PHYSICAL REVIEW E **86**, 026103 (2012)

Emergence of patterns in random processes

William I. Newman[*]
Departments of Earth and Space Sciences, Physics and Astronomy, and Mathematics, University of California, Los Angeles, California 90095-1567, USA

Donald L. Turcotte[†]
Department of Geology, University of California, Davis, California 95616, USA

Bruce D. Malamud[‡]
Department of Geography, King's College London, Strand, London WC2R 2LS, United Kingdom
(Received 3 April 2012; published 7 August 2012)

Sixty years ago, it was observed that any independent and identically distributed (i.i.d.) random variable would produce a pattern of peak-to-peak sequences with, on average, three events per sequence. This outcome was employed to show that randomness could yield, as a null hypothesis for animal populations, an explanation for their apparent 3-year cycles. We show how we can explicitly obtain a universal distribution of the lengths of peak-to-peak sequences in time series and that this can be employed for long data sets as a test of their i.i.d. character. We illustrate the validity of our analysis utilizing the peak-to-peak statistics of a Gaussian white noise. We also consider the nearest-neighbor cluster statistics of point processes in time. If the time intervals are random, we show that cluster size statistics are identical to the peak-to-peak sequence statistics of time series. In order to study the influence of correlations in a time series, we determine the peak-to-peak sequence statistics for the Langevin equation of kinetic theory leading to Brownian motion. To test our methodology, we consider a variety of applications. Using a global catalog of earthquakes, we obtain the peak-to-peak statistics of earthquake magnitudes and the nearest neighbor interoccurrence time statistics. In both cases, we find good agreement with the i.i.d. theory. We also consider the interval statistics of the Old Faithful geyser in Yellowstone National Park. In this case, we find a significant deviation from the i.i.d. theory which we attribute to antipersistence. We consider the interval statistics using the AL index of geomagnetic substorms. We again find a significant deviation from i.i.d. behavior that we attribute to mild persistence. Finally, we examine the behavior of Standard and Poor's 500 stock index's daily returns from 1928–2011 and show that, while it is close to being i.i.d., there is, again, significant persistence. We expect that there will be many other applications of our methodology both to interoccurrence statistics and to time series.

DOI: 10.1103/PhysRevE.86.026103 PACS number(s): 89.75.Kd, 05.45.Tp, 02.50.Ey

I. INTRODUCTION

It is natural to seek patterns in most everything that we encounter, prompting the question whether these patterns are real. Indeed, we normally expect that observations are due to some deterministic cause. Accordingly, we expect that "random uncorrelated events," in space and/or time, should be devoid of any underlying order. However, in a mathematical sense, it is plausible for some forms of order to exist in random data. A familiar example emerges from the normality (or Gaussian behavior) observed in many statistical distributions of values as a consequence of the central limit theorem [1]. Here we examine possible signatures of order that we detect with our eyes but do not satisfy currently established criteria for order. For example, it is natural to consider different points in space or time as being related if they are "close" to each other [2]. By making such an association, it is possible to identify "clusters" based on nearest-neighbor relationships. In addition, when examining data from experiments or field observations, it is natural to find patterns in the distribution of peaks in a data set. In this paper, we consider whether patterns associated with clusters and/or peak values could be an outcome of independent and identically distributed (i.i.d.) behavior. We will consider random events in time (e.g., a Poisson process), an i.i.d. time series (e.g., Gaussian white noise), and a variety of applications.

Many years ago, the Columbia biologist Lamont Cole consulted his mathematician colleague Mark Kac regarding possible periodicity in field-based observations of animal populations. The annual estimates of the population of a species constitutes a time series. Peak populations were associated with years in which the population was greater than the population in the previous or succeeding years. Cole [3] presented data obtained by others 30 years earlier for the Arctic fox and wolf populations in Canada. The data suggested the existence of a 3- to 4-year cycle in going from peak-to-peak. The prevailing view was that this was the outcome of a predator-prey cycle in a complex ecosystem. We denote the number of years before a new peak occurs by m. Cole [3] noted that the observed or sample mean of the m values was $\bar{m} \approx 3$ to 4 and matched the observed mean obtained from uniformly and independently distributed random numbers plotted in the same way. As a result of this consultation with

[*]win@ucla.edu
[†]dlturcotte@ucdavis.edu
[‡]bruce.malamud@kcl.ac.uk

Kac by Cole, Kac provided an elegant argument, presented in Cole [3] but neither extended nor published elsewhere. This argument demonstrated that the wildlife data could be entirely due to random processes. Kac's argument considered three consecutive points and observed that if the data are from independent and identically distributed (i.i.d.) random variables, then the probability that the middle point is the highest is 1/3 independent of the underlying probability distribution. Kac, in Ref. [3], therefore established that, on average, there are three events before a new peak occurs.

However, Kac's resolution of this problem is only a partial one. What distribution of sequence lengths can we expect? How would this distribution, in terms of event counts, differ from that obtained by some deterministic biological or physical model? Unfortunately, the data sets used by Cole [3] were not sufficiently long to provide good statistics. Consequently, more extensive data sets are essential. Imperial College, London, UK, and its NERC Center for Population Biology [4] has amassed a very comprehensive set of population biology-related data. Kac's observation indicates that the null hypothesis that the data are consistent with statistically independent random numbers cannot be dismissed. There exists substantial controversy, e.g., Refs. [5,6], regarding how well various predator-prey models fit observational data.

It is important, and a main purpose of this paper, to determine the distribution of peak-to-peak sequence lengths m for a random uncorrelated process. We then use this information to assess whether data sets in a variety of applications are the outcome of i.i.d. processes or are the product of a complex dynamic, possibly maintaining memory effects, that provide a deterministic foundation for the observed data.

In this paper, we will also consider the cluster statistics of *any* i.i.d. random variable that represents a point process in time. A cluster is defined to be a set of events that appear to be grouped in time, i.e., are mutually closest in time. The underlying theory for this cluster formation is given in Ref. [2] as well as elaborated on in the context of hierarchical and multidimensional behavior. The basic problem we will consider is as follows: Consider a sequence of events in time and we display each event as a point on a time line. We then draw an arrow from each point to its nearest neighbor in time, either the event immediately before it or immediately after it. The edge of a cluster is defined by an interval containing no arrows. In this way, we have constructed a set of directed graphs, each of which has the appearance of a distinct cluster. In practical terms, the edges of a cluster are bounded by time intervals that are longer than their two respective adjacent interval values. This condition for time intervals is identical to the local maximum value in the time series discussed above.

