} n_i \int_{B_N} \frac{[F(x) - H_N(x)]^2}{H_N(x)[1 - H_N(x)]} dH_N(x) \quad (5)$$

其中 $B_N = \{x \in R : H_N(x) < 1\}$；$A_{kN}^2$ 经过归一化处理得到 T_{kN}：$T_{kN} = [A_{kN}^2 - (k-1)] / \sigma_N$。式中，$\sigma_N$ 为 A_{kN}^2 的标准差。若 T_{kN} 小于高斯分布在置信度 α 下的临界值 $t_{k-1}(\alpha)$，则接受原假设 H_0。实际计算中可通过插值外推得到 Anderson-Darling 统计量 A_{kN}^2 的 P_{AD} 值，当 $P_{AD} > \alpha$，接受原假设；否则拒绝。计算中以逐个超限量样本的经验分布和理论累积分布为两样本，计算统计量 T_{2N}，进而计算 P_{AD} 值。详细计算原理和具体步骤可参见文献（Scholz et al, 1987）。

其次，选择一定的阈值区间，分别对区间内不同阈值的超限数使用 χ^2 假设检验其是否服从泊松分布：$P(x = k) = e^{-\lambda}\lambda^k / k!$，$k = 0, 1, 2, \cdots$，式中 λ 为发生超限的平均频次。对服从泊松分布样本的阈值再根据其他拟合优度检验指标做进一步的筛选。

再次，对各超限样本的 GPD 重现水平的推算结果分别采用均方根误差（RMSE）、经验频率和理论频率拟合误差平方和（Q）和概率点位相关系数（PPCC）检验其拟合优度。

2 实例研究

2.1 基本数据与研究背景

采用美国北卡罗来纳州 FRF（Field Facility Research）试验场 1985–2016 年观测的日最大波高数据。FRF 面向美国大西洋，经纬度坐标为 23.04° N，112.7° E。其间测量的海浪最大波高为 8.12 m。

2.2 波高阈值 ML

采用以下步骤确定极值波高阈值：

1）以历年中极值波高出现的最小值 3.15 m（2002 年）为 GPD 的初始阈值 u_0，组成阈值区间：$u_i = u_0 + (i-1)/5 + 0.1$，$i = 1, \cdots, 12$，阈值选择范围为 3.15 ~ 4.25 m。为满足抽样的超限样本之间的相互独立，分别以 10 天和 20 天为抽样时间间隔，构成 24 个超限极值波高序列；

2）在显著水平 0.05 下，对各超限量样本数的泊松分布做了 χ^2 检验，检验表明，间隔 10 天抽样的 24 个超限抽样系列中有 8 个序列超限量样本数符合泊松分布，间隔 20 天抽样的 24 个超限抽样系列中有 7 个序列超限量样本数符合泊松分布（见表 1 中原假设，H_0：样本服从泊松分布）；

3）24 个超限抽样系列的 Anderson-Darling 检验的 P 值都大于 $\alpha = 0.05$ 的临界值，多数超限量样本的频率分布图也显示样本点据和理论 CDF 拟合良好（图略），表明这些超限样本符合 GPD 分布。剔除不符合泊松分布的超限样本可见，阈值为 3.25 m 的超限样本的 PAD 值最大，表明此阈值可作为 FRF 极值波高序列的最优阈值；

4）对超限抽样系列分别使用具有统计特性良好的极大似然估计法（MLE）和概率权重矩法（PWM）估计 GPD 模型的参数。GPD 参数估计与各超限样本的 GPD 重现水平推算结果分别采用均方根误差（RMSE）、经验频率和理论频率拟合误差平方和（Q）和概率点位相关系数（PPCC）作拟合优度检验。

均方根误差（RMSE）：

$$RMSE = \sqrt{\sum_{i=1}^{n} ((\hat{x}_i - x_i)/x_i)^2 / n} \quad (6)$$

经验频率和理论频率拟合误差平方和（Q）：

$$Q = \sum_{i=1}^{n} (P_i - P_{ei})^2 \quad (7)$$

概率点位相关系数（$PPCC$）：

$$PPCC = \sum_{i=1}^{n} (x_i - x_m)(\hat{x}_i - \hat{x}_m) / \left[\sum_{i=1}^{n} (x_i - x_m)^2 \sum_{i=1}^{n} (\hat{x}_i - \hat{x}_m)^2 \right]^{\frac{1}{2}} \quad (8)$$

