一种确定极端事件阈值的新方法：随机重排去趋势波动分析方法*

侯 威[1)2)†] 章大全[1)] 周 云[1)3)] 杨 萍[4)]

1)(国家气候中心,北京 100081)

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)(扬州大学物理科学与技术学院,扬州 225002)

4)(中国气象局北京城市气象研究所,北京 100089)

(2010 年 11 月 15 日收到;2011 年 1 月 7 日收到修改稿)

将去趋势波动分析法(detrended fluctuation analysis,DFA)和替代数据法相结合,同时引入启发式分割算法和卡方检验,提出了一种确定极端气候事件阈值的新方法,称为随机重排去趋势波动分析(stochastic sort detrended fluctuation Analysis,S-DFA)方法. 该方法的基本物理思想是认为资料序列中出现概率非常小的数据点属于小概率事件,这些数据点包含的系统整体演化信息极少,它们所对应的状态是系统演化的异常状态或是系统受到外界扰动而出现的极端状态;而出现概率密度较大或者概率分布比较均匀的数据点则不属于小概率事件的范畴,这些数据点包含了系统演化的丰富信息,是系统演化的正常状态. 同百分位阈值方法相比,S-DFA 方法明确指出了极端事件和非极端事件之间的临界值,并通过数值试验从不同的角度对 S-DFA 方法进行检验,以验证 S-DFA 方法的有效性.

关键词: 去趋势波动分析法, 替代数据法, 极端气候事件, 阈值

PACS: 92.70.Aa

1. 引 言

20 世纪以来,对极端天气气候事件发生、发展机制的研究已经成为大气科学中非常重要的研究领域. 从概率分布的角度来看,对某一特定地点和时间,极端天气事件就是发生概率极小的事件[1,2]. 目前国际上多采用某个百分位值作为极端事件的阈值,超过或小于该阈值的事件被认为是极端天气气候事件;也有人对不同气候要素采用不同分布型的边缘值来确定极端天气气候事件的阈值.

随着非线性动力学在认识天气与气候系统过程和机理方面取得巨大进展[3—10],"通过何种方式减少对极端天气气候事件认识上的不确定性"正成为目前国内外极端气候事件研究的重点之一. 极端天气气候事件研究中最初的或最基本的不确定性来源于极端天气气候事件的定义. 无论是采用边缘值还是百分位定义来确定极端事件的阈值,都没有给出具体、明确的临界值,也就是概率究竟在何种情况下可以称之为"小"? 而且由于定义本身并没有考虑极端事件在整个动力系统中的作用,即没有从系统整体动力学特征的角度来讨论极端事件,这些定义给出的阈值确定方法并没有明确的物理背景.

本文基于系统的整体动力学特征,将去趋势波动分析法和替代数据法相结合,得到一种具有物理背景和统一定义的确定极端事件阈值的新方法,称为随机重排去趋势波动分析(S—DFA)方法. 该方法明确指出了极端事件和非极端事件之间的临界值,能将系统演化的极端状态或异常状态同常规状态区分开来. 进一步通过理想数据和实际数据从不同的角度对 S-DFA 方法进行了反复检验,验证了 S-DFA 方法的有效性.

* 国家自然科学基金(批准号:41005043,40905034)、全球变化研究国家重大科学研究计划(批准号:2010CB950504)和科技支撑项目(批准号:2007BAC29B01)资助的课题.

† E-mail:hou_w@sohu.com

2. 去趋势波动分析方法和替代数据方法介绍

定量描述气候系统复杂的非线性演化过程及其自相似结构特征的有效手段之一是多分形理论. 分形时间序列在不同的时间标度上有类似的统计特性,具有长程相关性. 分形时间序列可以有分数维数,使时间序列表现出局部的随机性和整体的相似性. 20 世纪 90 年代一些学者根据 DNA(脱氧核糖核酸)机理提出了一种全新的研究时间序列波动长程相关性的标度指数计算方法,即去趋势波动分析(DFA)方法[11],它是基于随机过程理论和混沌动力学新发展的一种分析方法,用来检测时间序列的物理特征,随后这一方法在自然科学和社会科学等领域得到了广泛的应用[12—16]. 去趋势波动分析法的具体过程及其物理意义可参见文献[12,17]. DFA 方法对不稳定信号和噪声信号不敏感,不同的尺度 DFA 指数 h_q 反映了不同的时间过程.

本文在 DFA 方法的计算中取参数 $q = 2$, 即对每个子区间 v,($v = 1,2,\cdots,2N_s$) 的数据进行二阶多项式回归拟合. 此时,对于平稳序列而言, h_q 就是赫斯特指数 H; 对于非平稳序列,当 $h_q = 0.5$ 时,该序列为一独立随机过程;当 $0.5 < h_q \leqslant 1$ 时,该序列存在长程相关性;当 $h_q < 0.5$ 时,该序列存在负长程相关,即反持久性[18—20].

为了获得与原始数据具有相同的均值、方差和自相关函数以及概率分布函数的替代数据,采用 Theiler 和 Prichar 提出的替代数据法[21—23],也就是相位随机化方法,该算法的基本思想是保留原始数据的功率谱值,但随机产生替代数据的功率谱相位[24—27]. 替代数据保留了原始数据的线性自相关函数,而非线性自相关性被相位随机化消除了.

值得注意的是得到的替代数据必须为实数,也就是写成复数后其虚部为零,从坐标象限来看替代数据全部落在 x 轴上,经过正的 Fourier 变换和扭转相位后的逆的 Fourier 变换,若其虚部不为零则替代数据必落在整个复平面内,这样会造成原始序列中的某些固有特性流失在虚部中. 为了减小这种信息流失,在相位随机化时应保证

$$\phi(1) = \phi(N/2 + 1) = 0,$$
$$\phi(k) = -\phi(N + 2 - k),\ k = 2,3,\cdots,N/2. \quad (1)$$

3. 利用去趋势波动分析法和替代数据法确定极端事件的阈值

3.1. 算法介绍

下面介绍将 DFA 方法和替代数据法相结合的具体过程. 对系统演化的某一组时间序列 $\{x_i, i = 1,\cdots,n\}$, 首先得到 $\{x_i\}$ 的最大值 $x_{\max}$ 和最小值 $x_{\min}$, 确定参考点 R,R 值可以是序列均值 x_{ave} 或者界于 $x_{\max}$ 与 $x_{\min}$ 之间的某一中值 x_{med}; 然后从 $x_{\max}$ 开始,对序列中位于第 k 个数据区间内的数据点 $\{x_i, x_i \geqslant x_{\max} - d \times k\}$ 的顺序进行随机化,其中 d 为区间间隔, d 的数量级取为序列 $\{x_i\}$ 数量级的 $\left[\frac{1}{10},\frac{1}{100}\right]$ 之间, $\text{int}(1,\cdots,(x_{\max} - R)/d)$, $\text{int}(\,)$ 表示取整,同时保留序列中其余数据的顺序不变,直到 $x_i = R$, 依次得到新序列 Y_J,$J = x_{\max} - d \times k$; 再从 $x_{\min}$ 开始,对第 k 个数据区间内的数据点 $\{x_i, x_i \leqslant x_{\min} + d \times k\}$ 的顺序进行随机化,同样, d 为区间间隔, $k = \text{int}(1,\cdots,(R - x_{\min})/d)$, 保留其余数据的顺序不变,直到 $x_i = R$, 依次得到新序列 Y_J,$J = x_{\min} + d \times k$;将 J 看作为各个重排区间下限或上限值,最后计算每个新序列 Y_J 的长程相关性指数 DFA_J,得到其随 J 的变化.

3.2. 算法物理含义

分别使用满足高斯分布、指数分布及平均分布的三组随机序列 $\{G_i\}$, $\{E_i\}$, $\{A_i\}$($i = 1,\cdots,10000$), 使用 3.1 中的算法对不同序列各个数据区间内的数据分别进行随机重排,得到各组新序列 Y_J, 计算各 Y_J 的长程相关性指数 DFA_J,观察其随 J 值的演化发展. 对上述过程重复计算 5 次,得到 5 个 DFA_J 随 J 的演化序列,分别用 S_1,S_2,S_3,S_4 和 S_5 来表示. 关于计算中参数的选择,对局部趋势函数 $y_v(i)$ 使用二阶多项式进行拟合,即取 $q = 2$; 不重叠等长度子区间的长度 s 取值为 $100 \leqslant s \leqslant n/10$,$n$ 为序列长度;区间间隔取 $d = 0.01$;参考点 R 取为.

从图 1 中可以看出,对于高斯分布随机序列 $\{G_i\}$ 的正值部分,随着 J 值的不断增加,重排区间内进行顺序随机重排的数据点个数不断减少,其所对应的概率密度也越来越小,进行随机重排后得到的各个新序列 Y_J 的 DFA 指数 DFA_J 也偏离原始序列的 DFA 指数(如图 1 中点线所示),并且不同 Y_J

的 DFA 指数之间也有很大的差异. 但当重排区间内的数据点个数越来越小,在 $J \geqslant 2.6$ 时,各个新序列 Y_J 的 DFA 指数开始收敛于原始序列的 DFA 指数(如图 1 中虚线框所示),并且各个 Y_J 的 DFA 指数之间的差异也愈来愈小. 对于 $\{G_i\}$ 的负值部分,随着 J 值的不断减小,重排区间内进行顺序随机重排的数据点个数也在不断减少,各个新序列 Y_J 的 DFA 指数也偏离于原始序列的 DFA 指数且彼此之间有较大差异,同样,当重排区间内的数据点个数越来越小时,在 $J \leqslant -2.6$ 时,各个新序列 Y_J 的 DFA 指数之间的差异也愈来愈小,并且都开始收敛于原始序列的 DFA 指数.

