第 59 卷 第 10 期 2010 年 10 月
1000-3290/2010/59(10)/7491-07

物 理 学 报
ACTA PHYSICA SINICA

Vol. 59, No. 10, October 2010

极端事件再现时间长程相关性与群发性研究*

王启光[1)2)3)] 侯 威[2)3)] 郑志海[1)] 冯爱霞[1)] 邓北胜[1)†]

1)(兰州大学大气科学学院,兰州 730000)

2)(中国科学院大气物理研究所东亚区域气候-环境重点实验室,北京 100029)

3)(国家气候中心,中国气象局气候研究开放实验室,北京 100081)

(2009 年 9 月 16 日收到;2009 年 12 月 14 日收到修改稿)

利用百分位阈值方法定义极端事件,从极端事件再现时间的角度,研究了极端事件发生时间间隔的长程相关性.发现若原时间序列具有长程相关性,则它的极端事件再现时间序列也具有长程相关性;计算表明两者的标度指数 α 相当接近,这一特性与随机产生的再现时间序列有着本质的差别,再现时间序列的长程相关性是由原序列的长程相关性决定的.具有长程相关性的时间序列再现时间的概率分布明显不同于随机序列,其小值再现时间的概率较大,反映出极端事件的群发现象.本文根据这一特征定义了再现时间的群发性指数,发现时间序列的长程相关性是导致极端事件群发性的原因.

关键词:再现时间,长程相关性,群发性,极端事件

PACC:9260X

1. 引 言

极端事件作为人们十分关注的问题,其传统的基本假设是时间序列中的各事件之间是无关联的[1],即假设极端事件之间是相互独立的,前一事件对后面发生的事件毫无影响,是完全随机的,并且遵循 Poisson 分布[2].这种情况下极端事件的统计规律仅仅取决于原序列的概率密度函数,在以往进行预测研究的时候完全采用随机模拟的方法[3].事实上,洪涝、地震和经济学等领域的极端事件并非是随机的,而是呈现一定的规律性[4—7],即极端事件序列具有长程相关性的特征,通过混沌序列分析发现这种长程相关性和极端事件的可预测性存在较好的定性关系[8],并且发现极端天气气候事件的长程相关性具有区域分布的特征[9].事实上对极端事件的描述,不仅有发生的频次、强度,还有十分关注的科学问题——重大异常的极端事件下次何时发生,即极端事件的再现时间也是一个值得研究的问题,对其研究有助于加深对极端事件发生规律的了解.

本文采用百分位阈值方法定义极端事件,对极端事件发生的再现时间进行研究,探讨极端事件再现时间序列的长程相关性、概率分布和群发性之间的联系.同时本文定义了群发性指数,研究了其与阈值之间的关系.

2. 极端事件再现时间和研究方法

所谓极端事件是指小概率事件.目前对于极端事件的阈值非参数确定方法应用较多的是 Bonsal 提出的一种百分位值方法[10—12]:若某一气象要素有 n 个值,先将这 n 个值按升序排列 $x_1, x_2, \cdots, x_m, \cdots x_n$,某个值小于或等于 x_m 的概率 P 为,

$$P = (m - 0.31)/(n + 0.38). \quad (1)$$

假如有 68 个值,那么第 90 个百分位上的值为排列后的 x_{61}($P = 88.8\%$)和 x_{62}($P = 90.2\%$)间的线性插值.

对于时间序列$\{x(j), j = 1, 2, 3\cdots, N\}$在给定的百分位阈值 q 条件下,就可以构建一条对应的极端事件序列,定义两次相邻极端事件发生的时间间隔为极端事件再现时间 $r_q(i)$($i = 1, 2, 3, \cdots, N_q$)(如图

* 国家自然科学基金(批准号:40875040,40775048)、科技部支撑计划(批准号:2007BAC29B01)和公益性行业(气象)科研专项基金(批准号:GYHY 200806005)资助的课题.

† 通讯联系人. E-mail:dengbeisheng@hotmail.com

1 中 r_1, r_2, r_3). 为了便于比较和探讨极端事件的分布规律,本文分别定义 q 取百分位 85,90,95 极端事件,并记为 E_{85}, E_{90} 和 E_{95},且极端事件再现时间序列分别记为 $\{R_{85}(i)\}, \{R_{90}(i)\}, \{R_{95}(i)\}, (i = 1,2,3\cdots)$.