式中，P_{ei} 为经验累积频率；P_i 为理论累积频率；n 为样本容量。在样本容量相同的情况下，均方根误差和 Q 值越小，相关系数 $PPCC$ 越大，样本与理论线形拟合越好。

综合上述结果，确定以 10 天间隔，阈值为 3.25 m，PWM 参数估计方法拟合的 GPD 模型，即：$G(x, \xi, \sigma, u) = 1 - (1 - 0.025 \times \frac{x - 3.25}{0.92})^{1/0.025}$ 为最优超限量波高的 GPD 模型。

表 1　波高 GPD 模型的阈值、参数估计与拟合优度检验结果

抽样间隔	序号	阈值 (m^3/s)	H_0	POT	λ	PWM 参数估计与优度检验						MLE 参数估计与优度检验					
						ξ	σ	P_{AD}	RMSE	Q	PPCC	ξ	σ	P_{AD}	RMSE	Q	PPCC
10 天	1	3.15	0	146	4.56	0.116	1.04	0.678	0.028	0.004	0.989	0.074	1.01	0.667	0.037	0.395	0.991
	2	3.25	0	137	4.28	0.025	0.92	0.748	0.043	0.000	0.991	0.035	0.92	0.747	0.040	0.030	0.991
	3	3.35	0	121	3.78	0.058	0.96	0.697	0.039	0.001	0.989	0.051	0.95	0.701	0.040	0.005	0.989
	4	3.45	0	108	3.38	0.064	0.96	0.708	0.040	0.002	0.988	0.055	0.95	0.710	0.042	0.005	0.988
	5	3.55	0	95	2.97	0.122	1.03	0.694	0.033	0.006	0.983	0.077	0.99	0.690	0.042	0.155	0.985
	6	3.65	0	85	2.66	0.157	1.07	0.640	0.030	0.019	0.978	0.085	1.00	0.630	0.044	0.284	0.983
	7	3.75	0	75	2.34	0.267	1.19	0.593	0.021	0.133	0.962	0.110	1.04	0.577	0.049	0.910	0.978
	8	3.85	1	71	2.22	0.178	1.05	0.589	0.029	0.054	0.970	0.073	0.96	0.577	0.048	0.404	0.980
	9	3.95	1	61	1.91	0.371	1.27	0.309	0.015	0.502	0.935	0.116	1.04	0.289	0.054	1.269	0.971
	10	4.05	1	59	1.84	0.210	1.04	0.479	0.027	0.147	0.958	0.057	0.91	0.461	0.052	0.595	0.976
	11	4.15	0	55	1.72	0.116	0.91	0.559	0.039	0.058	0.968	0.011	0.82	0.551	0.060	0.299	0.978
	12	4.25	1	51	1.59	0.029	0.80	0.568	0.055	0.019	0.975	−0.040	0.74	0.575	0.073	0.145	0.980
20 天	1	3.15	0	124	3.88	0.192	1.21	0.581	0.024	0.021	0.985	0.117	1.14	0.565	0.043	0.753	0.989
	2	3.25	0	120	3.75	0.050	1.00	0.732	0.043	0.001	0.990	0.061	1.01	0.732	0.040	0.029	0.990
	3	3.35	0	107	3.34	0.083	1.04	0.698	0.039	0.001	0.988	0.077	1.03	0.697	0.040	0.002	0.988
	4	3.45	0	95	2.97	0.127	1.09	0.674	0.034	0.004	0.985	0.095	1.06	0.672	0.041	0.072	0.986
	5	3.55	0	85	2.66	0.173	1.15	0.671	0.030	0.014	0.980	0.109	1.08	0.668	0.044	0.213	0.984
	6	3.65	0	76	2.38	0.237	1.22	0.618	0.024	0.052	0.971	0.124	1.11	0.604	0.048	0.474	0.980
	7	3.75	0	68	2.13	0.340	1.34	0.536	0.017	0.220	0.954	0.143	1.14	0.515	0.055	0.987	0.976
	8	3.85	1	64	2.00	0.293	1.24	0.528	0.020	0.171	0.957	0.119	1.07	0.525	0.051	0.712	0.976
	9	3.95	1	57	1.78	0.408	1.36	0.299	0.014	0.518	0.932	0.137	1.10	0.299	0.056	1.115	0.970
	10	4.05	1	55	1.72	0.265	1.14	0.419	0.023	0.207	0.952	0.084	0.97	0.414	0.052	0.595	0.974
	11	4.15	1	51	1.59	0.207	1.04	0.434	0.030	0.147	0.957	0.053	0.91	0.428	0.056	0.412	0.975
	12	4.25	1	49	1.53	0.050	0.84	0.562	0.053	0.023	0.974	−0.020	0.78	0.558	0.072	0.145	0.980