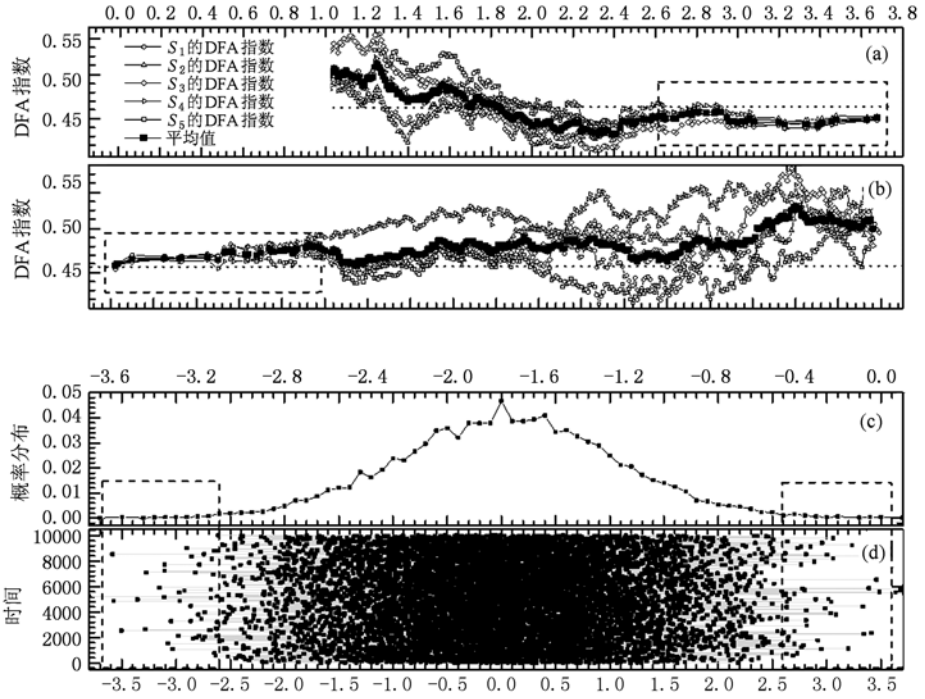

图 1 (a)对 $\{G_i\}$ 正值的不同区间数据点进行顺序随机化后 DFA 指数的变化;(b)对 $\{G_i\}$ 负值的不同区间数据点进行顺序随机化后 DFA 指数的变化;(c) $\{G_i\}$ 的概率分布;(d)原始序列 $\{G_i\}$

从图 2 中可以看出,对于指数分布随机序列 $\{E_i\}$ 的正值(负值)部分,随 J 的增大(减小),各个重排区间内数据点个数不断减少,其对应的概率密度也越来越小,随机重排之后得到的新序列 Y_J 的 DFA 指数偏离原始序列的 DFA 指数(如图 2 中点线所示),各个新序列 Y_J 的 DFA 指数之间差异较大,在 $J \geqslant 7.2$ 时,逐渐演变为各个 Y_J 的 DFA 指数向原始值收敛且彼此之间差异很小(如图 2 中虚线框所示).

由于 $\{A_i\}$ 满足均匀随机分布,所以随着 J 取值的不断增加,各个重排区间内数据点数目,即数据点概率密度的差异很小,不会出现小概率或大概率事件. 在图 3(a),(b)中,随着 J 取值的不断增加,对各个区间内的数据顺序进行随机重排之后得到的新序列 Y_J 的 DFA 指数,也都偏离于原始序列的 DFA 指数(如图 3 中点线所示)且彼此之间的差异也较大,所有 Y_J 的 DFA 指数均没有出现向原始值收敛的情况,这一点和图 1,图 2 均不一样.

综上所述,可以看出,当重排区间内数据点很少时,也就是其概率分布处于整个序列概率分布的尾部,改变这些数据点顺序得到的新序列 Y_J 的 DFA 指数收敛于原始值,即改变这些数据点的顺序对序列 DFA 指数影响不大或几乎没有影响,由于这部分数据点的概率密度非常小,其中所包含系统演化信息的极少,统计效应可以基本忽略,所对应的状态不属于系统的常规演化轨迹,而是系统演化的极端状态或是系统受到外界扰动而导致的极端异常状态,属于小概率事件的范畴;对于那些概率密度较大或者概率分布比较平均的数据点,改变其顺序对序列 DFA 指数的影响也较大,它们包含了丰富的系统演化信息,其统计效应也非常显著,属于系统演化的常规状态,不属于小概率事件的范畴.

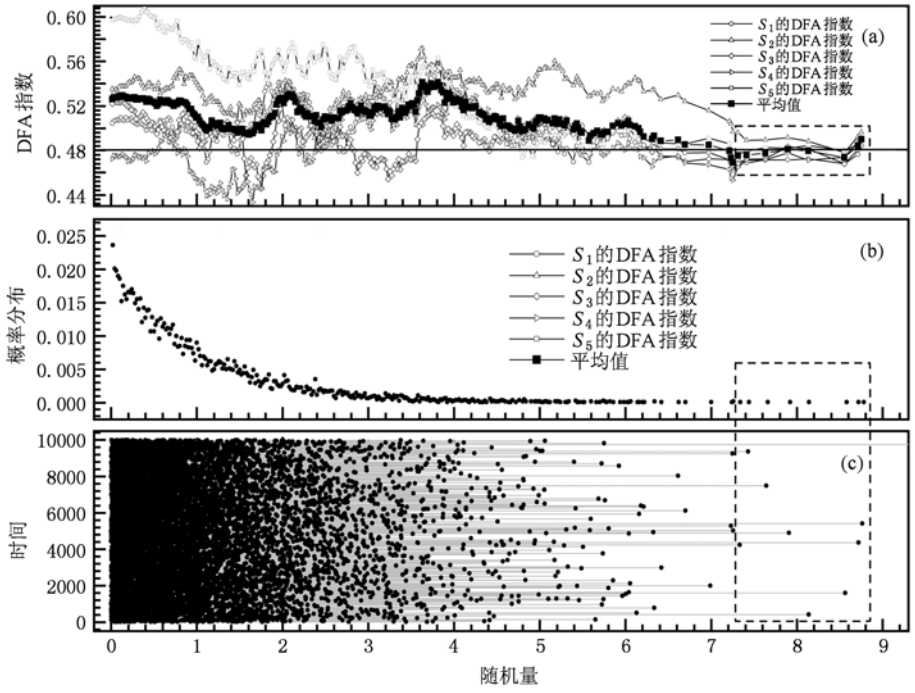

图2 (a)对 $\{E_i\}$ 的不同区间数据点进行顺序随机化后 DFA 指数的变化;(b) $\{E_i\}$ 的概率分布图;(c)原始序列 $\{E_i\}$

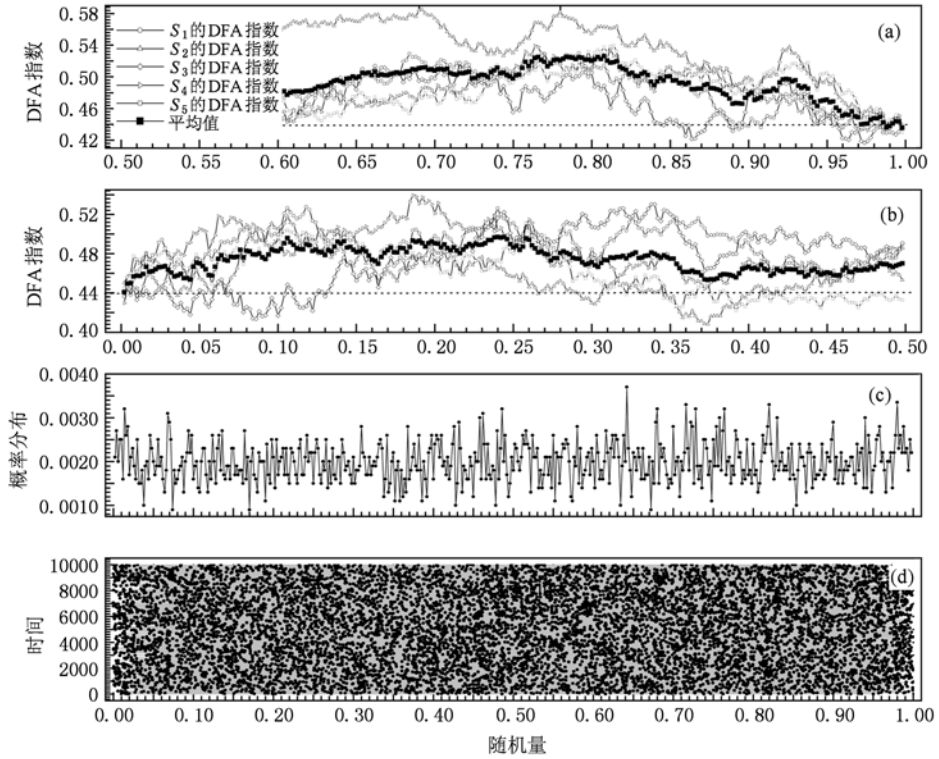

图3 (a)对 $\{A_i\}$ 正值的不同区间数据点进行顺序随机化后 DFA 指数的变化;(b)对 $\{A_i\}$ 负值的不同区间数据点进行顺序随机化后 DFA 指数的变化;(c) $\{A_i\}$ 的概率分布;(d)原始序列 $\{A_i\}$

由于极端事件或者极值事件属于小概率事件,同样可以认为此类事件所对应的演化状态是系统极端状态,或是系统的异常演化状态,不属于系统自身正常演化状态的范畴. 由前面的分析可知,将DFA方法和替代数据法相结合,通过确定序列 Y_J 的DFA指数何时开始收敛于原始值,此时所对应的 J 值就是所研究序列的小概率事件的阈值,即系统极端事件的阈值,所以可将上述算法运用于确定序列或系统极端事件或极值事件的阈值. 下面给出具体实例及判断收敛点的方法.