We will consider two i.i.d. processes in this paper. The first is a sequence of events in time in which intervals are selected randomly and independently from a statistical distribution. The statistical distribution's yields cluster sizes in a closed-form result for any population and the population mean has a value of $\langle m\rangle = 3$ as given by Newman [2]. The second process is an uncorrelated time series in which the values are selected randomly and independently from a statistical distribution. The statistical distribution of peak-to-peak sequence lengths also has a mean value $\langle m\rangle = 3$. We show that the sequence lengths for Monte Carlo simulations for a Gaussian white noise are in excellent agreement with our theory. We then extend our analysis to Brownian motion and show that the statistical distribution of peak-to-peak sequence lengths has a mean value $\langle m\rangle = 4$. We shall consider a variety of problems emerging from natural hazards to see if they are statistically consistent with the null hypothesis emerging from a random process. We also consider random processes that include "memory effects," such as those encountered in the Langevin equation [7,8] of statistical mechanics in the theory of Brownian motion.

II. PROBABILISTIC DESCRIPTION OF PROBLEM AND KAC'S SOLUTION

Suppose that we have three sequential events with amplitudes x_{n-1}, x_n, and x_{n+1}, where n designates the middle event, i.e., x_n is the amplitude of the n-th event. Without loss of generality, we will replace n with 0. Suppose, now, that these events are independent and can be described by the same cumulative distribution function $0 \leqslant P(x) \leqslant 1$ for $0 \leqslant x < \infty$. We can modify our definitions to allow for a doubly infinite or a finite domain for the random variable x. For convenience, although this is not necessary for our derivation, we shall assume that $P(x)$ is differentiable and has a probability density function $p(x)$ defined by

$$p(x) \equiv \frac{dP(x)}{dx} \geqslant 0 \quad \text{and} \quad P(x) = \int_0^x p(x')\,dx'. \tag{1}$$

It follows, therefore, that the probability $\mathcal{P}$ that the amplitude of the middle event, designated by x_0, is greater than that of its neighbors x_{-1} and x_{+1} is given by

$$\mathcal{P} = \int_0^\infty p(x_0)\,dx_0 \int_{x_{-1}=0}^{x_0} p(x_{-1})\,dx_{-1} \int_{x_1=0}^{x_0} p(x_1)\,dx_1. \tag{2}$$

We observe that the integrals over x_{+1} can be written $P(x_0)$, which can also be said for the integral over x_{-1}. Moreover, we can express $p(x_0)\,dx_0 = dP(x_0)$. Therefore, Eq. (2) becomes, where we no longer have a need for the subscript "0,"

$$\mathcal{P} = \int_{x=0}^\infty P^2(x)\,dP(x) = \int_0^1 P^2\,dP = \frac{1}{3}. \tag{3}$$

Remarkably, this result does not depend on the explicit nature of the underlying probability distribution function $P(x)$. While this result was immediately obvious using Kac's [3] argument, we have now introduced the methodology that we will employ in calculating the distribution of sequential event lengths, m.

In order to illustrate this behavior, consider Fig. 1. This figure has two components. We will begin by considering the first of these, Fig. 1(a), which contains a sequence of eight events in time. We will regard this as a time line which begins at time 0. Each point on the time line can be regarded as an event or a milestone. Moreover, each point corresponds to the sum of the previous point's location and a selected interval, namely the x_i, which we now identify as intervals τ_i. While we refer to this axis as representing time, it can also represent distance and other variables in different applications. From each point, we draw a line terminating in an arrow to its nearest neighbor. We observe that the configuration that arises displays gaps which are longer than the time intervals flanking the gaps on both

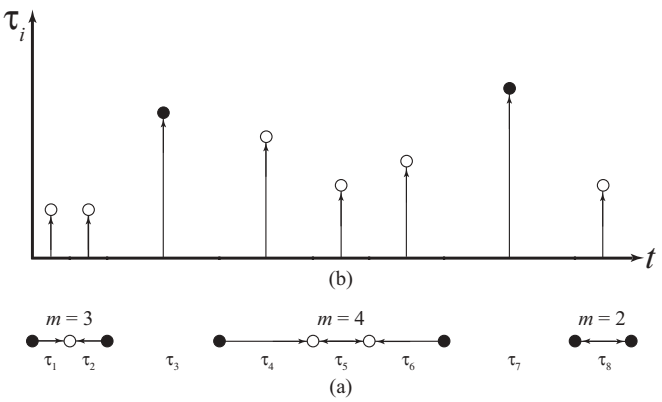

FIG. 1. In (a), we show a sequence of nine events in time t. The interoccurrence times are $\tau_1, \tau_2, \ldots, \tau_8$. We consider nearest neighbors in time clustering. The events at the ends of each cluster are solid circles, while the interior events are open circles. We show arrows from each point to its nearest neighbor. Intervals with no intervening arrow correspond to intervals that are longer than those on either side. The corresponding nearest-neighbor cluster sizes are $m = 3$, 4, and 2. In (b), we give a time-series representation of the sequence of interoccurrence times τ_i shown in (a). A peak value is defined to be larger than its two neighboring values. The peaks in the time series are shown with solid circles and lower values as open circles. The intervals τ_3 and τ_7 are peaks. The peak-to-peak sequence lengths are $m = 3$, 4, and 2. We see that the time series representation in (b) is identical to the time line representation in (a). The nearest-neighbor cluster statistics and the peak-to-peak sequence statistics are identical.

sides. Similarly, we observe intervals on the time line that are shorter than those of the intervals flanking them. These locally shortest intervals have two arrows, each pointing in the opposite direction. We refer to such intervals and the presence of two arrows as being "reflexive." We observe the emergence of groups of points, interconnected by these arrows, that we will refer to as "clusters." Newman [2] showed that the kind of construction we have presented at the base of this diagram is a (random) directed graph and showed, analytically, that the mean number of events in a cluster is $\langle m \rangle = 3$, so long as the distribution of time intervals is an i.i.d. random variable. For the example given in Fig. 1(a), we have cluster sizes $m = 3$, 4, and 2.