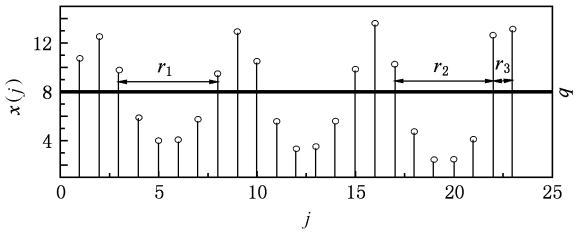

图 1 极端事件再现时间 r_1, r_2, r_3

由图 1 可知,一旦百分位阈值 q 确定,相应的极端事件数目 N_q 也就确定. 当 $N \to \infty$ 时,有 $\sum_{i=1}^{N_q} r_q(i) \approx N$,极端事件序列的平均再现时间 R_q 定义为,

$$R_q = \frac{1}{N_q}\sum_{i=1}^{N_q} r_q(i) \cong \frac{N}{N_q}, \tag{2}$$

在样本量 N 一定的情况下,如果平均再现时间 R_q 越小,说明发生的极端事件的数目越多. 若 $P(x)$ 为 $\{x(j), j = 1,2,3\cdots,N\}$ 的概率密度分布函数,则 R_q 表示为,

$$R_q \cong 1/\int_q^{\infty} P(x)\,\mathrm{d}x. \tag{3}$$

由此可见,R_q, q 和 $P(x)$ 具有紧密的关系,其实质是极端事件的阈值如何确定,如何真正揭示极端事件的分布规律,这个问题一直是极值理论关注的热点之一.

去趋势波动分析(DFA)方法目前是研究非线性时间序列长程相关性存在与否的工具,已广泛应用到自然科学甚至社会科学各个领域. 其具体计算方法如下[13—20]:

首先,计算时间序列 $\{x_i, i = 1, 2, \cdots, N\}$ 的累计离差,

$$Y(i) = \sum_{i=1}^{N} (x_i - \bar{x}), \tag{4}$$

其中 $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. 把 $Y(i)$ 分成 N_s 个不重叠的等间隔 s 的区间 ν,其中 $N_s = [N/s]$.

然后,对于每个区间 ν,用最小二乘法拟合数据,得到局部趋势. 滤去该趋势后的时间序列记为 $Y_s(i)$,表示原序列与拟合值之差,

$$Y_s(i) = Y(i) - P_\nu(i), \tag{5}$$

其中 $P_\nu(i)$ 为第 ν 区间的拟合多项式. 计算每个区间滤去趋势后的方差,

$$F^2(\nu, s) = \frac{1}{s}\sum_{i=1}^{s} Y_s^2[(\nu - 1)s + i] \quad (\nu = 1, 2, \cdots, N_s). \tag{6}$$

对所有等长度区间的方差求均值并开方,计算标准 DFA 涨落函数,

$$F(s) = \sqrt{\frac{1}{N_s}\sum_{\nu=1}^{N_s} F^2(\nu, s)}. \tag{7}$$

若时间序列是相关的,则 DFA 涨落函数 $F(s)$ 与滞后时间 s 成幂律关系

$$F(s) \propto s^{\alpha}. \tag{8}$$

在双对数坐标 $(F(s), s)$ 中,用最小二乘法拟合,其直线部分的斜率即为标度指数 α. 当 $0 < \alpha < 0.5$ 时,表示时间序列是非持久的,只有短期记忆性,当前事件不会对长期的未来事件产生影响;当 $\alpha = 0.5$ 时,表示原序列是白噪声;当 $0.5 < \alpha < 1.0$ 时,表示序列具有长程相关特征,即当前发生的事件和未来事件之间存在长程相关性,时间序列具有长期记忆性,且标度指数 α 越大记忆性越好,即时间序列可预测性也越强.

