二项–二维对数正态分布及其在极端海况预测中的应用

丁京华[1], 付殿福[2], 庞 亮[1*]

（1. 中国海洋大学 工程学院, 山东 青岛 266100；
2. 中海石油（中国）有限公司北京研究中心, 北京 100028）

摘 要：极值分析是进行海洋灾害概率预测和海洋工程工况计算的必要方法。样本资料缺乏会影响极值分析结果的合理性。阈值法和基于复合极值分布理论的过程取样法是扩充样本的有效方法，但依然依赖于长期海况资料。为了构建一种利用短期海况资料代替传统年极值序列的极值模型，本文将阈值取样法和复合极值理论结合，建立了基于短期海况资料的二项–二维对数正态复合极值分布模型。该模型既考虑了极端海况的出现频次，又能反映不同海况要素间的相关性，同时降低了对资料序列长度的要求。将该模型应用于黄海某区域设计波浪要素的计算，结果表明二项–二维对数正态分布对短期海况序列具有良好的拟合效果，与采用长期数据所得概率预测结果相差不大，能够合理反映极端海况的概率分布特征。该模型可在海洋水文资料匮乏条件下为海岸工程设计提供可靠依据，并适用于防灾减灾领域极值预测计算。

关键词：复合极值分布；对数正态分布；极端海况；短期数据；概率预测
中图分类号：P731.33 **文献标志码：**A **文章编号：**1002-3682（2023）02-0140-09
doi: 10.12362/j.issn.1002-3682.20220711002
引用格式：丁京华, 付殿福, 庞亮. 二项–二维对数正态分布及其在极端海况预测中的应用[J]. 海岸工程, 2023, 42（2）：140-148. DING J H, FU D F, PANG L. Binomial-bivariate log-normal compound model and its application in extreme sea condition prediction[J]. Coastal Engineering, 2023, 42（2）：140-148.

我国领海面积达 300 万 km^2，海岸线长约 1.8 万 km[1]，具有良好的海洋开发前景。沿海地区作为海洋开发的前沿和基地，同时也容易受到热带气旋、寒潮大风等自然灾害的侵袭。上述极端海况通常可以用极值理论描述。利用实测海洋水文数据或经验证的后报资料[2]，选择相符的极值模型进行拟合分析，就可以计算不同重现期的海洋要素设计值，作为工程设计依据。

在此类研究中，最常采用的模型是基于长期资料年极值取样的极值分布，如 Gumbel 分布、Weibull 分布及对数正态分布等。但传统极值模型受限于样本序列的长度，某些海区水文数据匮乏，传统模型的拟合结果往往无法满足工程要求。尤其在海洋工程设计中，更受关注的是高分位数问题。相同置信水平下样本容量越大置信区间越小，结果越精准。传统的极值理论采用的年极值取样方式舍弃了很多有用的信息，往往造成高分位数计算结果的误差。采用阈值取样（Peak Over Threshold, POT）的 GPD（Generalized Pareto Distribution）模型和采用过程取样的复合极值分布（Compound Extreme Value Distribution, CEVD）模型能够有效扩充样本序列，但仍然依赖于长期（通常不少于 20 a）观测或后报资料。

收稿日期：2022-07-11
资助项目：中国海洋石油集团有限公司课题（20210384）
作者简介：丁京华（1998—），男，硕士研究生，主要从事极端海洋环境预测方面研究.
E-mail：djh2020Tom@163.com
*** 通信作者：**庞 亮（1980—），男，副教授，主要从事极端海洋环境预测及风险评估方面研究.
E-mail：pang@ouc.edu.cn

（李 燕 编辑）

同时，海洋工程结构面临的是多种要素共同作用的复杂环境，并且多种海洋环境要素之间具有复杂的相关性[3]，多元极值模型逐渐成为发展的趋势。1967 年，Gumbel 和 Mustafi[4] 提出了多元极值理论——对称 Logistic 模型。此后 Galambos[5]、Leadbetter 等[6] 学者从概率统计、随机序列过程等方面将该理论进行了系统总结。在 Feller[7] 提出的复合分布概念基础上，1979 年，马逢时和刘德辅[8] 提出将离散分布与连续型极值分布结合，组成复合极值分布模型，得到了推求海洋要素设计值的新方法，在工程界得到了普遍的认可。21 世纪初，刘德辅等又将该理论由一维扩展到二维、多维，具有广泛的应用意义[9-10]。2008 年，刘德辅等[11] 提出双层嵌套概率模式，将多维复合极值分布与后报模型相结合，解决极值预测中资料序列匮乏的问题。