In Fig. 1(b), we give a time series representation of the sequence of interoccurrence times τ_i shown in Fig. 1(a). A peak value in the time series is defined to be a value larger than the two neighboring values. We define a sequence length to be the number of events that occur until the last event is another maximum. The sequence lengths in Fig. 1(b) are $m = 3$, 4, and 2, which is equal to the cluster sizes in Fig. 1(a).

In Fig. 1, we have presented two seemingly different constructions which, remarkably, appear to have a common feature, namely that there are, on average, three intervals per cluster and three events per sequence. To establish this connection and the geometries that make this rigorous, the two parts of this figure are intimately related. We began by selecting time intervals between events and produced the time line as a directed graph using the prescription above. In the middle of each time interval, we then introduced a point with an amplitude x_i proportional to the time interval τ_i. Hence, the role of time intervals, for the directed graph, and of amplitudes, for the line graph, are equivalent. We further note that the peaks (or local maxima) are situated directly above the middle of the gaps between the clusters. Further, the valleys (or local minima) are situated directly above the middle of the reflexive intervals in each cluster. Therefore, it should come as no surprise that the derivation we are about to provide for the distribution of peak-to-peak sequence lengths is essentially equivalent to that provided in Ref. [2] for one-dimensional clustering. The construction given in Fig. 1 directly relates the cluster statistics in the temporal occurrence of events to any time series, it provides a visual verification of Kac's [3] observation that there are on average three events per "cycle" (in the biological problem) or, more generally, peak-to-peak event sequence.

Having established this algebraic/geometrical relationship, we must now analyze the structure of all possible peak-to-peak sequences. In Fig. 2, we illustrate the "taxonomy" of a time series that emerges. The shortest such sequence is of length $m = 2$, where there is but one intervening point x_0 between peaks situated at x_{-1} and x_{+1}. However, for these points to be peaks, they must be higher than the points that flank them. Therefore, we require

$$0 < x_{-2} < x_{-1} > x_0 < x_{+1} > x_{+2} > 0. \tag{4}$$

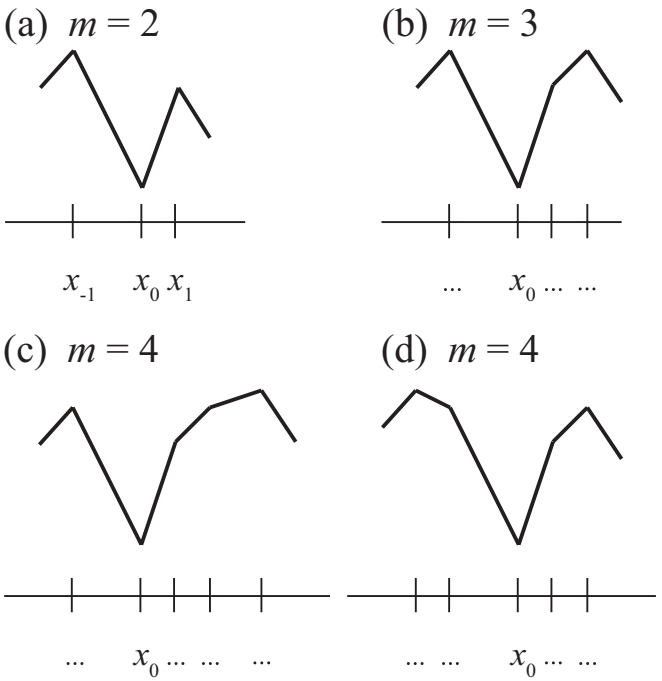

FIG. 2. Examples of sequence lengths associated with maxima in time series. (a) Illustration with $m = 2$. A maximum value is followed by one low value and then another maximum value. (b) Illustration with $m = 3$. A maximum value is followed by two low values and then another maximum value. [(c) and (d)] Illustrations with $m = 4$. A maximum value is followed by three low values and then another maximum value. The sequence in (d) is identical to the central sequence in Fig. 1.

We have shown strict inequalities in this expression; without loss of generality, we can include the equality in each term inasmuch as they present a situation with measure zero. In simulation or data-based applications of the theory, this potential for equal values requires some care.

Illustrations of sequence lengths m between maximum values in a time series are given in Fig. 2. In Fig. 2(a), a single low value lies between the two maximum values. Thus, the sequence length is $m = 2$. In Fig. 2(b), a maximum value is followed by a minimum value and then two events with increasing amplitudes. The second of these is a maximum value. This configuration, and its mirror image, describe a sequence of length $m = 3$. We expect, as a result, that we will need to identify configurations where the minimum value is separated by i events from its peak or local maximum value on its right and j events from its peak or local maximum value on its left. Both $i \geqslant 1$ and $j \geqslant 1$ with $m = i + j$. In Fig. 2(c), a maximum value is followed by one event with diminishing amplitude and three events with increasing amplitude so $j = 1$ and $i = 3$; the mirror image of this configuration must also be considered corresponding to $m = 4$ with $j = 3$ and $i = 1$. In Fig. 2(d), with have a maximum value followed by two events with diminishing amplitudes ($j = 2$) and two events with increasing amplitudes ($i = 2$) again with $m = 4$. In each case, the fourth event after a maximum value is another maximum value. The sequence in Fig. 2(d) is identical to the central sequence in Fig. 1.