3. 再现时间长程相关性及其概率分布

许多自然事件都呈现出长期持续性的特征[13—20],即长程相关性,这种特性在水文数据、气候资料、脱氧核糖核酸(DNA)序列和心电记录,甚至在股票市场等自然科学和社会领域中都有所体现[21—25]. 已有研究表明,当原序列有长程相关性时(标度指数 $0.5 < \alpha < 1.0$),在一定阈值条件下构建的极端事件序列也有长程相关性[26]. 但极端事件再现时间组成的序列的相关特性的研究目前还涉及甚少,本文采用数值模拟的方法对这一问题做初步探讨.

3.1. 再现时间长程相关性

采用 Fourier 滤波变换的方法生成标度指数 $\alpha = 0.7, 0.8, 0.9$ 的 3 条理想时间序列[27,28],样本量为 2×10^6,其均值为 0,方差为 1. 为了进行对比研究,同时生成样本量、均值和方差都相同的 Gauss 白噪声序列.

采用百分位为 85,90,95(分别记为 85th,90th,95th)的阈值组成极端事件再现时间序列 $\{R_{85}(i)\}$,

$\{R_{90}(i)\}$，$\{R_{95}(i)\}$，($i=1,2,3\cdots$)，并研究和对比它们的统计特征(如图 2 所示).

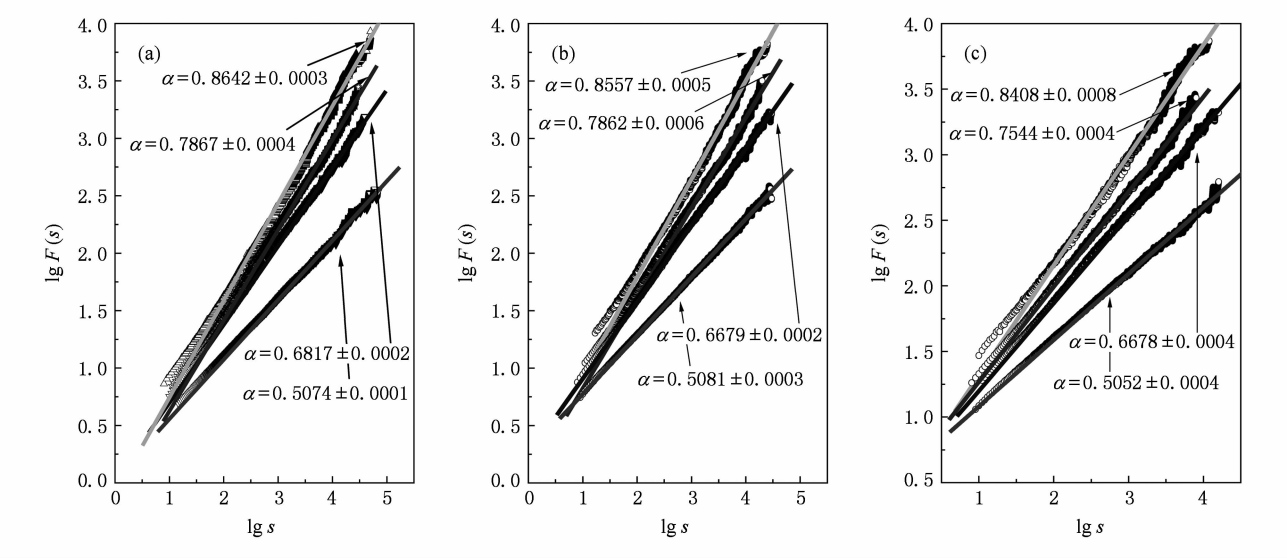

图 2　不同百分位阈值条件下再现时间序列长程相关性　(a)85th，(b)90th，(c)95th

对于理想时间序列而言，$\alpha=0.5,0.7,0.8,0.9$ 所对应的$\{R_{85}(i)\}$，$\{R_{90}(i)\}$，$\{R_{95}(i)\}$，($i=1,2,3\cdots$)标度指数 α 如表 1 所示.