综上所述，在极端海况概率预测领域，合理扩充样本序列和不同环境要素合理组合是两个关键问题。本文在复合极值分布理论的基础上，提出基于短期数据的二项−二维对数正态复合极值分布模型（Binomial-Bivariate Log-normal Compound Extreme Value Distribution，BBLCED），采用阈值法对极端海况取样，同时考虑不同海洋环境要素的相关性，可以用于求解不同重现期下海况要素组合设计值。

1　二项−二维对数正态复合模型

极值理论中常用的两种取样模型分别为 BMM（Block Maximum Method）和 POT（Peaks Over Threshold），这两者都只关注尾部分布。由于 BMM 划分区域时的偶然性，可能会忽略掉一些有效的信息，而 POT 则有效地使用了极端观察值，具有更良好的统计推断[12-13]。2001 年，Liu 和 Li[14] 建立的二项−对数正态分布是从日极值序列中简单挑选符合对数正态分布的样本，不确定性较大。为减少取样不确定性，本文将 POT 模型筛选超阈值得到的二项分布与二维对数正态模型进行组合，构建二项−二维对数正态复合模型。

1.1　二维复合极值分布

设(ξ, η)为二维连续型随机变量，(ξ, η)的联合概率分布函数记作$G(x, y)$，联合概率密度函数为$g(x, y)$，其中ξ的分布函数记作$G_x(x)$。记(ξ_i, η_i)为(ξ, η)的第i次观测值，设N为与(ξ, η)独立的取值为非负的整数的随机变量，其分布函数记作：

$$P(n=k)=p_k \qquad k=0,1,\cdots,\infty, \tag{1}$$

$\sum_{k=1}^{\infty} p_k = 1$。

定义随机变量(X, Y)：

$$(X, Y)=(\xi_i, \eta_i)|\xi_i=\max_{0\leqslant i\leqslant N}\xi_i \qquad N\geqslant 1, \tag{2}$$

(X, Y)通常作为观测数据中的台风过程极值序列。

当采用 POT 法进行(X, Y)取样时，得到

$$(X, Y)=(\xi_i, \eta_i)\,|\,\xi_i=\underset{0\leqslant i\leqslant N}{\mathrm{POT}}\xi_i \qquad N\geqslant 1, \tag{3}$$

经刘德辅等[9] 证明，(X, Y)的联合分布函数为：

$$F(x, y)=\sum_{k=1}^{\infty} P_k\cdot k\cdot\int_{-\infty}^{y}\int_{-\infty}^{x} G_x^{k-1}(u)\cdot g(u, v)\mathrm{d}u\mathrm{d}v。 \tag{4}$$

1.2　二项−二维对数正态分布

假设m天的观测数据为独立同分布的随机变量(ξ_i, η_i)（$i=1, 2, \cdots, m$），对于一个充分大的阈值μ，假设有n个数据超过阈值，满足$X_j>\mu$（$j=1, 2, \cdots, n$），则称X_j为超阈值，X_j和对应的Y_j组成超阈值序

列(X_j, Y_j)。超过μ的样本序列长度为k。n服从参数(m, p)的二项分布，$P(n=k)=C_m^k\bar{p}^k(1-\bar{p})^{m-k}$，$\bar{p}$为$p_k$的平均值。

年内出现超阈样本次数n符合二项分布，式(4)转换为以下形式：

$$F_0(x,y)=(1-p)^n+\sum_{k=1}^{\infty}\binom{n}{k}\bar{p}^k(1-p)^{n-k}\cdot k\cdot\int_{-\infty}^{y}\int_{-\infty}^{x}G_x^{k-1}(u)\cdot g(u,v)\mathrm{d}u\mathrm{d}v, \tag{5}$$

对式(5)求导得

$$f(x,y)=\sum_{k=1}^{\infty}\binom{n}{k}\bar{p}^k(1-p)^{n-k}\cdot k\cdot G_x^{k-1}(u)\cdot g(x,y), \tag{6}$$