This illustration establishes how we need to construct a taxonomy describing all possible peak-to-peak event sequences. In particular, we employ as our starting point the local minimum or "valley" in a peak-to-peak sequence. We need to consider the circumstance where the distance from the valley to its right-hand peak is i events and to its left-hand peak is j events. Let us now calculate the probability of the simplest case, the two-event sequence where $i = j = 1$ depicted in Fig. 2(a), using the probabilistic framework established previously for Kac's result in Eq. (2). Given the inequality in Eq. (4), it follows that the probability $\mathcal{P}_2$ that a five-event sequence can contain a peak-to-peak sequence of length 2 is

$$\mathcal{P}_2 = \int_0^\infty p(x_0)\,dx_0 \left[\int_{x_1=x_0}^\infty p(x_1)\,dx_1 \int_{x_2=0}^{x_1} p(x_2)\,dx_2\right] \times \left[\int_{x_{-1}=x_0}^\infty p(x_{-1})\,dx_{-1} \int_{x_{-2}=0}^{x_{-1}} p(x_{-2})\,dx_{-2}\right]. \quad (5)$$

We observe in the previous expression that the second term in square brackets is formally identical to the first. Moreover, each of these terms in brackets corresponds to the probability that the peak is removed by 1 event from the valley. It is convenient to express this term as

$$F(x_0,1) \equiv \int_{x_1=x_0}^\infty p(x_1)\,dx_1 \int_{x_2=0}^{x_1} p(x_2)\,dx_2 = \int_{x_1=x_0}^\infty P(x_1)\,p(x_1)\,dx_1. \quad (6)$$

Intuitively, this $F(x_0,1)$ term provides the probability that the valley with amplitude x_0 is removed by only one point from the peak in question. Since the latter integral runs from $x_1 = x_0$ to ∞, it is convenient to employ the complement of the distribution defined by

$$Q(x) \equiv 1 - P(x) \quad \text{and} \quad p(x)\,dx = -dQ(x). \quad (7)$$

This substitution proved to be invaluable in Ref. [2] whose details are very similar to the present problem. As a consequence, we observe that

$$F(x_0,1) = \int_\infty^{x_1=x_0} [1 - Q(x_1)]\,dQ(x_1) = \left[\frac{Q(x_1)}{1} - \frac{Q^2(x_1)}{2}\right]_\infty^{x_1=x_0} = \frac{Q(x_0)}{1} - \frac{Q^2(x_0)}{2}. \quad (8)$$

It is useful to think of this term $F(x_0,1)$ as designating the probability, given that the minimum amplitude is at x_0, of having the peak to the right 1 event later, i.e., at the next event with amplitude x_1 and with the succeeding point $x_2 < x_1$. (Momentarily, we will see how to exploit this result in a recursive definition.) It follows, therefore, that

$$\begin{aligned}\mathcal{P}_2 &= \int_0^\infty p(x_0)\,F^2(x_0,1)\,dx_0 \\ &= \int_{x_0=\infty}^0 \left[\frac{Q(x_0)}{1} - \frac{Q^2(x_0)}{2}\right]^2 dQ(x_0) \\ &= \int_0^1 \left[\frac{Q}{1} - \frac{Q^2}{2}\right]^2 dQ \\ &= \left[\frac{Q^3}{3} - \frac{Q^4}{4} + \frac{Q^5}{20}\right]_0^1 = \frac{2}{15}. \end{aligned} \quad (9)$$

What this means is that the number of two-event sequences will be numerically equal to $2/15$ times the total number of events. Remarkably, the functional dependence of P (and Q) on the amplitude x_0 simply disappeared. This is reminiscent of the "record-breaking" statistical theory [9–12] whose outcome is independent of the details resident in the underlying distribution function.

Suppose, now that we wanted to calculate the probability of the three-event sequence depicted in Fig. 2(b). It follows directly that each of the two (mirror-image) configurations that can produce a three-event sequence has the probability

$$\mathcal{P}_3 = \int_0^\infty p(x_0)\,F(x_0,1)\,F(x_0,2)\,dx_0. \quad (10)$$

We associate the term in $F(x_0,1)$ with the peak at x_{-1} and the reduced amplitude point x_{-2}, as before. However, we have now introduced a term $F(x_0,2)$ which corresponds to

$$F(x_0,2) = \int_{x_1=x_0}^\infty p(x_1)\,dx_1 \int_{x_2=x_1}^\infty p(x_2)\,dx_2 \int_{x_3=0}^{x_2} p(x_3)\,dx_3. \quad (11)$$

From the structure of the latter, we observe that

$$\begin{aligned}F(x_0,2) &= \int_{x_1=x_0}^\infty p(x_1)\,F(x_1,1)\,dx_1 \\ &= \int_{x_1=\infty}^0 \left[\frac{Q(x_1)}{1} - \frac{Q^2(x_1)}{2}\right] dQ(x_1) \\ &= \frac{Q^2(x_0)}{2!} - \frac{Q^3(x_0)}{3!}. \end{aligned} \quad (12)$$

Indeed, by induction, we can show that

$$F(x_0,n) = \int_{x_1=x_0}^{\infty} p(x_1)\,F(x_1,n-1)\,dx_1 = \frac{Q^n(x_0)}{n!} - \frac{Q^{n+1}(x_0)}{(n+1)!}, \tag{13}$$

which describes the probability, given that the minimum amplitude is at x_0, of having the peak to the right (or to the left) n events later (or earlier). We are now able to calculate the probability for all possible configurations of peaks and valleys.

In the foregoing, we began by evaluating the probability of a configuration where there was a minimum amplitude event x_0 flanked on each side by a peak. In other words, the valley or minimum amplitude event was related to one event on its left (x_{-1}) and one event on its right (x_{+1}). We now generalize this probability, incorporating i events on its left and j events on its right, using the definitions we have just established by defining a probability

$$\begin{aligned}\mathcal{P}(i,j) &= \int_{x=0}^{\infty} p(x)\,F(x,i)\,F(x,j)\,dx \\ &= \int_{x=\infty}^{0} \left[\frac{Q^i(x)}{i!} - \frac{Q^{i+1}(x)}{(i+1)!}\right]\left[\frac{Q^j(x)}{j!} - \frac{Q^{j+1}(x)}{(j+1)!}\right] dQ(x) \\ &= \int_0^1 \left[\frac{Q^i}{i!} - \frac{Q^{i+1}}{(i+1)!}\right]\left[\frac{Q^j}{j!} - \frac{Q^{j+1}}{(j+1)!}\right] dQ, \end{aligned} \tag{14}$$

where we no longer need to employ a subscript for the event amplitude x_0. Following a little algebra, we obtain

$$\mathcal{P}(i,j) = \frac{1}{i!\,j!}\left[\frac{1}{i+j+1} - \frac{1}{(i+1)(j+1)} + \frac{1}{(i+1)(j+1)}\,\frac{1}{i+j+3}\right]. \tag{15}$$