表 1　理想时间序列的$\{R_{85}(i)\}$，$\{R_{90}(i)\}$，$\{R_{95}(i)\}$，($i=1,2,3\cdots$)的标度指数

	$\alpha_{\{R_{85}(i)\}}$	$\alpha_{\{R_{90}(i)\}}$	$\alpha_{\{R_{95}(i)\}}$
$\alpha=0.5$	0.507	0.508	0.505
$\alpha=0.7$	0.681	0.667	0.667
$\alpha=0.8$	0.786	0.786	0.754
$\alpha=0.9$	0.864	0.855	0.841

从表 1 中可以发现，当原序列具有长程相关性时，在确定阈值条件下的极端事件再现时间序列也存在长程相关性，并且两者的标度指数比较接近；而对于随机时间序列，其极端事件再现时间序列标度指数 α 始终为 0.5，不存在长程相关性这一特点. 由长程相关性的物理含义可知[7,12—14]，标度指数表明了数据间的关联程度，反映了系统间内禀性特征. 如果 $0.5<\alpha<1.0$，说明系统具有长程相关性，并且随着 α 的增大，系统可预测性越强. 因此极端事件的检测和预测，尤其预测极端事件的再现时间，要充分研究序列的统计特征. 为进一步揭示具有长程相关性的再现时间序列的内在规律，下面采用随机序列进行对比，研究它们概率分布特征.

由(2)和(3)式知，在样本量、均值和方差相同的情况下，只要极端事件的百分位阈值确定，构建的 4 条理想序列所得到的再现时间个数是相同的，因此对应的 85th，90th，95th 的 R_q 的理论值分别为 6.67，10.00，20.00，随阈值的增大而增大. 图 3(a)验证了理论结果的正确性，阐明了极端事件平均再现时间 R_q 与序列是否具有长程相关性无关. 因此，R_q 仅能描述极端事件总体特征，对体系的长程相关性特征，不能很好地描述，但是观察图 3(b)可以发现极端事件再现时间的最大值存在很大的差异.

比较图 3(a)和 3(b)，在相同阈值条件下，各条序列对应的再现时间的最大值 $r_q(\max)$ 差别明显. 具体表现为标度指数越大，$r_q(\max)$ 也就越大，但高斯白噪声对应的 $r_q(\max)$ 始终最小. 例如百分位阈值取 90th 时，$\alpha=0.5,0.7,0.8,0.9$ 对应的 $r_q(\max)$ 分别是 138，219，545，1716(此时平均再现时间 R_q 都是 10)，但从第 2 节分析可知，r_q 的个数对于每条理想时间序列是相同的，因此在确定的百分位阈值条件下，标度指数越大的序列，极端事件再现时间较小值所占的比例要越大. 换而言之，标度指数大，极端事件的发生会相对集中于某一时段，另一时段相对比较少，即存在群发现象. 因此本文定义 $r_q/R_q<1$ 为小值再现时间. 将各条序列小值再现时间的概率 $P(r_q<R_q)$ 进行量化比较(如图 3(c)所示)，可以明显发现随着 α 增大，$P(r_q<R_q)$ 也随之增大. 以 90th 的阈值为例，$\alpha=0.5,0.7,0.8,0.9$ 对应的 $P(r_q<R_q)$ 分别是 0.63，0.71，0.75，0.78. 在图 3(c)中还可以发现各条序列本身在不同百分位条件下对应的 $P(r_q<R_q)$ 基本相同，其中 Gauss 白噪声序列的 $P(r_q<R_q)$ 相对于其他 3 条序列明显要小得多. 因此，