4. 利用去趋势波动分析法和替代数据法确定极端事件阈值

4.1. 确定理想序列极端事件阈值

使用混沌系统Lorenz方程 x 分量的10000个数据点作为原始序列 $\{x_i\}$,且已知 $18.57 > \{x_i\} > -18.667$. 从序列 $\{x\}$ 中随机选择任意点 x_i 并改变其值大小,使得 $x_i > 18.57$ 或 $x_i < -18.667$,使其成为极值点,得到含极值的理想序列 $\{x'_i\}$,如图4(c)所示. 算法如3.1中所述,并重复计算5次,分别用 S_1, S_2, S_3, S_4 和 S_5 来表示5个 DFA_J 随 J 值的变化. 取区间间隔 $d = 0.1$,其余参数选择同3.2节.

原始序列 $\{x_i\}$ 的DFA指数为0.704300,理想序列 $\{x'_i\}$ 的DFA指数为0.700062,二者只在小数点后第三位发生变化,说明小概率事件对系统DFA指数的影响非常小,这也验证了本方法的可靠性.

图4和图5中的点线代表理想序列 $\{x'_i\}$ 的DFA指数. 可以看到,在图4(a)和(b)中,当 $x'_i \geqslant R$ 时(此时 $J = R = 0$),也就是对理想序列中所有大于等于零的数据点的顺序进行5次重排后,得到的5个 Y_J 的DFA指数都和原始序列的DFA指数有很大差异,而且各个 Y_J 的 DFA_J 彼此之间也存在着很大变化;随着 J 值的逐渐增加,不同 Y_J 的DFA指数逐渐收敛于原始序列的DFA指数值. 下面介绍确定收敛点的方法.

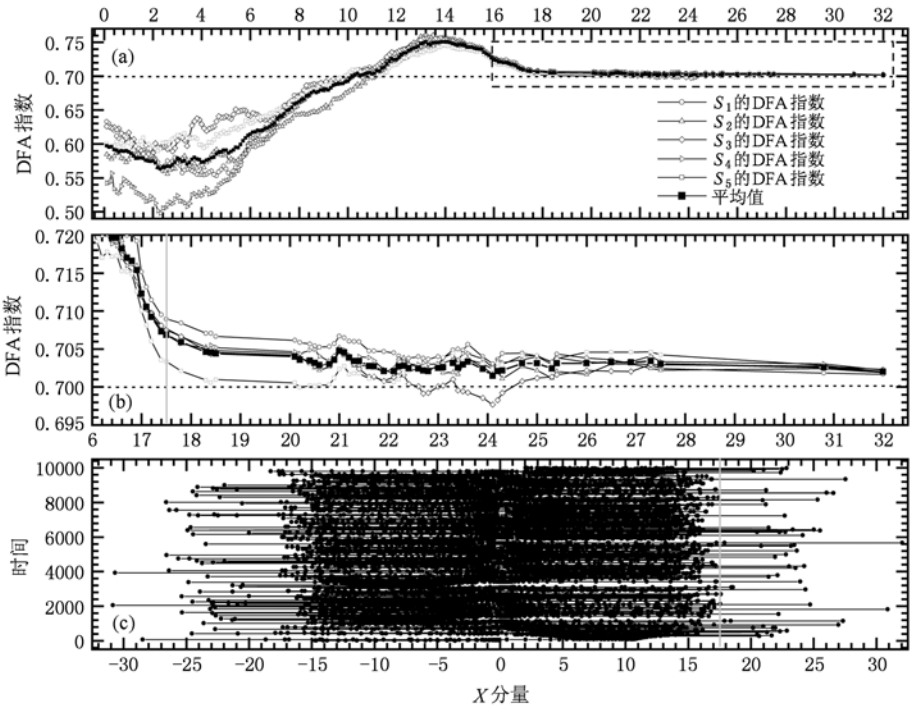

图4 (a)对 $\{x'_i\}$ 正值的不同区间数据点进行顺序随机化后DFA指数的变化;(b)收敛区域放大图;(c)理想序列 $\{x'_i\}$

不同 Y_J 的DFA指数序列 $\{\mathrm{DFA}_J\}$ 随时间的演化逐渐收敛于原始序列DFA指数值,由于随着 J 值逐渐变大,$\{\mathrm{DFA}_J\}$ 的收敛并不是完全一致的逼近收敛形式,而是围绕原始序列DFA指数值有着振幅非常小的波动,因此通过计算 $\{\mathrm{DFA}_J\}$ 的方差序列 $\{\mathrm{var}_J\}$ 来消除这一波动对判断收敛点带来的影响. 随着

$\{DFA_J\}$ 逐渐收敛于原始序列 DFA 指数值,序列 $\{DFA_J\}$ 的平均值也越来越接近于原始序列 DFA 指数值,不同 DFA_J 值之间的差异逐渐减小,其方差 $\{var_J\}$ 也逐渐减小最终收敛于 0. var_J 值越小,说明 $\{DFA_J\}$ 收敛程度越高,反之亦然,如图 5(b)所示. 并且 $\{var_J\}$ 与 $\{DFA_J\}$ 的收敛具有同步性,也就是同时开始收敛. 由于序列 $\{var_J\}$ 向着零值逐级收敛,随着收敛程度的变化,在此过程中 $\{var_J\}$ 存在一系列的转折点 $\{x_J\}$,每个转折点表示 $\{var_J\}$ 的收敛程度在此点前后是不一样的,同时这些转折点也代表 $\{DFA_J\}$ 向着原始序列 DFA 指数值收敛程度的变化. 一种情况是 $\{var_J\}$ 值增大,此时 $\{var_J\}$ 背离于收敛,越来越远离收敛值;另一种情况是 $\{var_J\}$ 值减小,此时 $\{var_J\}$ 进一步收敛,越来越接近收敛值. 因为要确定收敛点,所以只考虑序列 $\{var_J\}$ 中愈来愈收敛(即 $\{var_J\}$ 值越来越小)的区域即收敛区(如图 5(a)内虚线框所示),在收敛区域内,所有的转折点均代表了序列 $\{var_J\}$ 向收敛值进一步逼近,收敛程度进一步提高. 采用 BG 算法确定序列 $\{var_J\}$ 中的转折点 $\{x_J\}$. Bernaola-Galvan 在 2001 年提出的启发式分割算法(BG 算法)是一种有别于传统理论的突变检测方法,其详细算法过程可参见文献[4,27,28]. BG 算法是一种检测转折或突变的有效方法,该方法可以将非平稳序列分割为多尺度的自平稳子序列,检测的尺度和精度具有可变性,能够检测不同尺度和不同幅度的转折或突变;白噪声和尖峰噪声对该方法的影响较小.

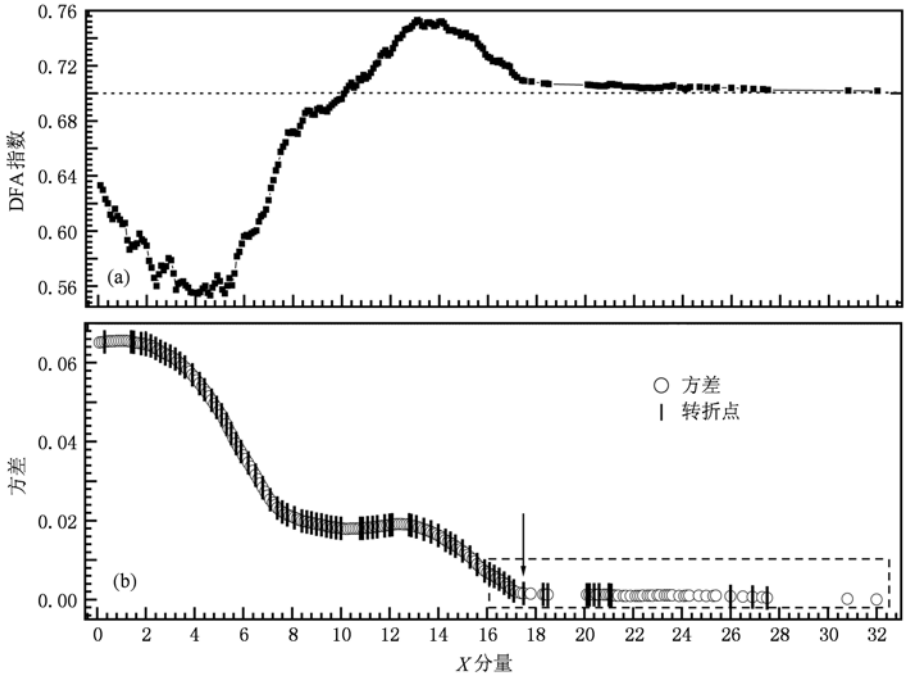

图 5 (a) DFA 指数平均值的方差及其突变点;(b) DFA 指数平均值

由于转折点 $\{x_J\}$ 中有些点所对应的 DFA 值远远偏离了原始序列 DFA 指数值,这些点不是收敛点,所以对收敛区域内的转折点采用总体方差显著性检验(卡方检验)方法来检验各个转折点是否就是收敛点. 如果在收敛区域内存在 $A \to F$ 一系列转折点,且已知 A 点不是收敛点和 D 点为收敛点,则收敛点 D 之前的转折点因为还没有达到收敛,其收敛程度将继续提高,由于各转折点的收敛情况不一致,所以转折点 A,B,C 之间应具有显著差异;而 D 点之后的转折点因为已经达到收敛,其收敛程度已经非常高且可以提高的余地非常小,各转折点的收敛情况一致,素以转折点 D,E,F 之间应无显著差异. 采用卡方检验对收敛区域内各个转折点之间的差异进行显著性检验,据此就可以确定出收敛点为 D.