Now, suppose that we want to consider all event sequences where the interval from peak-to-peak is m. We then must consider the sum of all $i \geqslant 1$ and $j \geqslant 1$ such that $m = i + j$. For this purpose, we define by $\mathcal{F}(m)$ the fraction of clusters of length $m \geqslant 2$ formed from an individual point by

$$\mathcal{F}(m) \equiv \sum_{i=1}^{m-j} \mathcal{P}(i,m-j) = \frac{2^m\,(m-1)}{(m+1)!\,(m+3)}. \tag{16}$$

This expression is identical to that obtained in Ref. [2] for the fraction or rational number of clusters of length m formed from an individual event. Finally, we can show, following a little algebra, that

$$\sum_{m=2}^{\infty} \mathcal{F}(m) = \frac{1}{3}. \tag{17}$$

This is equivalent to Kac's result and Eq. (3). In words, when we sum this quantity over $m \geqslant 2$, we obtain the fraction or rational number of sequences formed from an individual event in a time series.

In these calculations, we have been exploring the probability $\mathcal{F}(m)$ that a given peak-to-peak sequence has length m. To convert this to the probability that a given event belongs to a sequence of length m, we multiply the former sequence-based fraction $\mathcal{F}(m)$ by the number of events m associated with it, namely

$$\Pi(m) = m\,\mathcal{F}(m) = \frac{m\,2^m\,(m-1)}{(m+1)!\,(m+3)}, \tag{18}$$

and the sum of all such probabilities,

$$\sum_{m=2}^{\infty} \Pi(m) = 1, \tag{19}$$

as we expect.

The quantity $\mathcal{F}(m)$, when multiplied by the number of events N in a data set, provides the expected number of event sequences of length m (apart from end effects). Since that evaluation corresponds to a counting experiment, the uncertainty or standard error in such an estimate is $\sqrt{\mathcal{F}(m) \times N}$.

We observed that $\mathcal{F}(m)$ was proportional to the probability of a sequence being of length m. In order to convert it to a probability, we must normalize it according to Eq. (17). Therefore, we now define the normalized probability quantity $f(m)$ by

$$f(m) = \frac{\mathcal{F}(m)}{\sum_{n=2}^{\infty} \mathcal{F}(n)} = \frac{3 \times 2^m\,(m-1)}{(m+1)!\,(m+3)}, \tag{20}$$

where $f(m)$ is the normalized probability that a sequence has a length m. The (population) mean sequence length $\langle m \rangle$ is given by

$$\langle m \rangle = \sum_{m=2}^{\infty} m f(m) = 3. \tag{21}$$

Thus, under very general conditions, the mean sequence length $\langle m \rangle$ for an i.i.d. process is 3. When we consider a data set with N elements that contains M sequences (or clusters), we will calculate the sample mean $\bar{m}$ according to

$$\bar{m} = \frac{N}{M}, \tag{22}$$

where N is the number of events in the time series and we have introduced the "overbar" to designate a sample average. Furthermore, the variance σ^2 of the population may be expressed,

$$\sigma^2 = \sum_{m=2}^{\infty} [m - \langle m \rangle]^2 f(m) = 3\,e^2 - 21 \approx 1.167\,168\,30. \tag{23}$$

Accordingly, the standard error in a sample will be approximately the square root of the population variance given in Eq. (23), namely 1.080 355 636, divided by $\sqrt{N-1}$. The probabilities $f(m)$ that a sequence has a length m from Eq. (20) are given in Table I.

Finally, it is important to establish for simulations an empirical estimate or sample mean of the fraction of sequences that have length m. We will designate this (sample) estimate of the fraction using an "overbar," i.e., as $\bar{f}(m)$.

It is common in probability theory [1] to construct a generating function $g(x)$ according to

$$g(x) \equiv \sum_{m=2}^{\infty} f(m)\,x^m. \tag{24}$$

TABLE I. Probability $f(m)$ that an i.i.d. sequence has a length m as obtained from Eq. (20).

m	$f(m)$	Decimal
2	$\frac{2}{5}$	0.400 000
3	$\frac{1}{3}$	0.333 333
4	$\frac{6}{35}$	0.171 429
5	$\frac{1}{15}$	0.066 667
6	$\frac{4}{189}$	0.021 164
7	$\frac{1}{175}$	0.057 143
8	$\frac{2}{1\,485}$	0.013 468
9	$\frac{4}{14\,175}$	0.002 822
10	$\frac{4}{75\,075}$	0.000 533

In order to do this, we insert Eq. (20) for $f(m)$ into the expression for the generating function Eq. (24). We inspect the resulting expression for terms in the power series that have the appearance of an exponential and proceed to deconstruct it accordingly. Following a derivation given in Newman [2], we observe that

$$g(x) = \frac{3\,(x-1)^2\,\exp(2x)}{2\,x^3} + \frac{2\,x^3+3\,x^2-3}{2\,x^3} = \frac{2}{5}\,x^2 + \frac{1}{3}\,x^3 + \frac{6}{35}\,x^4 + \frac{1}{15}\,x^5 + \cdots. \tag{25}$$

As a consequence of our existing definition Eq. (20), $g(x)$ varies smoothly from 0 to 1 as x goes from 0 to 1. We now extend our analysis to problems associated with random processes with memory, utilizing the Langevin equation of kinetic theory and Brownian motion as our paradigm.

III. BROWNIAN MOTION AND THE LANGEVIN EQUATION

Our discussion thus far has focused on problems where the underlying physics has no memory; it constitutes an independent and identically distributed random process. There are, however, many problems in statistical physics that do possess memory and an appreciation of its influence is important here. Brownian motion describes the irregular motions exhibited by small grains or particles immersed in a fluid and undergoing rapid agitation by collisions with much smaller particles. The velocity of a free particle $u(t)$ in the absence of an external field of force is generally given by Langevin's equation [7,8]

$$\frac{du(t)}{dt} = -\beta\,u(t) + A(t). \tag{26}$$

We will employ only part of the classic derivation essential to our discussion in one dimension. Here, $-\beta\,u(t)$ describes the dynamical friction force with β related to the collision rate with smaller particles, and $A(t)$ describes the fluctuations associated with the Brownian motion and is regarded as being temporally decorrelated and can be regarded as an i.i.d. variable.