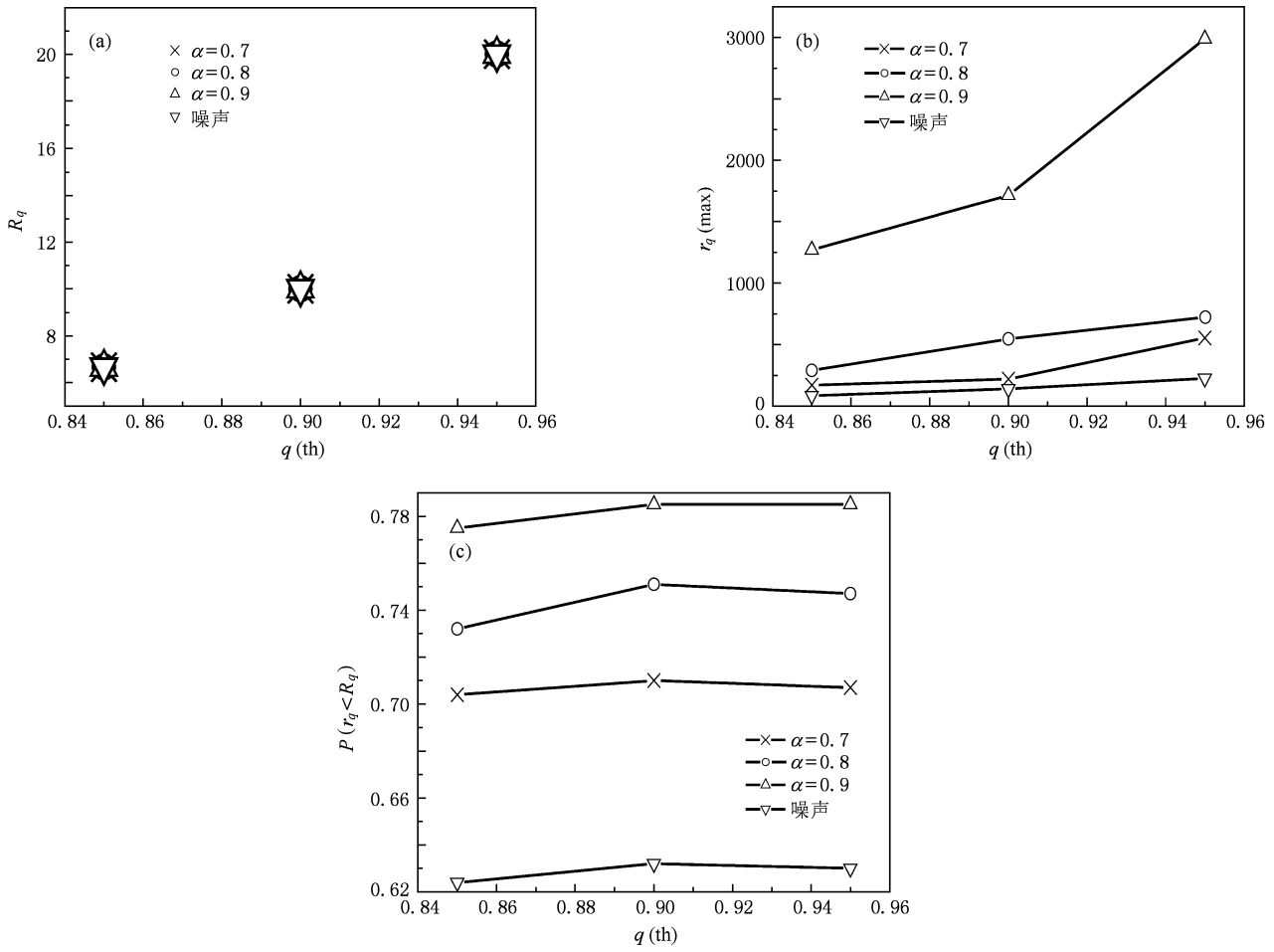

图 3 4 条理想序列不同百分位阈值条件下平均再现时间、最大再现时间以及小值再现时间概率 (a)平均再现时间 R_q,(b)最大再现时间 $r_q(\max)$,(c)小值再现时间的 r_q 所占概率

α 决定 r_q 的疏密程度,即 $\alpha=0.5$ 时,时间序列为随机序列,r_q 几乎平均分布在整个时间轴上,R_q 是一个很好的量度;当 $0.5<\alpha<1$ 时,系统具有长程相关性,极端事件的发生具有群发性,具有明显的阶段性特征,即某一时段极端事件集中发生,而另一个时段为间隙期.

3.2. 再现时间概率分布

4 条理想时间序列 85th,90th,95th 阈值条件下再现时间概率分布如图 4 所示.

当 $\alpha=0.7,0.8,0.9$ 时,在不同百分位阈值条件下的概率分布与 Gauss 白噪声随机序列存在明显差异. 在单对数坐标中,随机序列的再现时间分布符合 Poisson 分布,而具有长程相关性的序列概率分布随着标度指数的增大,偏离随机序列分布情况就越明显,因此传统的基于随机理论的极端事件发生时间概率预测理论存在一定的缺陷. 具有长程相关性的序列在相同百分位阈值条件下,小值再现时间出现的概率 $P(r_q<R_q)$ 比随机序列大得多(图 4). 同时当具有长程相关性序列在某一百分位阈值条件下再现时间较大时($r_q/R_q>4$),其分布概率也大于随机序列. 这充分说明了对于具有长程相关性的系统,其再现时间有向两端(极小或极大)聚集的倾向,即此时百分位阈值条件下极端事件相对更易集中发生.