又有 (X_j,Y_j) 服从二维对数正态分布如下：

$$g(x,y)=\frac{1}{2\pi\sqrt{1-\rho_{xy}^2}\sigma_x\sigma_y xy}\exp\left[\frac{-1}{2\left(1-\rho_{xy}^2\right)}\left(\frac{(\ln x-\mu_x)^2}{\sigma_x^2}-\frac{2\rho_{xy}(\ln x-\mu_x)(\ln y-\mu_y)}{\sigma_x\sigma_y}+\frac{(\ln y-\mu_y)^2}{\sigma_y^2}\right)\right]\quad(x>0,\ y>0), \tag{7}$$

式中：μ_x、μ_y、σ_x、σ_y、ρ_{xy}均为常数；μ_x、μ_y分别为X_j、Y_j序列的均值；σ_x、σ_y分别为X_j、Y_j序列的标准差；ρ_{xy}为X_j、Y_j序列的相关系数。

对数正态的边缘分布：

$$G_x(u)=\int_0^x\frac{1}{\sqrt{2\pi}\sigma_x x}\cdot\mathrm{e}^{-\frac{(\ln x-\mu_x)^2}{2\sigma_x^2}}\mathrm{d}u, \tag{8}$$

将式(7)、式(8)代入式(5)中即得二项−二维对数正态分布。

2 算 例

2.1 阈值选取

阈值的合理选取是 POT 模型成功拟合的关键[15]。将高于阈值的波高数据及对应(伴随)的周期数据作为二项−二维对数正态复合模型的样本(X_j,Y_j)。

本文以黄海某海洋站 1992—1996 年日最大波高及周期数据为初始序列，运用 Hill 图法选取阈值，选取原则为观察尾部指标α，找到图形尾部相对稳定的线段作为起始点，则将该点的横坐标对应的数据作为阈值。绘制 Hill 图，如图 1 所示。

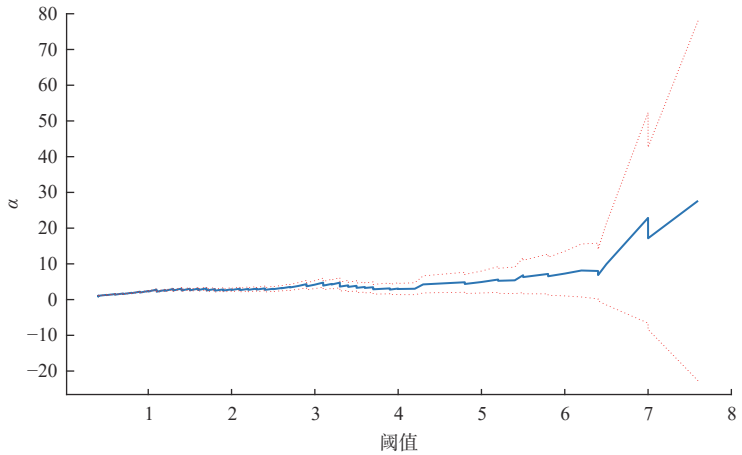
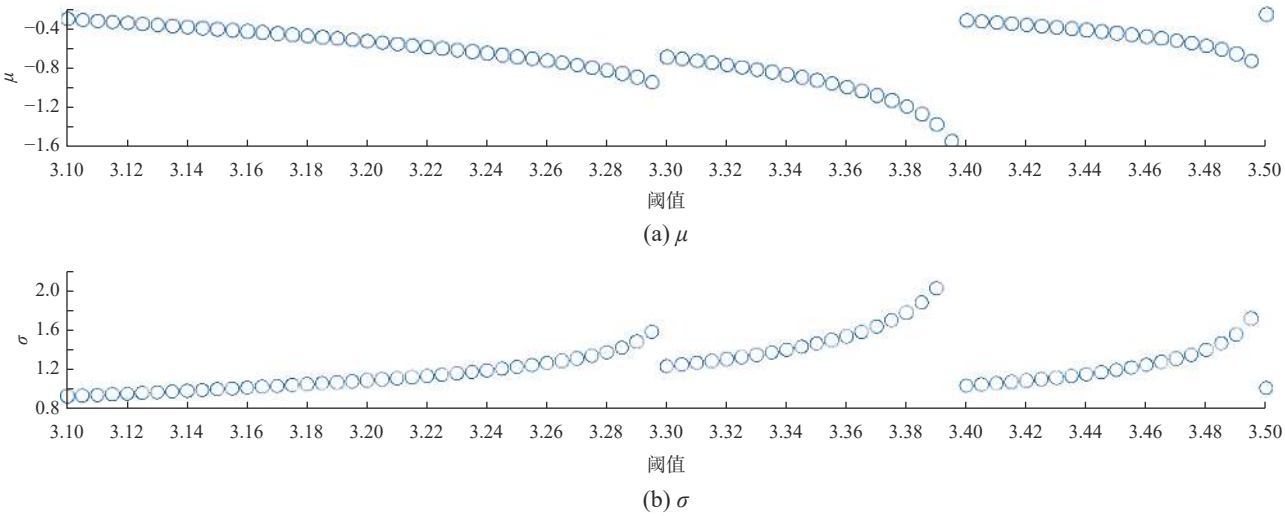