2.3　不同概率分布模型对比分析

对两种极值分布与当前广泛应用的 P-Ⅲ型和 GEV 模型推算的参数与分布函数拟合指标值作对比，其中，P-Ⅲ型分布参数估计使用常规矩（OME）法和线性矩（L-M）法，GEV 模型分布参数估计使用概率权重矩（PWM）法和极大似然（MLE）法。最优 GPD 模型、GEV 模型和 P-Ⅲ型模型参数和拟合拟合优度检验指标见表 2，三种分布模型推算的极值波高重现水平见表 3 和图1-图3。

表 2　最优 GPD、GEV、P-Ⅲ分布参数与拟合优度检验对比

模型				PWM						MLE			
GPD	ξ	σ	u	RMSE	Q	PPCC	ξ	σ	u	RMSE	Q	PPCC
	0.025	0.92	3.25	0.020	0.0001	0.990	0.035	0.92	3.25	0.016	0.030	0.990
GEV	ξ	σ	μ	RMSE	Q	PPCC	ξ	σ	μ	RMSE	Q	PPCC
	0.082	0.712	4.756	0.258	0.120	0.972	−0.086	0.849	4.812	0.283	0.084	0.962
模型				L-M						OME			
P-Ⅲ	ξ	σ	u	RMSE	Q	PPCC	ξ	σ	u	RMSE	Q	PPCC
	1.473	2.189	3.744	0.239	0.212	0.971	2.048	4.314	3.123	0.242	0.013	0.970

三种概率分布的拟合优度指标对比显示，超限量抽样在满足超限量数服从泊松分布，超限彼此相互独立条件下构建的 GPD 模型，其精度优于 GEV 和 P-Ⅲ型，尤以 PWM 参数估计推算的 GPD 模型为最佳。

由同频率设计值对比可见，设计频率小于 10%（重现期大于 10 年）时，GPD 模型设计值小于其余二者模型（表 3）。此说明，由 BM 模型推算的极值波高年设计值有可能偏大。

表 3 三种概率分布函数设计波高/m

T/年	GPD		P-Ⅲ		GEV	
	PWM	MLE	L-M	OME	PWM	MLE
100	8.40	8.30	8.49	8.27	8.74	8.04
50	7.85	7.77	7.93	7.79	8.03	7.63
20	7.11	7.06	7.17	7.13	7.15	7.04
10	6.53	6.51	6.57	6.59	6.52	6.55
5	5.95	5.94	5.94	6.00	5.89	6.01
2	5.16	5.16	5.01	5.07	5.02	5.12