4. 再现时间的群发性

从上节分析可知,小值再现时间($r_q/R_q<1$)的出现分布概率约比大值的再现时间($r_q/R_q>4$)概率约大两个数量级,因此本文着重研究极端事件的小值再现时间群发情况. 为探讨长程相关性和群发性间的关系,我们采用随机扰动方法将时间序列的长

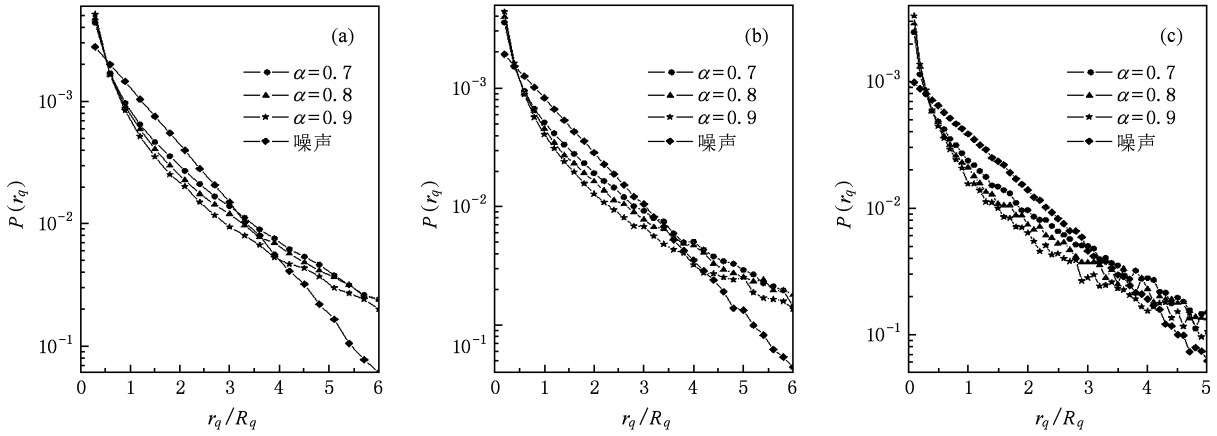

图 4　不同百分位阈值条件下理想时间序列再现时间概率分布　(a)85th,(b)90th,(c)95th

程相关性破坏,即在不改变序列数据量和数值的情况下,将原序列的排列顺序变成随机排列,研究此时时间序列的再现时间概率分布.

以 $\alpha=0.8$ 的时间序列为例,当随机扰动重新排列后,时间即变成随机序列,此时其标度指数值变为 0.5. 这表明时间序列长程相关性被破坏,扰动后的时间序列在样本量和数值不变的情况下,变为随机时间序列. 计算表明,扰动后在不同百分位阈值条件下再现时间序列的标度指数都约为 0.5,不具备长程相关性特点(图略),且此时的不同百分位阈值条件下的极端事件的再现时间序列概率分布符合 Poisson 分布. 因此证明了系统中存在长程相关性是极端事件具有群发性的原因. 为了定量描述极端事件的群发性,定义群发性指数 C_1 为:

$$C_1 = \frac{P_{\mathrm{cor}}(r_q < R_q)}{P_{\mathrm{shu}}(r_q < R_q)} - 1 \qquad (9)$$

其中,$P_{\mathrm{cor}}(r_q < R_q)$ 是长程相关性序列的小值再现时间($r_q < R_q$)的概率,$P_{\mathrm{shu}}(r_q < R_q)$ 是该序列随机扰动后的小值再现时间($r_q < R_q$)的概率,为将 C_1 值限定在[0,1]之间,公式后面减 1. 由此可以定量地比较极端事件群发性的强弱,评估极端事件对社会经济的影响,对极端事件发生进行预警. 已知 $C_1=0$ 时,系统是随机的,表示发生极端事件的概率是均匀的,群发性相对最弱;图 5 中分别计算了 $\alpha=0.7$, 0.8,0.9 三种不同长程相关性时间序列在不同阈值条件下的群发性指数 C_1,其平均值分别为 0.126, 0.184,0.244,随着标度指数的增大而增大. 为探究群发性指数 C_1 的演变规律,图 5 中分别计算了每条时间序列从 85th—95th 阈值条件下 11 个 C_1 值,可以发现对于同一条时间序列而言,C_1 基本一致,表明了群发指数 C_1 在一定阈值条件下是稳定的.