图 1 Hill 图

Fig. 1 Hill plot

由图 1 可以看出，中尾部首次达到平稳状态时横坐标对应的值约为 3.3。为了检验阈值是否合理，结合参数稳定性估计图进一步判断。判断方法为：在 Hill 图确定的阈值范围内，以规定的步长选取若干点进行参数μ(波高超阈值序列均值)和σ(波高超阈值序列方差)的试算，选取令 2 个参数的极大似然估计值达到稳定的最大值作为最终确定的阈值。依据前文的计算，将 μ=3.3 作为初始阈值，在(3.1，3.5)之间均匀选取 81 个值作为阈值，观察计算得到的 2 个参数的极大似然估计值在区间内的稳定性。在上述阈值的选取范围内，参数的变化趋势图如图 2 所示。



(a) μ

(b) σ

图 2　不同波高阈值下参数μ、σ的最大似然估计值

Fig. 2　The maximum likelihood estimates of parameters μ, σ at thresholds of different wave heights

由图 2 可以看出，各参数均在(3.100, 3.295)阈值内保持相对稳定，为确保 POT 模型拟合的准确性，本文选取相对稳定区间内的较大值作为最终阈值，即μ=3.295。将大于阈值的波高数据提取出来，并提取出这些波高数据伴随的周期数据。

2.2　参数估计

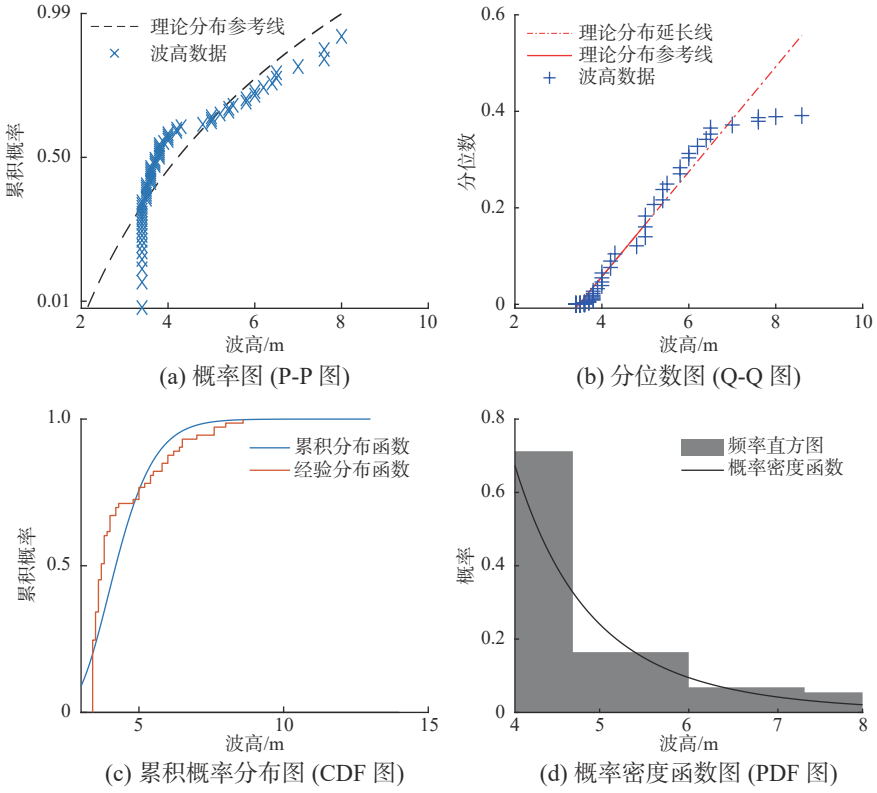
确定阈值后，利用极大似然估计法对二项−二维对数正态复合模型中的各个统计参数进行估计，结果见表 1。将表 1 中参数代入式(4)中，可得日最大波高周期联合分布函数。