卡方检验是一种较为常用的数理统计方法,在测量、天气预报、工农业生产统计中有较为广泛的应用与良好的应用效果[29—31]. 卡方检验在统计学上又称为 χ^2 检验,是指通过从某一总体中抽取出的样本来判断这一总体的方差与某一已知值的大小关系,因此零假设表示

$$H_0: s^2 = \partial_0^2, \tag{2}$$

根据抽样分布理论, n 倍的样本方差与总体方差

之比服从自由度为 $(n-1)$ 卡方分布,检验公式为

$$\chi^2 = ns^2/\partial_0^2 \propto \chi^2(n-1), \quad (3)$$

定义显著性水平为 α, 则当 $\chi^2 > \chi^2_{(\alpha/2)}$ 或 $\chi^2 < \chi^2_{(1-\alpha/2)}$ 时, s^2 和 ∂_0^2 差异显著; $\chi^2_{(\alpha/2)} > \chi^2 > \chi^2_{(1-\alpha/2)}$ 时, s^2 和 ∂_0^2 差异不显著.

根据自由度 $(n-1)$ 可以从 χ^2 表中查到 $\chi^2_{(\alpha/2)}$ 和 $\chi^2_{(1-\alpha/2)}$ 的值. 据此可以对收敛区内的各转折点(图 5(a)内虚线框所示)之间的差异进行显著性检验,确定收敛点.

以图 5 为例,具体算法为:从序列 $\{\mathrm{var}_J\}$ 的末尾开始计数,令自由度 $n_{J=32} = 1$, 收敛区域内转折点为 $x_{J=16.3}$, $x_{J=16.5}$, $x_{J=16.7}$, $x_{J=16.9}$, $x_{J=17.1}$, $x_{J=17.5}$, $x_{J=18.3}$, $x_{J=18.5}$, $x_{J=20.1}$, $x_{J=20.2}$, $x_{J=20.4}$, $x_{J=20.6}$, $x_{J=21}$, $x_{J=21.1}$, $x_{J=26}$, $x_{J=26.9}$ 和 $x_{J=27.5}$. 因为 $\mathrm{DFA}_{J=16.3}$ 明显偏离于原始序列 DFA 指数值,所以点 $x_{J=16.3}$ 不是收敛点,其所对应的 $\mathrm{var}_{J=16.3}$ 为序列 $\{\mathrm{DFA}_J, J=16.3\}$ 的方差,以其作为整体方差 ∂_0^2; 后一个转折点为 $x_{J=16.5}$, 对应的 $\mathrm{var}_{J=16.5}$ 为序列 $\{\mathrm{DFA}_J, J=16.5\}$ 的方差,以其作为样本方差 s^2, 自由度 $n_{J=16.5} = 57$, 根据卡方检验来判定这两个序列 $\{\mathrm{DFA}_J, J=16.3\}$ 和 $\{\mathrm{DFA}_J, J=16.5\}$ 的差异是否显著,如显著,则点 $x_{J=16.3}$ 和 $x_{J=16.5}$ 处的收敛情况也有显著差异. 取显著性水平为 99%,即 $\alpha = 0.01$,通过计算得到

$$\chi^2_{16.316.5} = n_{J=16.5}s^2/\partial_0^2 = 43.684,$$

查表得 $\chi^2_{(\alpha/2)} = 91.952$, $\chi^2_{(1-\alpha/2)} = 35.534$, 此时 $\chi^2_{(1-\alpha/2)} < \chi^2_{16.316.5} < \chi^2_{(\alpha/2)}$, 说明点 $x_{J=16.3}$ 和 $x_{J=16.5}$ 处的收敛无显著差异,两点的收敛情况一致;下一个转折点为 $x_{J=16.7}$, 同样以 $x_{J=16.5}$ 对应的 $\mathrm{var}_{J=16.5}$ 作为整体方差 s^2, $x_{J=16.7}$ 对应的 $\mathrm{var}_{J=16.7}$ 为样本方差 ∂_0^2, $n_{J=16.7} = 55$, 计算得

$$\chi^2_{16.516.7} = n_{J=16.7}s^2/\partial_0^2 = 36.211,$$

查表得 $\chi^2_{(\alpha/2)} = 91.952$, $\chi^2_{(1-\alpha/2)} = 35.534$, 此时 $\chi^2_{(1-\alpha/2)} < \chi^2_{16.516.7} < \chi^2_{(\alpha/2)}$ 则点 $x_{J=16.5}$ 和 $x_{J=16.7}$ 处的收敛也无显著差异;其余各个转折点之间差异的显著性检验结果见表 1.

表 1 各个转折点之间差异的显著性检验

转折点	n	χ^2	$\chi^2_{(\alpha/2)}$	$\chi^2_{(1-\alpha/2)}$	差异是否显著
$x_{J=16.5}$	57	$\chi^2_{16.316.5} = n_{J=16.5}s^2/\partial_0^2 = 43.684$	91.952	35.534	否
$x_{J=16.7}$	55	$\chi^2_{16.516.7} = n_{J=16.7}s^2/\partial_0^2 = 36.211$	91.952	35.534	否
$x_{J=16.9}$	53	$\chi^2_{16.716.9} = n_{J=16.9}s^2/\partial_0^2 = 31.172$	82.23	30.254	否
$x_{J=17.1}$	51	$\chi^2_{16.917.1} = n_{J=17.1}s^2/\partial_0^2 = 24.191$	80.18	28.67	是
$x_{J=17.5}$	48	$\chi^2_{17.117.5} = n_{J=17.5}s^2/\partial_0^2 = 23.399$	76.945	26.534	是
$x_{J=18.3}$	46	$\chi^2_{17.518.3} = n_{J=18.3}s^2/\partial_0^2 = 34.029$	74.400	25.077	否
$x_{J=18.5}$	44	$\chi^2_{18.318.5} = n_{J=18.5}s^2/\partial_0^2 = 38.469$	71.9	23.6	否

由表 1 可以看出,从转折点 $x_{J=17.5}$ 开始,位于点 $x_{J=17.5}$ 后面的各个转折点之间的差异都不显著,各点处的收敛情况一致,可以认为序列 $\{\mathrm{DFA}_J\}$ 在点 $x_{J=17.5}$ 处开始收敛,点 $x_{J=17.5}$ 即为收敛点(图 5 中箭头表示). 当 $J \geqslant 17.5$ 时, $\{\mathrm{DFA}_J\}$ 序列收敛于原始序列 DFA 指数值,且 $\{\mathrm{DFA}_J\}$ 中各值之间差异也很微小,可以确定理想序列 $\{x'_i\}$ 的极大值阈值为 17.5(图 5 中灰色竖线表示). 对于理想序列 $\{x'_i\}$ 的负值部分,使用相同的计算方法和步骤,发现在点 $x_{J=-16.4}$ 之前,各个收敛点之间的差异明显,而在 $x_{J=-16.4}$ 之后,各个收敛点之间的差异不明显,可以认为点 $x_{J=-16.4}$ 即为收敛点,确定序列 $\{\mathrm{DFA}_J\}$ 的收敛点为 $J = -16.4$, 当 $J \leqslant -16.4$ 时, $\{\mathrm{DFA}_J\}$ 收敛于原始值,不同 Y_J 的 DFA 指数之间差异也很微小,可以确定理想序列 $\{x'_i\}$ 的极小值阈值是 -16.4.

4.2. 确定原始序列极端事件阈值

下面对原始序列 $\{x_i\}$ 采用同 4.1 节中相同的方法和步骤进行计算,分析其极大值阈值和极小值阈值. 参数选择同 4.1 节. 初始序列 $\{x_i\}$ 的 DFA 指数为 0.704300(图 6 和图 7 中点线所示).

用 BG 算法来确定 $\{\mathrm{VAR}_J\}$ 中的转折点,并用卡方检验来判定这些转折点之间的显著性是否明显,以此来确定收敛点,即极端事件的阈值. 图 7 中,对原始序列 $\{x_i\}$ 的正值部分,在点 $x_{J=17.2}$ 之前,各个收敛点之间的差异明显,其后各个收敛点之间的差异不明显,可以认为点 $x_{J=17.2}$ 即为收敛点(图 7 中箭头表示),确定序列 $\{\mathrm{DFA}_J\}$ 的收敛点为 $J = 17.2$ (图 6 中灰色竖线表示),其余各转折点差异的显著