The solution of Eq. (26) can be written

$$u(t) = u_0\,\exp[\beta(t_0-t)] + \int_0^{t-t_0} A(t-\tau)\,\exp(-\beta\tau)\,d\tau, \tag{27}$$

where $u_0 = u(t_0)$ is the initial condition. Supposing that $\beta(t-t_0) \gg 1$, this can be approximated as

$$u(t) \approx \int_0^{\infty} A(t-\tau)\,\exp(-\beta\tau)\,d\tau. \tag{28}$$

Suppose, now, that we discretize Eq. (26), forming a first-order finite difference equation, where $t_i = t_0 + i\,\Delta t$. In other words, we set set $u_i = u(t_i)$, $A_i = A(t_i)$ and obtain

$$\frac{u_{i+1}-u_i}{\Delta t} \approx -\beta u_i + A_i \tag{29}$$

so

$$u_{i+1} = (1-\beta\,\Delta t)\,u_i + \Delta t\,A_i. \tag{30}$$

For convenience, we will replace $\Delta t\,A_i$ with the i.i.d. variable η_i and introduce a quantity α defined by

$$\alpha \equiv 1-\beta\,\Delta t \approx \exp(-\beta\,\Delta t), \tag{31}$$

thereby allowing us to write the recursion relation

$$u_{i+1} = \alpha\,u_i + \eta_i. \tag{32}$$

We observe that this finite difference equation can be solved immediately to give

$$u_{n+1} = \sum_{i=0}^{n} \alpha^i\,\eta_{n-i} + \alpha^{n+1}\,u_0. \tag{33}$$

We normally regard $\alpha < 1$ and note that $\alpha^k \approx \exp(-k\beta\,\Delta t)$ for $k = 0,1,\ldots,$ which renders Eq. (33) as a discretized version of Eq. (27). We observe that Eq. (32) is equivalent to a first-order autoregressive process [13,14] which is frequently encountered in time-series analysis. Interestingly, Cole [3] and other biologists have speculated on the possibility that observed time series for wildlife populations might be better described by running sums of random numbers.

There are two limits in which we wish to consider Eq. (33), as well as the intervening range for α. In the first limit, we let $\alpha = 0$, which implies that $\beta\,\Delta t \gg 1$ by the exponential in Eq. (31). In other words, $\alpha = 0$ implies that the velocity u_{n+1} maintains no memory of its past value and becomes the current value of the fluctuation η_n. In this case, we expect that the fluctuating force is i.i.d. and Gaussian, i.e., has the character of Gaussian white noise. We will consider this topic in the next section. For $0 < \alpha < 1$, the sum in Eq. (33) remains well defined and is associated with statistical mechanical principles, such as the "fluctuation-dissipation theorem." However, in the second limit, we let $\alpha = 1$ which is equivalent to assuming that there is no dissipation. In this limit, the current velocity u_{n+1} is simply the sum of all previous velocities. The variance of the velocity increases linearly in time and the velocity values diverge. This limit corresponds to a Brownian motion that has an infinite range of correlations.

In what follows, we will simulate the Langevin equation via Eq. (32) with α varying from 0, the i.i.d. Gaussian noise case, to 1, the Brownian motion case which preserves memory of past history over all time.

IV. PROBABILISTIC DESCRIPTION AND SOLUTION FOR BROWNIAN MOTION

Brownian motion is described by Eq. (32) with $\alpha = 1$. In words, what distinguishes one value of the time series, namely u_i, from the next, namely u_{i+1}, is the addition of the i.i.d. fluctuating noise η_i. Therefore, what determines whether u_{i+1} is greater or less than u_i is solely the sign of η_i. We note, therefore, that the random variable that emerges is the difference between two successive time-series elements. Since, as Kac observed, extrema and, hence, the emergence of peak-to-peak sequences requires three random elements on average, it follows that four time-series events will, on average, produce a peak-to-peak sequence, in contrast with the three i.i.d. events which, on average, separate peak-to-peak sequences. We shall, using methods similar to those we employed earlier, derive both the mean number of events in a Brownian sequence and the distribution of sequence lengths that arise from Brownian processes.

In analogy to Sec. II, we want to calculate the probability that a time-series element u_i is a local minimum with but one element to its right. In other words, what is the probability p_1 that $u_i < u_{i+1}$ and $u_{i+1} > u_{i+2}$, i.e., that the peak is removed by only one point? From Eq. (32), it follows that this is the product of the probability that $\eta_i > 0$ with the probability that $\eta_{i+1} < 0$. Assuming that the median value of η is 0, as it would be with Gaussian noise, then each of those probabilities would be $1/2$ and

$$p_1 = \left(\tfrac{1}{2}\right)^2. \tag{34}$$

If the median value was nonzero, then u_i would on average undergo a uniform drift. In many problems, the mean and the median of a distribution are the same but not always. Remarkably, unlike our previous discussion of i.i.d. processes, we do not need to incorporate the details implicit to the distribution function associated with the η_i. Similarly, the probability that the peak on the right is removed by two points from u_i is

$$p_2 = \left(\tfrac{1}{2}\right)^3 \tag{35}$$

and, by induction, for $k = 3,4,\ldots,$

$$p_k = \left(\tfrac{1}{2}\right)^{k+1}. \tag{36}$$

It also follows that we get the same expression describing the probability that the peak on the left is removed by j points from u_i. Thus, the probability $\hat{P}(j,k)$ that the "valley" u_i has a peak j points to the left and k points to the right is

$$\hat{P}(j,k) = \left(\tfrac{1}{2}\right)^{j+k+2}. \tag{37}$$

Here, we have introduced a "$\hat{}$" to designate quantities associated with Brownian motion, unlike the i.i.d. situation where the symbols employed are not identified in this way. It follows, therefore, that the probability that Brownian motion will have a peak-to-peak sequence with $j + k$ elements will be proportional to $\hat{P}(j,k)$.