表 1　二项−二维对数正态分布复合模型统计参数

Table 1　Statistical parameters of BBLCED model

统计参数	μ_x	μ_y	σ_x	σ_y	ρ_{xy}	p
估计值	1.422 3	1.904 8	0.255 2	0.137 9	0.767 7	0.009 0

图 3、图 4 分别为对数正态分布模型拟合上一步中得到的超阈值序列的波高和周期数据诊断图。由 P-P 图(图 3a、图 4a)和 Q-Q 图(图 3b、图 4b)看出，实测数据的经验分布较为贴近理论分布。CDF 图(图 3c、图 4c)和样本的累计分段函数基本一致，PDF 图(图 3d、图 4d)和样本的频率直方图也是吻合的。因此 4 个诊断图都支持拟合的对数正态分布模型，即该模型拥有较好的拟合效果。

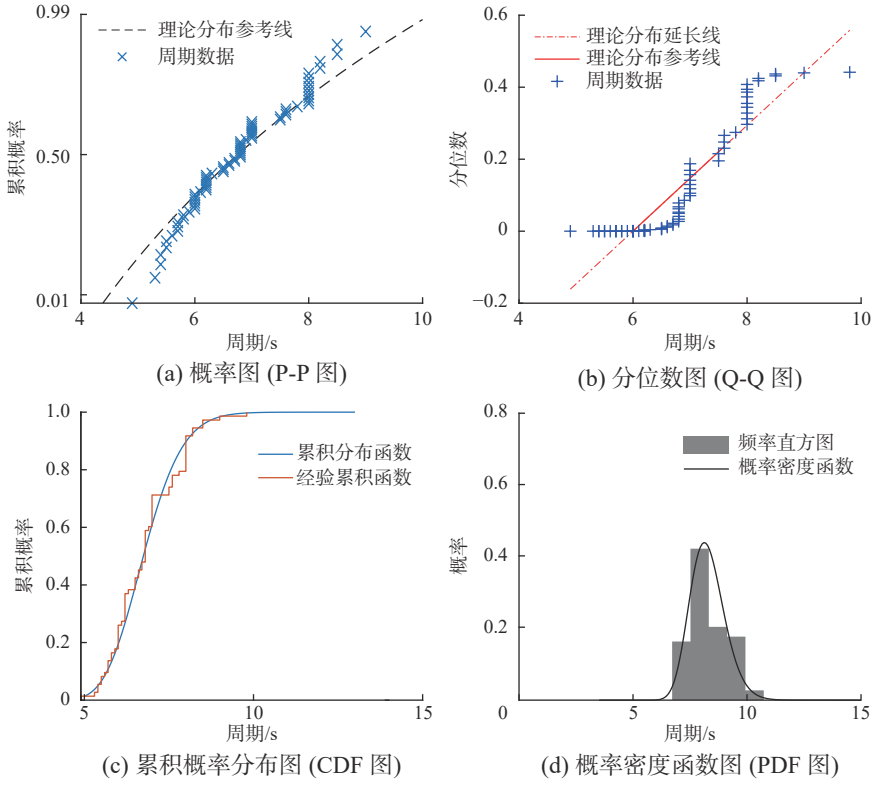


(a) 概率图 (P-P 图) (b) 分位数图 (Q-Q 图)

(c) 累积概率分布图 (CDF 图) (d) 概率密度函数图 (PDF 图)

图 3 波高数据诊断结果

Fig. 3 Diagnostic results of wave height data



(a) 概率图 (P-P 图) (b) 分位数图 (Q-Q 图)

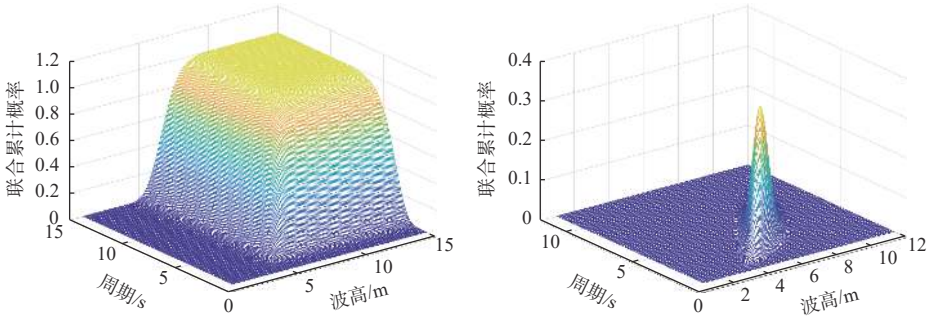
(c) 累积概率分布图 (CDF 图) (d) 概率密度函数图 (PDF 图)