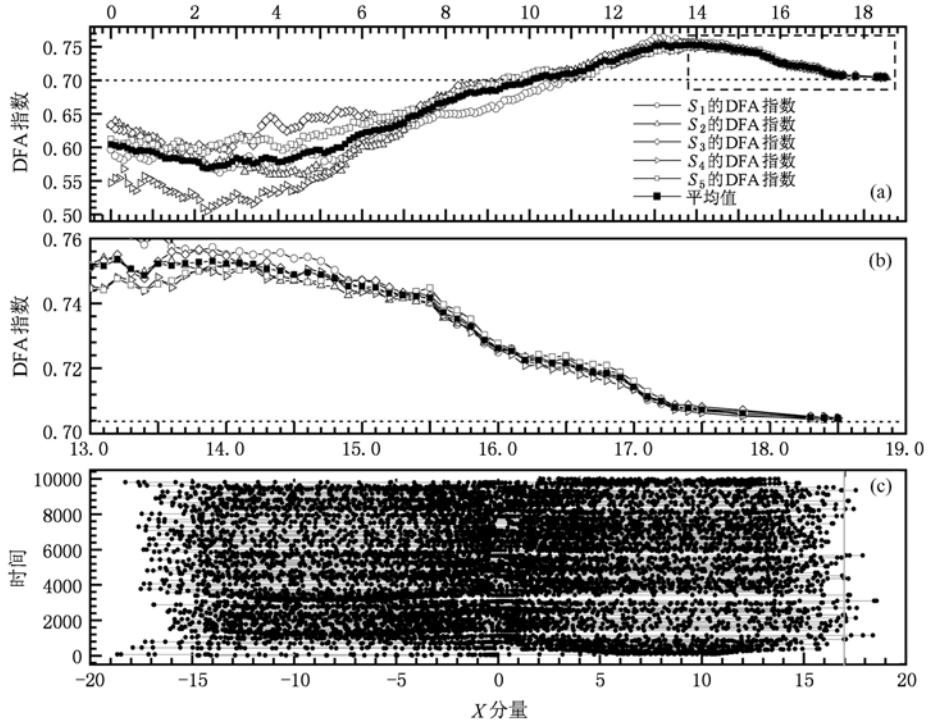

图6 (a) 对 $\{x_i\}$ 正值的不同区间数据点进行顺序随机化后 DFA 指数的变化; (b) 收敛区域放大图; (c) 原始序列 $\{x_i\}$

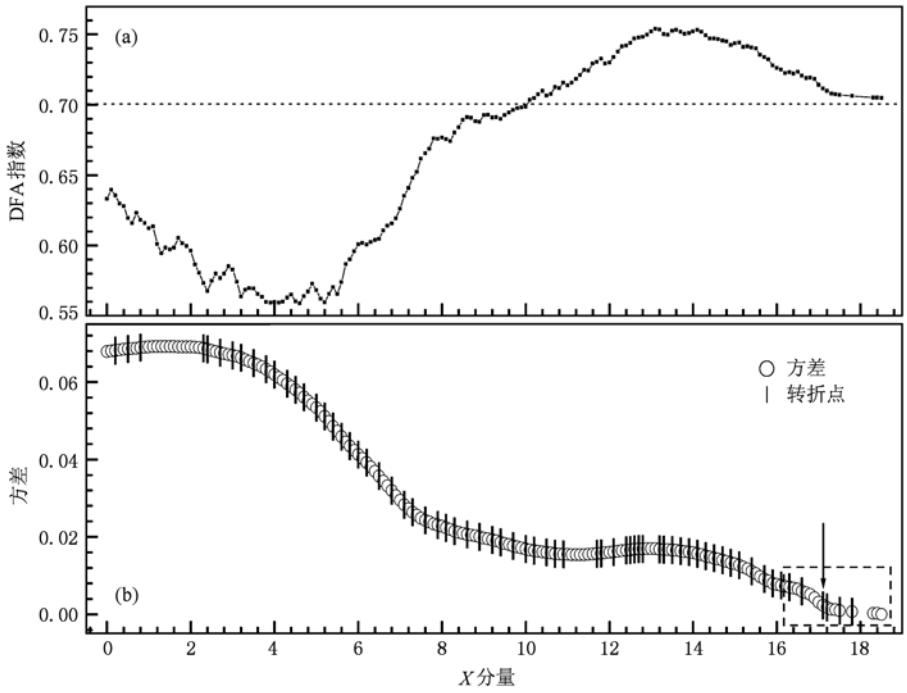

图7 (a) DFA 指数平均值的方差及其突变点; (b) DFA 指数平均值

性检验见表 2. 当 $J \geqslant 17.2$ 时, $\{\mathrm{DFA}_J\}$ 收敛于原始值, 不同 Y_J 的 DFA 指数之间差异也很微小, 可以确定原始序列 $\{x_i\}$ 的极大值的阈值是 17.2.

对原始序列 $\{x_i\}$ 的负值部分, 在点 $x_{J=-16.4}$ 之

表 2 各个转折点之间差异的显著性检验

转折点	n	χ^2	$\chi^2_{(\alpha/2)}$	$\chi^2_{(1-\alpha/2)}$	差异是否显著
$x_{J=16.6}$	14	$\chi^2_{16.316.6} = n_{J=16.6}s^2/\partial_0^2 = 10.013$	31.3	4.07	否
$x_{J=16.8}$	12	$\chi^2_{16.616.8} = n_{J=16.8}s^2/\partial_0^2 = 8.702$	23.6	1.73	否
$x_{J=17.2}$	8	$\chi^2_{16.817.2} = n_{J=16.8}s^2/\partial_0^2 = 0.975$	22.0	1.34	是
$x_{J=17.5}$	5	$\chi^2_{17.217.5} = n_{J=17.5}s^2/\partial_0^2 = 1.470$	16.7	0.412	否
$x_{J=17.8}$	4	$\chi^2_{17.517.8} = n_{J=17.8}s^2/\partial_0^2 = 1.987$	14.9	0.207	否

前,各个收敛点之间的差异明显,其后各个收敛点之间的差异不明显,可以认为点 $x_{J=-16.4}$ 即为收敛点,确定序列 $\{\mathrm{DFA}_J\}$ 的收敛点为 $J=-16.4$. 当 $J\leqslant -16.4$ 时,$\{\mathrm{DFA}_J\}$ 收敛于原始值,不同 Y_J 的 DFA 指数之间差异也很微小,可以确定原始序列 $\{x'_i\}$ 的极小值的阈值是 -16.4.

理想序列 $\{x'_i\}$ 是由原始序列 $\{x_i\}$ 改造而来,来自于同一个动力学系统,因此两个序列的极端事件阈值应该是相同或十分接近的. 序列 $\{x'_i\}$ 中 $x'_i>18.57$ 和 $x'_i<-18.67$ 是 $\{x_i\}$ 受到外界随机扰动而产生的极值点,由 4.1 节分析可知,$\{x'_i\}$ 的极大值阈值为 17.5,极小值的阈值为 -16.4,而 $\{x_i\}$ 的极大值阈值为 17.2,极小值的阈值为 -16.4,二者极端事件阈值几乎完全一致,仅有微小的差别,说明本方法所是可靠的,可以准确给出潜在动力学系统相同的、不同表现形式的序列的极端事件阈值.

4.3. 方法检验-理想序列去除极值

根据 4.1 节中确定的极端事件阈值,消除理想序列 $\{x'_i\}$ 中的极值点或极端事件,也就是去除序列 $\{x'_i\}$ 中 $x'_i\geqslant 17.5$ 或 $x'_i\leqslant -16.4$ 点,以达到消除极端事件的目的,得到新序列 $\{x1''_i\}$,其中 $\{x1''_i, x1''_i = x'_i - 17.5, x'_i\geqslant 17.5\}$ 或 $\{x1''_i, x1''_i = x'_i + 16.4, x'_i\leqslant -16.4\}$,对序列 $\{x1''_i\}$ 采用同样的思路和算法进行计算,以检验对于一个不含有极大值或极小值的序列,本方法是否仍然有效,即能确定该序列中无极端事件,以此对本方法进行进一步检验. 同样采用 4.1 中所述的计算方法. 参数选择同 4.1 节. 计算得到序列 $\{x1''_i\}$ 的 DFA 指数为 0.72137(图 8 和图 9 中点线所示).

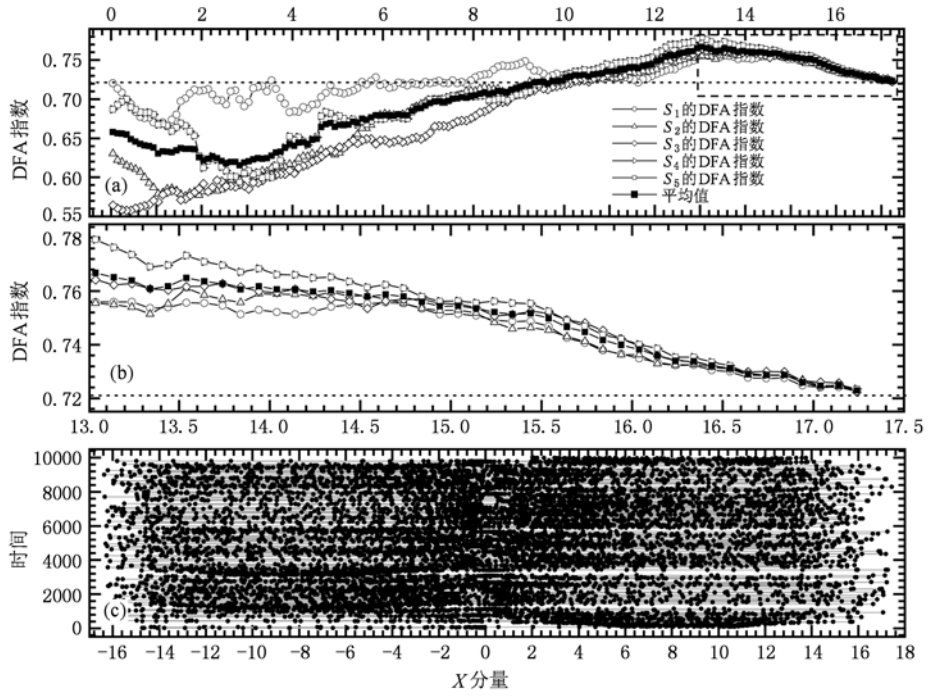

图 8 (a)对 $\{x1''_i\}$ 正值的不同区间数据点进行顺序随机化后 DFA 指数的变化;(b)收敛区域放大图;(c) 不含极端值的序列 $\{x1''_i\}$