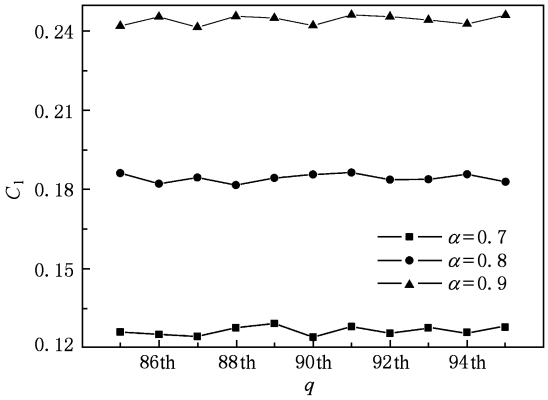

图 5　$\alpha=0.7,0.8,0.9$ 的 3 条序列在不同百分位条件下群发性指数 C_1

5. 结　　论

本文研究了百分位阈值条件下理想时间序列极端事件再现时间序列的长程相关性,发现具有长程相关性的序列,其再现时间序列也有长程相关性. 对于同一时间序列而言,不同阈值条件下极端事件再现时间序列的标度指数非常接近,但比原序列要略小. 对比 Gauss 白噪声序列发现,当原序列没有长程相关性时,其极端事件的再现时间序列也没有长程相关性. 由此得知,再现时间序列的长程相关性是由原序列的长程相关性决定的. 具有长程相关性的时间序列再现时间的概率分布明显不同于随机序列. 各理想序列在相同的阈值条件下,其平均极端事件再现时间不变,但是具有长程相关性序列再现时间序列的小值出现概率 $P(r_q < R_q)$ 增大,

导致极端事件的群发现象.证明了时间序列的长程相关性是导致该序列极端事件群发的原因.因为对于同一时间序列,小值再现时间在不同阈值条件下再现时间序列中所占概率基本不变,所以通过长程相关性序列和随机扰动后序列的比较,定义了群发性指数,计算了群发性指数并讨论了它在一定阈值条件下的稳定性,为评估极端事件可能影响提供参考.

[1] Gumbel E J 1958 *Statistics of Extremes* (New York: Columbia University Press)

[2] Fisher R A, Tippett L H C 1928 *Proc. Camb. Philos. Soc.* **24** 180

[3] Muzy B, Reg C 1997 *Adv. Atmos. Sci.* **14** 177

[4] Feng G L, Cao H X 1998 *Quart. J. Appl. Meteorol.* **9** 219 (in Chinese) [封国林、曹鸿兴 1998 应用气象学报 **9** 219]

[5] Bunde E K, Bunde A, Havlin S, Roman H E, Goldreich Y, Schellnhuber H J 1998 *Phys. Rev. Lett.* **81** 729

[6] Talkner P, Weber R O 2000 *Phys. Rev.* E **62** 150

[7] Eichner J F, Bunde E K, Bunde A, Havlin S 2003 *Phys. Rev.* E **68** 046133

[8] Wang Q G, Zhi R, Zhang Z P 2008 *Acta Phys. Sin.* **57** 5343 (in Chinese) [王启光、支 蓉、张增平 2008 物理学报 **57** 5343]

[9] Arneodo A, Bacry E, Graves P V, Muzy J F 1995 *Phys. Rev. Lett.* **74** 16

[10] Bonsal B R, Zhang X B, Vincent L A 2001 *J. Clim.* **5** 1959

[11] Zhang D Q, Feng G L, Hu J G 2008 *Chin. Phys.* B **17** 736

[12] Folland C K, Miller C, Bader D 1999 *Clim. Change* **42** 31

[13] Bunde E K, Bunde A, Havlin S, Goldreich Y 1996 *Physica* A **231** 393

[14] Peng C K, Mietus J, Hausdorff J M, Havlin S, Stanley H E, Goldberger A L 993 *Phys. Rev. Lett.* **1343** 70