图 4 周期数据诊断结果

Fig. 4 Diagnostic results of wave period data

2.3　多年一遇波高与周期预测以及方法比较

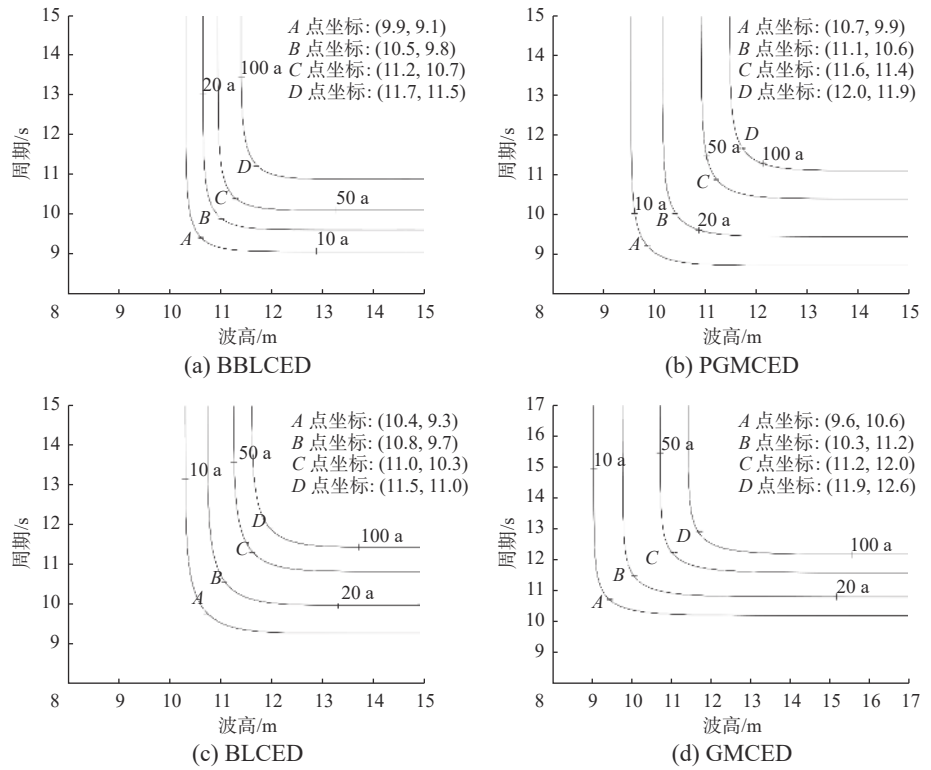
将前文求得的模型参数，代入到二项−二维对数正态复合模型计算多年一遇波高和周期设计值。另外，选用泊松−混合冈贝尔模型(Poisson-Gumbel Mixed Compound Extreme Value Distribution，PGMCED)计算所得结果作为复合极值分布的对照。同时，基于黄海某区域 20 a 的年极值波高和周期数据资料，选用混合冈贝尔模型(Gumbel Mixed Extreme Value Distribution，GMED)、对数正态模型(Bivariate Log-normal Extreme Value Distribution，BLED)计算多年一遇波高和周期设计值作为传统方法的对照。从各模型的联合累积分布图的概率值为 0.90、0.95、0.98、0.99(分别对应 10 年一遇、20 年一遇、50 年一遇、100 年一遇)的等高线图提取出概率密度最大的 4 个点，分别记作 *A*、*B*、*C*、*D*，这 4 个点对应的波高与周期值即为推求的波高与周期联合设计值。这 4 种模型所得的结果如图 5 和图 6 所示，统计结果如表 2 所示。



(a) BBLCED 累积分布　　(b) BBLCED 概率密度

图 5　波高和周期联合概率累积分布图、波高和周期联合概率密度分布图

Fig. 5　Cumulative distribution and density distribution plots of wave height and period joint probability



(a) BBLCED　　(b) PGMCED

(c) BLCED　　(d) GMCED

图 6　4 种模型在不同重现期下的波高和周期的设计值

Fig. 6　The wave height and period values designed by using 4 models under different return periods

表 2 4 种模型的不同重现期下的波高和周期重现水平

Table 2 Wave height and period recurrence levels obtained by 4 models under different return periods