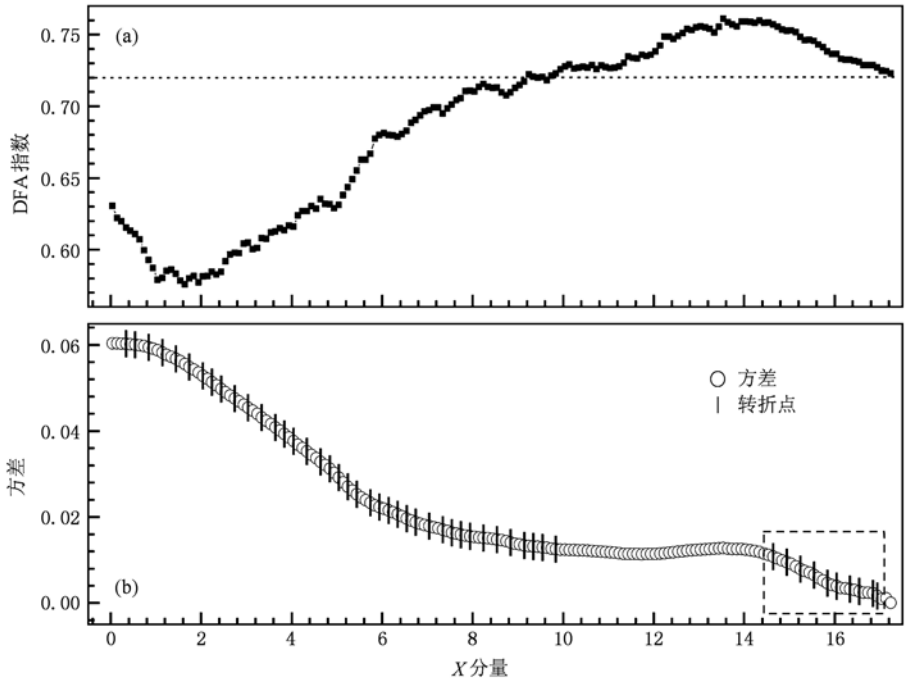

图9 (a) 序列$\{x1''_i\}$DFA 指数平均值的方差及其突变点;(b) 序列$\{x1''_i\}$DFA 指数平均值

图8(a)和(b)中的 DFA_J 序列随着 J 值的增大也表现出收敛于原始 DFA 指数的趋势,计算序列$\{DFA_J\}$的方差序列$\{VAR_J\}$,用 BG 算法来确定$\{VAR_J\}$中的转折点,并用卡方检验来判定这些转折点之间的显著性是否明显,以此来确定收敛点,即极端事件的阈值. 对图9(a)中最接近收敛区(图9(b)内虚线框所示)的转折点进行检验,从序列$\{var_J\}$的末尾开始计数,令自由度 $n_{J=17.88}=1$. 收敛区域内转折点依次为 $x_{J=14.64}$, $x_{J=14.94}$, $x_{J=15.24}$, $x_{J=15.54}$, $x_{J=15.84}$, $x_{J=16.04}$, $x_{J=16.34}$, $x_{J=16.54}$, $x_{J=16.84}$ 和 $x_{J=16.94}$. 第一个转折点为 $x_{J=14.64}$,此点并未开始收敛,是一虚假收敛点,以 $x_{J=14.64}$对应的 $\mathrm{var}_{J=14.64}$作为整体方差 s^2,下一个转折点为 $x_{J=14.94}$,其对应的 $\mathrm{var}_{J=14.94}$为样本方差 ∂_0^2,$n_{J=14.94}=24$,计算得

$$\chi^2_{14.6414.94}=n_{J=14.94}s^2/\partial_0^2=18.827,$$

查表得 $\chi^2_{(\alpha/2)}=45.6$,$\chi^2_{(1-\alpha/2)}=9.89$,$\chi^2_{(1-\alpha/2)}<\chi^2_{14.6414.94}<\chi^2_{(\alpha/2)}$,点 $x_{J=14.64}$ 和 $x_{J=14.94}$ 处的收敛没有显著差异,两点处收敛情况一致;其余各转折点差异的显著性检验见表3.

表3 各个转折点之间差异的显著性检验

转折点	n	χ^2	$\chi^2_{(\alpha/2)}$	$\chi^2_{(1-\alpha/2)}$	差异是否显著
$x_{J=14.94}$	24	$\chi^2_{14.6414.94}=n_{J=14.94}s^2/\partial_0^2=18.827$	45.6	9.89	否
$x_{J=15.24}$	21	$\chi^2_{14.9415.24}=n_{J=15.24}s^2/\partial_0^2=14.415$	41.4	8.03	否
$x_{J=15.54}$	18	$\chi^2_{15.2415.54}=n_{J=15.54}s^2/\partial_0^2=12.135$	37.2	6.26	否
$x_{J=15.84}$	15	$\chi^2_{15.5415.84}=n_{J=15.84}s^2/\partial_0^2=7.476$	32.8	4.60	否
$x_{J=16.04}$	13	$\chi^2_{15.8416.04}=n_{J=16.04}s^2/\partial_0^2=9.452$	29.8	3.57	否
$x_{J=16.34}$	10	$\chi^2_{16.0416.34}=n_{J=16.34}s^2/\partial_0^2=6.912$	25.2	2.16	否

通过以上分析可以看出,收敛区内各个转折点的收敛程度彼此之间均无显著差异,收敛情况一致,这些点均为虚假收敛点,序列$\{DFA_J\}$和$\{VAR_J\}$只有收敛的趋势,但未真正到达收敛时刻,不存在收敛点即极端事件的阈值,说明对去除极大值后的序列$\{x1''_i\}$使用本方法进行分析时,可以判定其不存在极大值事件. 因为序列$\{VAR_J\}$越来越小,说明$\{DFA_J\}$的收敛程度越来越高,当$\{VAR_J\}$达到最小

时$\{\mathrm{DFA}_J\}$最接近于收敛点,此时所对应的x_J是$\{x1''_i\}$的最大值而非极端值. 采用相同方法对序列$\{x1''_i\}$的负值区间进行计算分析,同样,对于序列$\{x1''_i\}$的负值区间的$\{\mathrm{DFA}_J\}$和$\{\mathrm{VAR}_J\}$序列而言,也只有收敛的趋势,而无收敛时刻,不存在收敛点和极端事件的阈值. 对去除极端小值后的序列$\{x1''_i\}$使用本方法也可以判定其不存在极端小值事件. 随着J值的逐渐减小,序列$\{\mathrm{VAR}_J\}$越来越趋向于零值,表明$\{\mathrm{DFA}_J\}$的收敛程度越来越高,当$\{\mathrm{VAR}_J\}$最小时$\{\mathrm{DFA}_J\}$最接近于收敛点,此时所对应的x_J是$\{x1''_i\}$的最小值而不是极端值.

4.4. 方法检验——原始序列去除极值

根据4.2节中确定的极端事件阈值,将原始序列$\{x_i\}$中的极端事件点消去,也就是去除序列$\{x_i\}$中$x_i \geqslant 17.2$或$x_i \leqslant -16.4$点,以达到消除极端事件的目的,得到新序列$\{x2''_i\}$,其中$\{x2''_i, x2''_i = x_i - 17.2, x_i \geqslant 17.2\}$或$\{x2''_i, x2''_i = x_i + 16.4; x_i \leqslant -16.4\}$,对序列$\{x2''_i\}$采用与4.3节相同的思路和算法进行计算,对不含有极端大值或极端小值的序列,检验本方法是否能确定该序列中无极端事件.

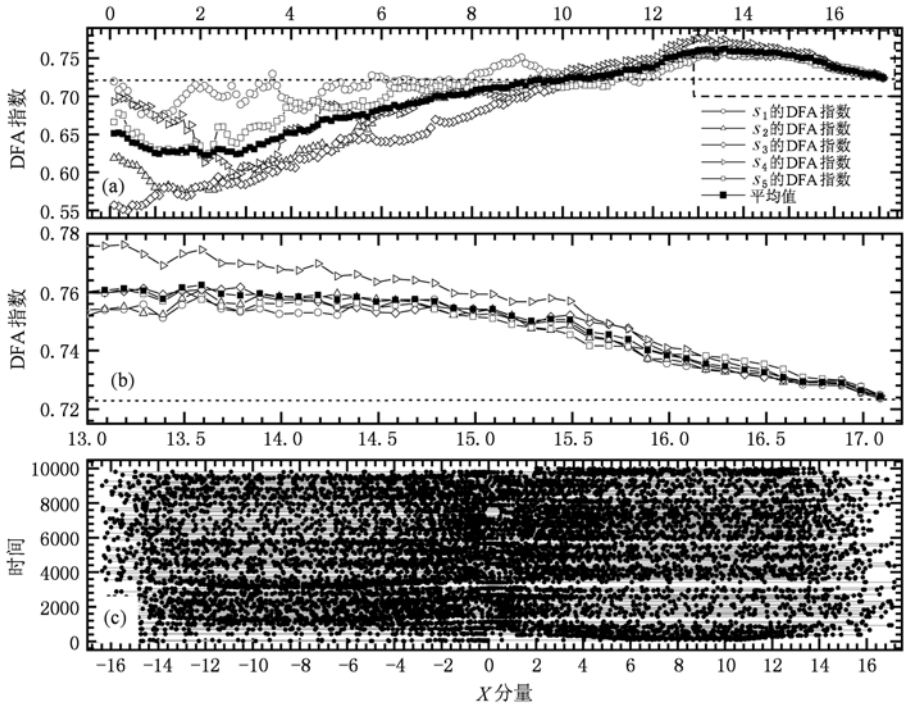

图10 (a)对$\{x2''_i\}$正值的不同区间数据点进行顺序随机化后DFA指数的变化;(b)收敛区域放大图;(c) 不含极端值的序列$\{x2''_i\}$