[15] Feng G L, Gong Z Q, Dong W J, Li J P 2005 *Acta Phys. Sin.* **54** 5494 (in Chinese) [封国林、龚志强、董文杰、李建平 2005 物理学报 **54** 5494]

[16] Buldyrev S V, Goldberger A L, Havlin S 1995 *Phys. Rev.* E **51** 5084

[17] Feng G L, Dai X G, Wang A H, Chou J F 2001 *Acta Phys. Sin.* **50** 606 (in Chinese) [封国林、戴新刚、王爱慧、丑纪范 2001 物理学报 **50** 606]

[18] Hu K, Ivanov P C, Chen Z, Carpena P 2001 *Phys. Rev.* E **64** 011114

[19] Kantelhardt J W, Bunde E K, Rego H A 2001 *Physica* A **295** 441

[20] Cao H X, Richard B, Yu H Y, He H Z 2006 *Trans. Sci.* **9** 10

[21] He W P, Feng G L, Wu Q, Wan S Q, Chou J F 2008 *Nonlin. Proc. in Geophys.* **15** 601

[22] Altmann E G, Kantz H 2005 *Phys. Rev.* E **71** 056106

[23] Feng G L, Dong W J, Gong Z Q, Hou W, Wan S Q, Zhi R 2006 *Nonlinear Theories and Methods on Spatial-temporal Distribution of the Observational Data* (Beijing: Metrological Press) (in Chinese) [封国林、董文杰、龚志强、侯 威、万仕全、支 蓉 2006 观测数据非线性时空分布理论和方法(北京:气象出版社)]

[24] Gilberto E P, Jose A R, Alejandro V 2006 *Ann. Nucl. Energy* **33** 1309

[25] Feng G L, Dong W J 2003 *Chin. Phys.* **12** 1076

[26] Feng G L, Hou W, Gong Z Q, Zhi R 2009 *Acta Phys. Sin.* **58** 4 (in Chinese) [封国林、侯 威、龚志强、支 蓉 2009 物理学报 **58** 4]

[27] Makse H A, Havlin S, Schwartz M, Stanley H E 1996 *Phys. Rev.* E **53** 5445

[28] Schreiber T, Schmitz A 2000 *Physica* D **142** 346

Long-range correlation and group-occurrence of return intervals of extreme events*

Wang Qi-Guang[1)2)3)]　Hou Wei[2)3)]　Zheng Zhi-Hai[1)]　Feng Ai-Xia[1)]　Deng Bei-Sheng[1)†]

1)(*College of Atmospheric Sciences, Lanzhou University, Lanzhou* 730000, *China*)

2)(*Key Laboratory of Regional Climate-Environment Research for Temperate East Asia of the Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing* 100029, *China*)

3)(*Laboratory for Climate Studies, National Climate Center, China Meteorological Administration, Beijing* 100081, *China*)

(Received 16 September 2009; revised manuscript received 14 December 2009)

Abstract

This paper studies the long range correlation of the return intervals of extreme events using the method of percent threshold. In view of return intervals of the extreme events, the long range-correlations of return intervals are studied by simulated time series when the extreme events happened. It is found that the constructed extreme event time series also have the characteristics of long-range correlation if the original time series have one. Meanwhile, the scaling exponents of the two time series are very close, it is very different from the return intervals of stochastic series. The probability density function of time series with long-range correlation has significant difference from the stochastic time series. It was found that there is a group-occurrence phenomenon in the return time series with long-range correlation, so a so-called group occurrence index is defined and computed, the stability of the index is also studied. The results show that the reason of the group occurrence of extreme events series is attributed to the long-range correlation of time series.

Keywords: return interval, long-range correlation, group-occurrence, extreme events

PACC: 9260X

* Project supported by the National Natural Science Foundation of China (Grant Nos. 40875040, 40775048), the National Key Technology Research and Development Program of Ministry of Science and Technology of China (Grant No. 2007BAC29B01) and the Special Scientific Research Fund of Meteorological Public Welfare Profession of China (Grant No. GYHY200806005).

† Corresponding author. E-mail: dengbeisheng@ hotmail. com