计算模型	10 年一遇		20 年一遇		50 年一遇		100 年一遇	
	波高/m	周期/s	波高/m	周期/s	波高/m	周期/s	波高/m	周期/s
BBLCED	10.4	9.3	10.8	9.7	11.0	10.3	11.5	11.0
PGMCED	9.9	9.1	10.5	9.8	11.2	10.7	11.7	11.5
BLED	10.7	9.9	11.1	10.6	11.6	11.4	12.0	11.9
GMED	9.6	10.6	10.3	11.2	11.2	12.0	11.9	12.6

根据表 2 可以看出，二项−二维对数正态复合模型下的不同重现期的波高重现水平与其他 3 种模型差距不大，相对来说结果更接近二维对数正态模型拟合结果。而该模型的不同重现期的周期重现水平与其他 3 种模型相比结果偏小，但差异很小，相对来说其更接近泊松−混合冈贝尔模型拟合结果。而本文所建立的二项−二维对数正态复合模型选用了 5 a 的海况资料，可以与选用 20 a 海况资料的其他 3 种模型达到类似的效果，即为该模型最大的优点。

2.4 稳定性分析

本文提出滑动重现期这一概念：在固定序列长度内，按样本顺序依次改变序列起始值，得到一系列重现期结果。从 1991—1998 年的海况资料滑动选取 5 a 的数据用二项−二维对数正态模型推求 50 年一遇的波高和周期，结果如表 3 所示。

由表 3 可以看出，波高的滑动重现期水平较为平稳，且波动都在 0.5 以内，而周期的滑动重现期整体也是较为平稳的，不确定性较小，所以该模型是适用的。

表 3 不同时间段的 50 年重现期的波高和周期水平

Table 3 Wave height and period levels of 50-year return period under different years

时间段	50 年一遇波高/m	50 年一遇周期/s
1991—1995 年	11.4	11.9
1992—1996 年	11.0	10.3
1993—1997 年	11.0	10.3
1994—1998 年	11.0	10.3

3 结 论

本文基于 POT 模型筛选出超阈值样本序列，其年内出现超阈值的次数符合二项分布。对极端海况出现时的波高及其“伴随”的周期数据分别进行边缘分布拟合，发现均很好地符合对数正态分布，由此建立采用 POT 取样的二项−二维对数正态复合模型，用于短期波高及其伴随的周期数据的概率预测，并得出以下主要结论。

运用该模型对黄海某海洋站 5 a(1992—1996 年)的波高和周期数据进行实例分析，模型拟合结果较好；分别预测了不同重现期下的波高和周期，并自定义了滑动重现期对其验证。结果显示：二项−二维对数正态分布对短期海况资料具有良好的拟合效果，与采用长期海况资料所得概率预测结果相差不大，能够合理反映极端海况的概率分布特征，且在 1991—1998 年的滑动重现期水平较为平稳，即二项−二维对数正态复合模型在短期海况预测方面是适用的。因此，用 5 a 的序列构造出的基于短期海况资料的极值分布的结果能代替使用 20 a 长期海况资料的传统极值方法。

同时还发现，现有的数据很难达到完全精确，并且模型本身也存在一定的不准确性，导致各种复杂的统计模型也会产生误差，因此没有一种方法是可以在所有情况下适用的。

本文提出的二项−二维对数正态复合极值模型基于阈值法取样，能够充分利用有效数据，从而

降低了对资料序列长度的要求。同时 BBLCED 考虑了极端海况出现频次与各要素间的相关性，可为工程设计提供更为合理可靠的依据。此外，该模型也适用于各类工程设计和防灾减灾领域极值预测的计算。

参考文献 (References):

[1] 韩婕妤. 中国近岸海域环境质量演变及驱动因素研究[J]. 资源开发与市场, 2019, 35(9): 1133-1137. HAN J Y. Study on the evolution and driving factors of environmental quality in China's coastal waters[J]. Resource Development & Market, 2019, 35(9): 1133-1137.

[2] YAN Z D, WANG Z L, PANG L. Study of the wind conditions in the South China Sea and its adjacent sea area[J]. Journal of Ocean University of China, 2022, 21(2): 264-276.

[3] CHENG Y J, YAN Z D, PANG L, et al. Probability analysis on the typhoon induced sea states of the South China Sea[C]//New York: 2018 OCEANS-MTS/IEEE KOBE TECHNO-OCEANS (OTO), 2018.