We now find the probability $\hat{f}(m)$ corresponding to all possible combinations $m = j + k$ of such configurations of peak-to-peak sequences with length $m = 4$. It follows, therefore, that $\hat{f}(m)$ can be written for $m \geqslant 2$ as

$$\hat{f}(m) = \gamma \sum_{j=1}^{m-1} \hat{P}(j,m-j) = \gamma(m-1)\left(\frac{1}{2}\right)^{m+2}, \tag{38}$$

where γ is a constant to be determined that will assure that $\sum_{m=2}^{\infty} \hat{f}(m) = 1$. After some algebra, we identify $\gamma = 1/4$ so

$$\hat{f}(m) = \frac{m-1}{2^m}, \quad \text{for} \quad m = 2,3,\ldots. \tag{39}$$

This is the normalized fraction of segment lengths of length m. This expression for Brownian motion dramatically differs from Eq. (20) for i.i.d. time series. Importantly, we observe that the decay rate with respect to m of $\hat{f}(m)$ is much slower than for $f(m)$ as a consequence of the role of memory. Moreover, for a Brownian process, it follows, after some algebra, that

$$\langle \hat{m} \rangle = \sum_{m=2}^{\infty} m\, \hat{f}(m) = 4, \tag{40}$$

as expected from the argument presented earlier. Analogously to Table I, we present in Table II the corresponding values for Brownian motion, further confirming the memory effect.

TABLE II. Probability $\hat{f}(m)$ that a Brownian sequence has a length m as obtained from Eq. (39).

m	$\hat{f}(m)$	Decimal
2	$\frac{1}{4}$	0.250 000
3	$\frac{1}{4}$	0.250 000
4	$\frac{3}{16}$	0.187 500
5	$\frac{1}{8}$	0.125 000
6	$\frac{5}{64}$	0.078 125
7	$\frac{3}{64}$	0.046 875
8	$\frac{7}{256}$	0.027 344
9	$\frac{1}{64}$	0.015 625
10	$\frac{9}{1\,024}$	0.008 789

Moreover, we have calculated and present below the generating function $\hat{g}(x)$ for Brownian motion data, namely

$$\hat{g}(x) = \frac{x^2}{(2-x)^2} = \frac{1}{4}x^2 + \frac{1}{4}x^3 + \frac{3}{16}x^4 + \frac{1}{8}x^5 + \ldots, \tag{41}$$

which, as before, varies smoothly from 0 to 1 as x goes from 0 to 1.

In the theory, m designates the number of events in a peak-to-peak sequence or in a cluster. In realizations of the process, $\bar{m}$ describes the sample average of m while $\langle m \rangle$ describes its (theoretical) formal average or population mean. We designate by the letter f the distribution of of sequence lengths or cluster sizes m, which represents the fraction that are m long. We designate by the letter g the generating function for the distribution functions f. For the i.i.d. theory, we use $f(m)$ and $g(x)$ and, for the Brownian theory, we use $\hat{f}(m)$ and $\hat{g}(x)$. Finally, in the analysis of simulation or observational data, we

utilize the sample average of the fraction, namely $\bar{f}(m)$. We employ this notation through the rest of this paper.

V. GAUSSIAN WHITE NOISE

As our first example, we will consider the maximum value statistics of a Gaussian white noise. The values in this time series are selected randomly from a Gaussian distribution of values. The values are uncorrelated in time. A maximum value is defined as a value larger than the two neighboring values. We determine the sequence of lengths m between maximum values for 4096 data points in each of 10 time series.

In Fig. 3, we give the fraction $\bar{f}(m)$ of the sequences that have a length m as a function of m. In addition, we present the standard error for the simulation results computing from the 10 data sets. Also included in the figure are the values given in Table I for the i.i.d. theory. Clearly, there is good agreement. For the i.i.d. theory, the mean sequence length is $\langle m \rangle = 3$ (exactly); for our analysis of the Gaussian white noise, we have $\bar{m} = 2.9916$. Since Gaussian white noises are symmetrically distributed, we would get the same result for the statistics of minimum values.

We calculate for each m the theoretical value for the number of sequences $\mathcal{F}(m) \times 4096$ that we expect, the sample mean from our 10 Monte Carlo simulation sets and the standard error estimated from those simulations. Finally, we calculate the observed discrepancy between the sample mean and i.i.d. theory in terms in standard error units, which we call the relative error. Since this is an i.i.d. random process, we do not expect that the last column will depart significantly from the range of -1 to 1. This is shown in Table III. Importantly, we performed many independent checks on our simulation results, particularly regarding convergence, in establishing the validity of our results. Moreover, we showed in Ref. [2] explicitly how convergence properties conformed with our expectations from probability theory.

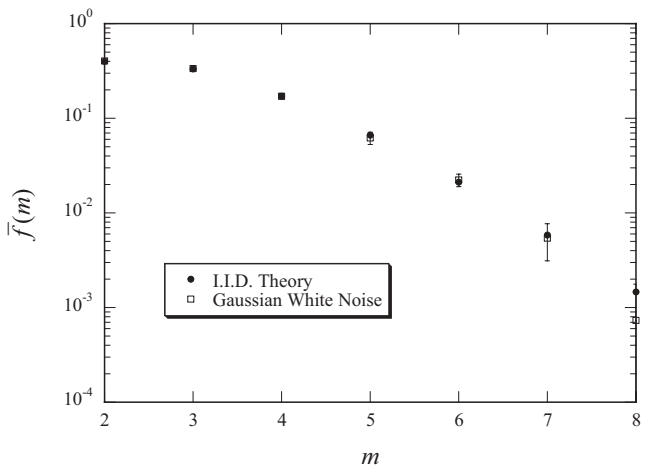

FIG. 3. Dependence of the fraction of sequences $\bar{f}(m)$ with sequence length m on m for the maximum values of a Gaussian white noise (open squares) and for the i.i.d. theory, also showing estimates of the standard error.