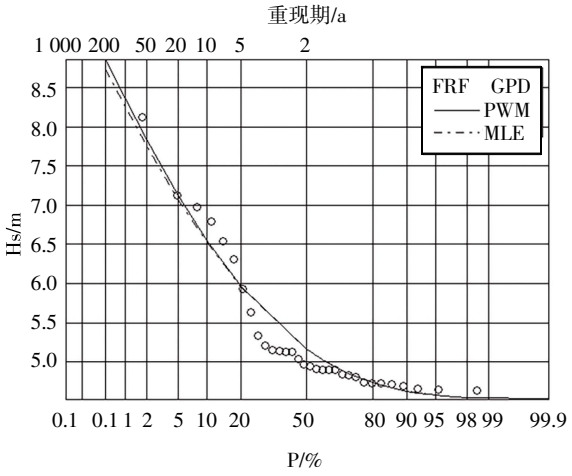

图 1 FRF 超定量波高频率曲线

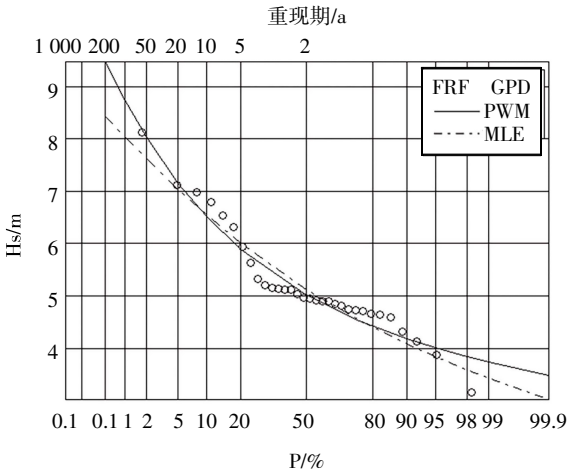

图 2 FRF 年最大波高 GEV 分布频率曲线

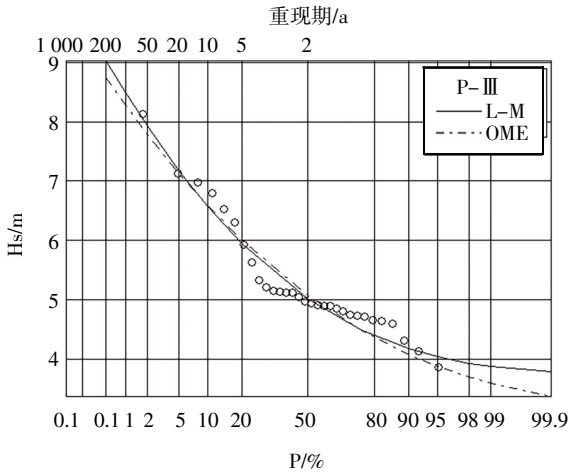

图 3 FRF 年最大波高 P-Ⅲ分布频率曲线

3 结论

对 FRF 试验场波浪数据的 GPD 分析有以下主要认识：

（1）GPD 的形态参数表明 FRF 极值波高的 GPD 模型属于右短尾型，波高有上限符合自然界的物理属性，符合海浪形成与发展理论。

（2）对存在多个满足 GPD 模型要求的阈值，需要通过对超阈值数的泊松分布检验和超阈值样本的拟合优度等做综合评判以构建最优 GPD 模型。

（3）多个概率分布模型对比表明，GPD 模型普遍优于 GEV 和 P-Ⅲ型，由于超阈值抽样能获取更多波高信息，GPD 更适用于观测年限较短站点的极值分析。

（4）极值模型的参数估计方法对于模型推算结果有较大影响。GPD 模型的 PWM 参数估计方法略优于 MLE 参数估计方法，推算的结果略大于 ML 法估计结果。

致谢：本文使用 FRF 提供的波浪观测数据，在此表示衷心感谢！

参 考 文 献

Coles S, 2001. An introduction to statistical modeling of extreme values. New York: Springer Verlag, 36-78.

Hosking J R M, Wallis J R, 1987. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 29, 339-349.

Leadbetter M R, 1991. On a basis for peaks over threshold modeling. Statistics and Probability Letters. 12(4): 357-362.

Madsen H, Rasmussen P F, Rosbjerg D, 1997. Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events, at −site modeling. Water Resources Research, 33(4): 747−757.

Rosbjerg D, Madsen H, 1992. Prediction in partial duration series with generalized Pareto distribution exceedances. Water Resources Research, 28(11): 3001−3010.

Scholz F W, Stephens M A, 1987, . K−sample Anderson−Darling tests. Journal of the American Statistical Association, 82(399): 918−924.

Van Montfort M A J, Witter J V, 1985. Testing exponentiality against generalized Pareto distribution. Journal of Hydrology, 78: 305−315.

Wang Q J, 1991. The pot model described by the generalized Pareto distribution with Poisson arrival rate. Journal of Hydrology , 129: 263−280.

岑泰林, 韦程东, 张晓东, 等, 2018. 复合 LINEX 对称损失下广义 Pareto 分布形状参数 θ 的 Bayes 估计.广西师范学院学报 (自然科学版), 35(3): 27−31.

陈子燊, 路剑飞, 于吉涛, 2017. 基于非对称 Archimedean Copula 的三变量风浪重现水平分析. 海洋通报, 36(6): 631−637.

戴昌军, 梁忠民, 栾承梅, 等, 2006. 洪水频率分析中 PDS 模型研究进展. 水科学进展, 17(1): 136−140.

方彬, 郭生练, 柴晓玲, 等, 2005. FPOT 方法在洪水频率分析中的应用研究. 水力发电, 31(2): 9−12.

王剑峰, 宋松柏, 2010. 广义 Pareto 分布在超定量洪水序列频率分析中的应用. 西北农林科技大学学报 (自然科学版), 38 (2): 191−196.

王善序, 1999. 洪水超定量系列频率分析. 人民长江, 30(8): 23−25.

颜亦琪, 易建军, 孙华安, 2010. 泊松分布在水文频率计算中的应用. 人民长江, 41(12): 92−94.

（本文编辑：王少朋）