同样采用4.1中所述的计算方法,参数选择同4.1节. 各转折点差异的显著性检验见表4. 从表4可以看出,收敛区内各个转折点的收敛程度彼此之间均无显著差异,这些点都是虚假收敛点,即对于序列$\{x2''_i\}$的正值区间的$\{\mathrm{DFA}_J\}$和$\{\mathrm{VAR}_J\}$序列而言,也只有收敛的趋势,而无收敛时刻,不存在收敛点和极端事件的阈值. 对去除极端大值后的序列$\{x2''_i\}$使用本方法也可以判定其不存在极端大值事件. 随着J值的逐渐增大,序列$\{\mathrm{VAR}_J\}$越来越趋向于0值,表明$\{\mathrm{DFA}_J\}$的收敛程度越来越高,当$\{\mathrm{VAR}_J\}$最小时$\{\mathrm{DFA}_J\}$最接近于收敛点,此时所对应的x_J是$\{x2''_i\}$的最大值而不是极端值. 同样对于序列$\{x2''_i\}$的负值区间的$\{\mathrm{DFA}_J\}$和$\{\mathrm{VAR}_J\}$序列而言,也只有收敛的趋势,而无收敛时刻,不存在收敛点和极端事件的阈值. 对去除极端小值后的序列$\{x2''_i\}$使用本方法也可以判定其不存在极端小值事件. 随着J值的逐渐减小,序列$\{\mathrm{VAR}_J\}$越来越趋向于0值,表明$\{\mathrm{DFA}_J\}$的收敛程度越来越高,当$\{\mathrm{VAR}_J\}$最小时$\{\mathrm{DFA}_J\}$最接近于收敛点,此时所对应的x_J是$\{x2''_i\}$的最小值而不是极端值.

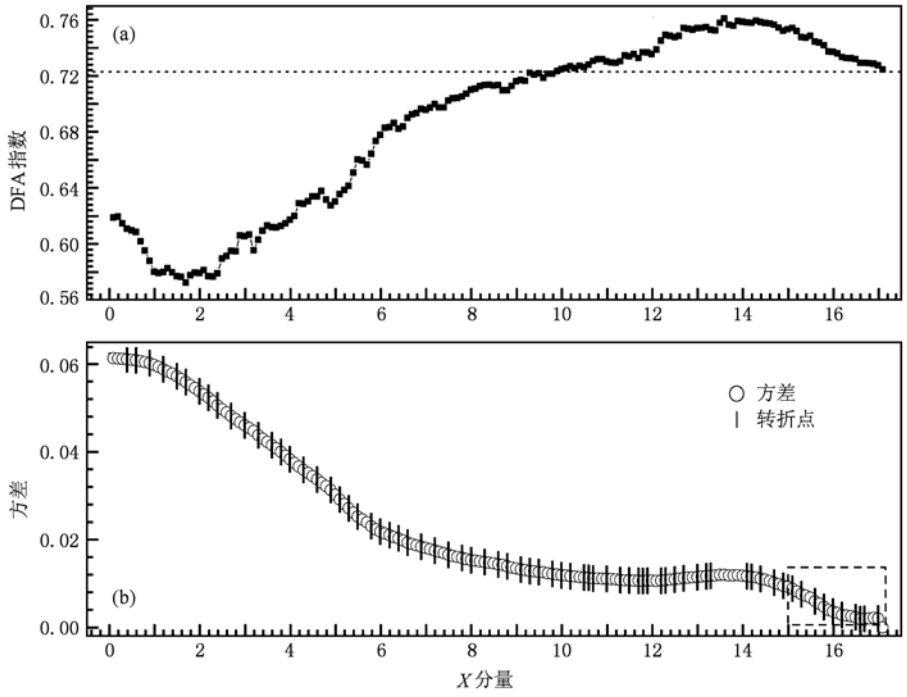

图 11 (a)序列$\{x2''_i\}$DFA 指数平均值的方差及其突变点;(b)序列$\{x2''_i\}$DFA 指数平均值

表 4 各个转折点之间差异的显著性检验

转折点	n	χ^2	$\chi^2_{(\alpha/2)}$	$\chi^2_{(1-\alpha/2)}$	差异是否显著
$x_{J=15.289}$	19	$\chi^2_{15.08915.289} = n_{J=15.289}s^2/\partial_0^2 = 13.882$	38.6	6.84	否
$x_{J=15.589}$	16	$\chi^2_{15.28915.589} = n_{J=15.589}s^2/\partial_0^2 = 9.896$	34.3	5.14	否
$x_{J=15.789}$	14	$\chi^2_{15.58915.789} = n_{J=15.789}s^2/\partial_0^2 = 8.675$	31.3	4.07	否
$x_{J=15.989}$	12	$\chi^2_{15.78915.989} = n_{J=15.989}s^2/\partial_0^2 = 7.272$	28.3	3.07	否
$x_{J=16.189}$	10	$\chi^2_{15.98916.189} = n_{J=16.189}s^2/\partial_0^2 = 5.939$	25.2	2.16	否
$x_{J=16.489}$	7	$\chi^2_{16.18916.489} = n_{J=16.489}s^2/\partial_0^2 = 4.656$	20.3	0.989	否

4.5. 使用 S-DFA 方法确定北京极端高事件的阈值

对北京 1961—2006 年逐日平均温度序列$\{x_i, i = 1,\cdots,n\}$使用替代数据方法共进行重排计算,为了更加突出图像的变化,图 12 和表 5 中涉及温度值的地方均乘以 10. 计算时局部趋势函数 $y_v(i)$ 使用二阶多项式, s 取值为 $250 \leqslant s \leqslant n/15$, n 为序列长度;取区间间隔 $d = 1.0$, $R = 0$. 计算得到序列$\{x_i\}$的 DFA 指数为 0.71425(图 12 中点线所示).

取显著性水平为 99%,即 $\alpha = 0.01$, 采用同样的计算步骤,各转折点之间的卡方检验结果如表 5 所示.

可以看出从最后一个转折点 $x_{J=388}$ 开始,与其前一个转折点 $x_{J=375}$ 之间的差异不显著,两点处的收敛情况一致;$x_{J=375}$ 与其前一个转折点 $x_{J=365}$ 之间的差异显著,并且从 $x_{J=365}$ 开始各个转折点之间也无显著差异,因此可以认为序列$\{\mathrm{DFA}_J\}$在点 $x_{J=375}$ 处开始收敛,点 $x_{J=375}$ 即为收敛点(图 12(b)中箭头表示),当 $J \geqslant 375$ 时,$\{\mathrm{DFA}_J\}$序列收敛于原始值,且不同 DFA 指数之间差异也很微小,可以确定序列 $\{x_i\}$ 极端高温事件的阈值是 37.5 ℃. 据确定的极端高温事件阈值,消除序列 $\{x_i\}$ 中的极值点或极端事件,也就是去除序列 $\{x_i\}$ 中 $x_i \geqslant 37.5$ 点,以达到消除极端事件的目的,得到新序列$\{x'_i\}$,其中$\{x'_i, x'_i = x_i - 37.516.7; x_i \geqslant 37.5\}$,对序列$\{x'_i\}$采用同样的思路和算法进行计算,以检验对于一个不含极小值的序列,本方法是否仍然有效,即能确定该序列中无极端事件,以此对本方法进行进一步检验. 通过分析,收敛区内各个转折点的收敛程度彼此之间均无显著差异,收敛情况一致,这些点均为虚假收敛点,

序列$\{\mathrm{DFA}_J\}$和$\{\mathrm{VAR}_J\}$只有收敛的趋势,但未真正到达收敛时刻,不存在收敛点即极端事件的阈值,说明对去除极大值后的序列$\{x_i'\}$使用本方法进行分析时,可以判定其不存在极大值事件.

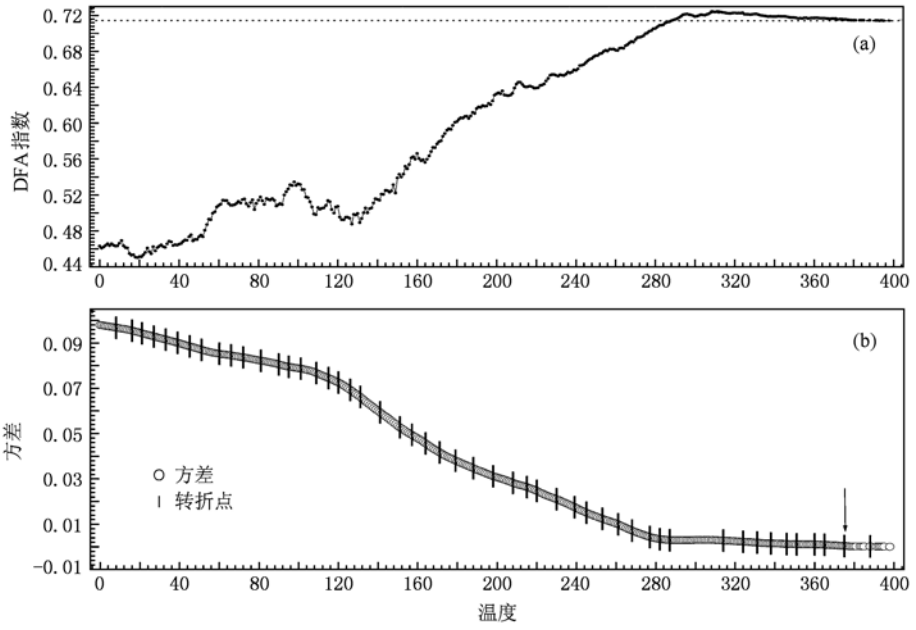

图 12 (a) 对日平均温度不同正值区间数据进行顺序随机化后 DFA 指数的变化; (b) DFA 指数的方差及其突变点

表 5 各个转折点之间差异的显著性检验

转折点	n	χ^2	$\chi^2_{(\alpha/2)}$	$\chi^2_{(1-\alpha/2)}$	差异是否显著
$x_{J=388}$	7	$\chi^2_{375388} = n_{J=388}s^2/\partial_0^2 = 1.012$	20.3	0.989	否
$x_{J=375}$	17	$\chi^2_{365375} = n_{J=375}s^2/\partial_0^2 = 3.484$	35.7	5.70	是
$x_{J=365}$	27	$\chi^2_{361365} = n_{J=365}s^2/\partial_0^2 = 19.697$	49.6	11.8	否
$x_{J=361}$	31	$\chi^2_{357361} = n_{J=361}s^2/\partial_0^2 = 28.056$	55.0	14.5	否
$x_{J=357}$	35	$\chi^2_{349357} = n_{J=357}s^2/\partial_0^2 = 31.012$	60.3	17.2	否