[4] GUMBEL E J, MUSTAFI C K. Some analytical properties of bivariate extremal distributions[J]. Publications of the American Statistical Association, 1967, 62(318): 569-588.

[5] GALAMBOS J. The asymptotic theory of extreme order statistics[M]. New York: John Wiley & Sons, Inc., 1978.

[6] LEADBETTER M R, LINDGREN G, ROOTZEN H. Extreme and related properties of random sequences and series[M]. New York: Springer, 1983.

[7] FELLER W. An introduction to probability theory and its applications[M]. New York: John Wiley & Sons, Inc., 1968.

[8] 马逢时, 刘德辅. 复合极值分布理论及其应用[J]. 应用数学学报, 1979, 2(4): 366-375. MA F S, LIU D F. The theory of compound extreme value distribution and its application[J]. Acta Mathematicae Applicatae Sinica, 1979, 2(4): 366-375.

[9] 刘德辅, 温书勤, 王利萍. 泊松-混合冈贝尔复合极值分布及其应用[J]. 科学通报, 2002, 47(17): 1356-1360. LIU D F, WEN S Q, WANG L P. Poisson-Gumbel mixed compound extreme value distribution and its application[J]. Chinese Science Bulletin, 2002, 47(17): 1356-1360.

[10] 刘德辅, 董胜. 随机工程海洋学[M]. 青岛: 中国海洋大学出版社, 2004: 98-117. LIU D F, DONG S. Stochastic engineering oceanography[M]. Qingdao: China Ocean University Press, 2004: 98-117.

[11] 刘德辅, 庞亮, 谢波涛, 等. 中国台风灾害区划及设防标准研究——双层嵌套多目标联合概率模式及其应用[J]. 中国科学: E辑 技术科学, 2008, 38(5): 698-707. LIU D F, PANG L, XIE B T, et al. Study on typhoon disaster zoning and fortification criteria in China: double nested multi-objective joint probability model and its application[J]. Science in China: Series E Technological Sciences, 2008, 38(5): 698-707.

[12] 刘莎莎. POT模型阈值的选取及应用[D]. 吉林: 吉林大学, 2014. LIU S S. The selection and application of the threshold of POT model[D]. Jilin: Jilin University, 2014.

[13] YAN Z D, PANG L, DONG S. Analysis of extreme wind speed estimates in the northern South China Sea[J]. Journal of Applied Meteorology and Climatology, 2020, 59(10): 1625-1635.

[14] LIU D F, LI H J. Prediction of extreme significant wave height from daily maxima[J]. China Ocean Engineering, 2001, 15(1): 97-106.

[15] 孙丽莉. 金融数据的极端风险度量及应用[D]. 重庆: 重庆大学, 2014. SUN L L. Measurements and application of the extreme risk of financial data[D]. Chongqing: Chongqing University, 2014.

Binomial-Bivariate Log-normal Compound Model and Its Application in Extreme Sea Condition Prediction

DING Jinghua[1], FU Dianfu[2], PANG Liang[1]

（1. *College of Engineering*, *Ocean University of China*, Qingdao 266100, China；
2. *CNOOC China Limited*, Beijing 100028, China）

Abstract: Extreme value analysis is a necessary method for predicting the probability of marine disasters and calculating the conditions of marine engineering. However, the rationality of extreme value analysis results can easily be affected by the lack of sample data. Although the peaks over threshold (POT) method and the process sampling method based on the compound extreme value distribution (CEVD) theory are effective methods to expand samples, they still rely on long-term sea condition data. In order to construct an extreme value model which uses short-term sea condition data instead of traditional annual extreme value series, a Binomial-Bivariate Log-normal Compound Extreme Value Distribution (BBLCED) model which is based on short-term data is established by combination of POT and CEVD. This model not only considers the frequency of extreme sea conditions, but also can reflect the correlation between different sea condition elements and reduce the requirement for the data series length. The model has been applied to the calculation of design wave elements in a certain area of the Yellow Sea and the results indicate that the BBLCED model has a good fitting to the short-term sea condition sequence. There is not much different from the probability prediction results obtained by using the long-term data, indicating that the model can reasonably reflect the probability distribution characteristics of the extreme sea conditions. This model can provide a reliable basis for the coastal engineering design under the condition of lack of marine hydrological data, and it is also suitable for the extreme value prediction calculation in the field of disaster prevention and mitigation.

Key words: compound extreme value distribution；Log-normal distribution；extreme sea condition；short-term data；probability prediction

Received: July 11, 2022