5. 结 论

由于极端事件或者极值事件属于小概率事件,可以认为此类事件所对应的演化状态是系统演化的极端状态,或是系统演化的异常状态,不属于系统自身正常演化状态的范畴. 基于这一物理概念,本文提出将去趋势波动分析(DFA)方法和替代数据法相结合来确定极端事件的阈值,即寻找一个临界值,当大于(小于)该临界值的数据点位置不变时,无论小于(大于)该临界值的数据点的位置如何变化,对整个序列的 DFA 指数无影响,认为该临界值为阈值. 方法的基本思路对序列中不同值域区间内数据点的位置进行随机重排,当重排数据点很少时,改变这些数据点顺序后得到的新序列的 DFA 指数收敛于原始序列的 DFA 指数. 由于这部分数据点的概率非常小,属于小概率事件,其中所包含的系统演化信息极少,其统计效应可以基本忽略,所对应的状态不属于系统的常规演化轨迹,而是系统演化的极端状态或是系统受到外界扰动而导致的极异常状态. 对于序列中那些概率密度较大的序列或者具有均匀概率分布的序列,改变其中不同值域区间内数据点的位置对序列 DFA 指数的影响也较大,它们包含了丰富的系统演化信息,不属于小概率事件的范畴,其统计效应也非常显著,属于系统演化的常规状态,以此来将系统的极端状态或异常状态同常规演化状态区分开来. 由于该方法主要将去趋势波动分析(DFA)方法和替代数据法结合使用,将其称为随机重排去趋势波动分析(S-DFA)方法.

同百分位阈值方法相比,S-DFA 方法明确指出了极端事件和非极端事件之间的临界值,并通过数

值试验和使用北京市 1961—2006 年实际温度观测资料从不同的角度对 S-DFA 方法进行了反复检验,说明本方法可以准确给出潜在动力学系统相同的、不同表现形式下的序列的极端事件阈值,得到的阈值是唯一的、可靠的,验证了 S-DFA 方法的有效性.基于 S-DFA 方法而得到的极端事件,更多关注于"极端事件是偏离系统常规演化状态的异常状态"这一特征,从系统的动力学性质出发,基于系统的动力学不变量,以此将影响和不影响该动力学不变率的事件甄别开来,进而从系统整体动力学行为的角度对极端事件加以描述和研究,因此,S-DFA 方法不仅可以用于气象领域极端天气气候事件的研究,还可用于由时间序列演化来表征系统潜在动力学系统演变的其他学科,如水文、金融、生物等.

[1] IPCC 2001 *Climate Change* (New York: Cambridge University Press) p155

[2] Soloman S, Qin D H, Manning M, Alley R B, Berntsen T 2007 *Climate Chang* (New York: Cambridge University Press) p316

[3] Xiong K G, Yang J, Wang S Q, Feng G L, Hu J G 2009 *Acta Phys. Sin.* **58** 2843 (in Chinese) [熊开国、杨 杰、万仕全、封国林、胡经国 2009 物理学报 **58** 2843]

[4] Feng G L, Dong W J, Gong Z Q, Hou W, Wan S Q, Zhi R 2006 *Nonlinear Theories and Metho-ds on Spatial-Temporal Distribution of the Observational Data* (Beijing: Metrological Press) p84 (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)]

[5] Zhang D Q, Yang J, Wang Q G, Feng G L 2009 *Acta Phys. Sin.* **58** 4354 (in Chinese) [章大全、杨 杰、王启光、封国林 2009 物理学报 **58** 4354]

[6] Easterling D R, Evans J L, Groisman P Y 2000 *Bulletin of the American Meteorological Socie-ty* **81** 417

[7] Render S, Petersen M R 2006 *Phys. Rev.* E **74** 061114

[8] Zhang D Q, Zhang L, Yang J, Feng G L 2010 *Acta Phys. Sin.* **59** 655 (in Chinese) [章大全、张 璐、杨 杰、封国林 2010 物理学报 **59** 655]

[9] Yang J, Hou W, Feng G L 2010 *Acta Phys. Sin.* **59** 664 (in Chinese) [杨 杰、侯 威、封国林 2010 物理学报 **59** 664]

[10] Wan S Q, Gu C H, Kang J P 2010 *Acta Phys. Sin.* **59** 676 (in Chinese) [万仕全、顾承华、康建鹏等 2010 物理学报 **59** 676]

[11] Peng C K, Buldyrev S V, Havlin S 1994 *Phys. Rev.* E **49** 1685

[12] He W P, Feng GL, Wu Q, Wan S Q, Chou J F 2009 *Non. Proc. Geophys.* **15** 601

[13] Janosi I M, Janecsko B, Kondor I 1999 *Physica* A **269** 111

[14] Ausloos M 2000 *Physica* A **285** 48

[15] Fraedrich K 2002 *Stoch. Dynam.* **2** 403

[16] Lux T, Marehesi M 1999 *Nature* **397** 498

[17] Yang P, Hou W, Feng G L 2008 *Acta Phys. Sin.* **57** 5333 (in Chinese) [杨 萍、侯 威、封国林 2008 物理学报 **57** 5333]

[18] Panlov A N, Sosnovtseva O V, Ziganshin A R 2002 *Physica* A **316** 233

[19] Lee J M, Kin D J, Kim I Y 2002 *Computers in Biology and Medicine* **32** 37

[20] Ott E 1993 *Chaos in dynamical systems* (Cambridge, UK: Cambridge University Press) p305

[21] Theiler J, Linsay P S 1993 *M Rubin Time Series Prediction: Forecasting the Future and Understanding the Past* (Addison-Wesley, Reading Mass. Press) p429

[22] Timmer J 1998 *Phys. Rev.* E **58** 5153

[23] Theiler J, Eubank S, Longtin A 1992 *Physica* D **58** 77

[24] Kugiumtzis D 2000 *Phys. Rev.* E **62** 25

[25] Timmer J 2000 *Phys. Rev. Lett.* **85** 2647

[26] Stam C J, Mpijn J P, Pritchard W S 1998 *Physica* D **112** 361

[27] Bernaola G P 2001 *Phys. Rev. Lett.* **87** 168

[28] Oliver J L, Bernaola G P, Carpena P, Román R R 2001 *Gene* **276** 47

[29] Chernoff H, Lehmann E L 1954 *The Annals of Mathematical Statistics* **25** 579

[30] Plackett R L 1993 *International Statistical Review* **51** 59

[31] Greenwood P E, Nikulin M S 1996 Wiley, New York. ISBN 047155779X

Stochastially re-sorting detrended fluctuation analysis: a new method to define the threshold of extreme event*

Hou Wei[1)2)†] Zhang Da-Quan[1)] Zhou Yun[1)3)] Yang Ping[4)]

1) (*National Climate Center*, *Beijing* 100081, *China*)

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

3) (*College of Physical Science and Technology*, *Yangzhou University*, *Yangzhou* 225002, *China*)

4) (*Institute of Urban Meteorology of Beijing*, *China Metrology Administration*, *Beijing* 100089, *China*)

(Received 15 November 2010; revised manuscript received 7 January 2011)

Abstract

By combining detrended fluctuation analysis (DFA) method with surrogate data method, and using the Heuristic segmentation algorithm as well as Chi-Square statistics, we develop a new method to determine the threshold of extreme events, e. g. stochastically re-sorting detrended fluctuation analysis (S-DFA) method. The S-DFA method has a certain phsical background, when the occurrence rate of the data is small, then these data belong to little-probability events and they contain so little information about the dynamic system, the states corresponding to these data are abnormal states or extreme states of the system. When the occurrence rate of the data is large or even in distribution these data do not belong to little-probability events and they contain much information about the system, the states corresponding to these data are normal states of the system. Compared with the Percentile curves method, the S-DFA method gives the critical value between extreme event and non-extreame event, which is definite and unique. We also extensively validate the effectiveness of S-DFA method through extreme event detection.

Keywords: detrended fluctuation analysis, surrogate data, extreme event, threshold

PACS: 92.70.Aa

* Project supported by the National Natural Science Foundations of China (Grant Nos. 41005043, 40905034), the Global Change Research of Major National Scientific Research Plan of China (Grant No. 2010CB950504), and the State Key Program of Science Technology of China (Grant No. 2007BAC29B01).

† E-mail: hou_w@sohu.com