TABLE III. Sequence counts for the white noise time series and for the i.i.d. theory. The white noise data are the average sequence length statistics for the 10 time series, each with 4096 points. Also shown are estimates of the uncertainty established from the theory, i.e., the standard error, and the departure between the white noise data and i.i.d. theory in units of the standard error.

m	i.i.d. theory	White noise data	Standard error	Relative error
2	546	551.0	24.23	0.206
3	455	459.3	20.09	0.214
4	234	234.2	15.48	0.013
5	91	84.8	12.47	−0.497
6	29	30.5	4.58	0.328
7	8	7.4	3.13	−0.191
8	2	1.0	1.41	−0.707

VI. SEQUENCE AND CLUSTER STATISTICS OF GLOBAL SEISMICITY

We now consider the statistics of the occurrence of earthquakes on a global scale. One question we address is whether these earthquakes occur randomly in magnitude. The second question we address is their statistics of (time) clustering. Specifically, we ask whether clusters of three in time tend to dominate. The moment magnitude m_w is the standard measure of the intensity of a strong earthquake. This quantity provides a simple representation for the stress tensor's action over a distance using dimensionless units that are equivalent to the original definition by Gutenberg and Richter of earthquake magnitude. An example of the temporal clustering of large, catastrophic earthquakes are the $m_w = 9.1$ (2004) Sumatra earthquake, the $m_w = 8.3$ (2010) Chile earthquake, and the $m_w = 9.0$ (2011) Japan earthquake. Because of the many problems associated with the magnitudes of large earthquakes, we will restrict our study to the global Centroid Moment Tensor (CMT) catalog, which is available at Ref. [15] on the Internet. It is necessary to specify a minimum magnitude for which the catalog is complete. For the CMT catalog, we take this threshold moment magnitude to be $m_w = 5.5$ [16,17]. We consider the set of global earthquakes for the period 1 January 1977 to 31 December 2011. This set includes 14 014 earthquakes. The moment magnitudes of these earthquakes constitute a time series. Again, a maximum moment magnitude earthquake is defined to have a moment magnitude larger than the moment magnitudes of the two global earthquakes adjacent in time. We determine the sequence lengths of global earthquakes between these maximum magnitude earthquakes in our time series.

In Fig. 4, we give the fraction $f(m)$ of the sequences that have a length m as a function of m. Also included in the figure are the values given in Table I for the i.i.d. theory, as well as the expected standard error. The earthquake magnitude statistics are well approximated by the i.i.d. theory. The mean sequence length for the earthquakes is $\bar{m} = 3.0006$, compared with the value of $\langle m \rangle = 3$ for the i.i.d. theory.

In the case of global seismicity, we also present in Table IV our results quantifying the departure of moment magnitude statistics from an i.i.d. process. As in the case of a Gaussian

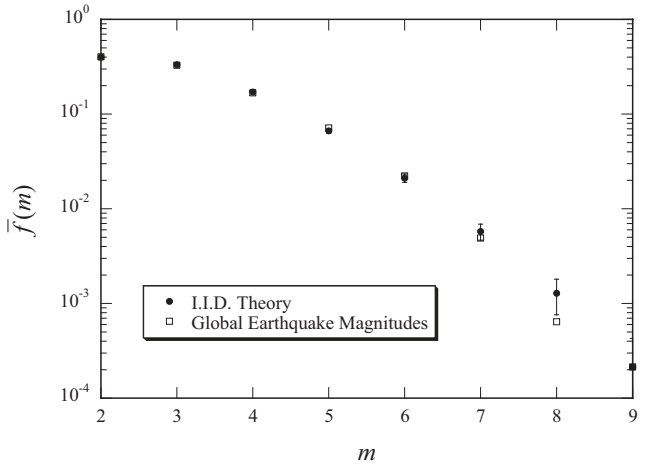

FIG. 4. Dependence of the fraction of sequences $\bar{f}(m)$ with the sequence length m on m for the maximum moment magnitudes of global earthquakes (open squares) and for the i.i.d. theory (closed circles). Standard error estimated from the i.i.d. theory.

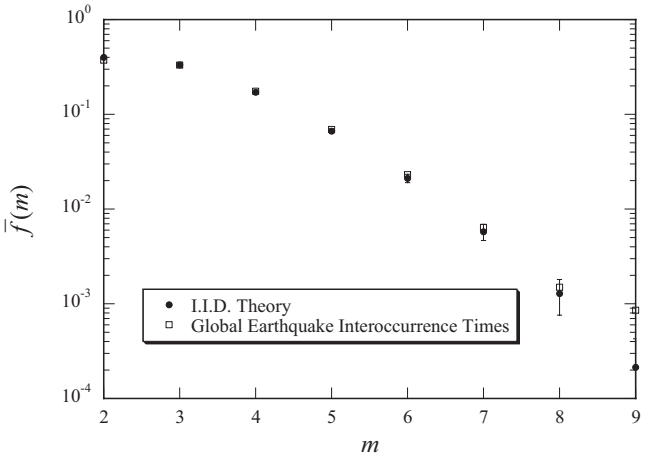

FIG. 5. Dependence of the fraction of clusters $\bar{f}(m)$ with the cluster size m on m for the interoccurrence times of global earthquakes (open boxes) and for the i.i.d. theory (solid circles). In a temporal cluster of occurrence times, each earthquake's nearest neighbor in time is a member of the cluster. Standard error estimated from the i.i.d. theory.

white noise, we observe no statistically significant departure of the CMT moment magnitude data from a random process.

We next turn to the interoccurrence time statistics of the same set of global earthquakes. Specifically, we look at the temporal clustering of events. We consider a cluster in which one of the adjacent earthquakes is the nearest in time as illustrated in Fig. 1. The interoccurrence time defining the break between two clusters is longer than the two adjacent interoccurrence times. Thus, it is a local maximum as discussed above.

In Fig. 5, we give the fraction $\bar{f}(m)$ of the clusters that have a size m as a function of m. Also included in the figure are the values given in Table I for the i.i.d. theory. As can be seen, the results in Fig. 5 are very similar to those in Fig. 4. The mean cluster size for the earthquakes is $\bar{m} = 2.9964$ compared with the value $\langle m \rangle = 3$ for the i.i.d. theory. We also tabulate the results from the interoccurrence time calculations in Table V.

There is a widely held superstition that bad events happen in sequences of three in time. Great earthquakes are certainly bad events. Our analysis shows that, on average, they do occur in sequences of three, even though they are uncorrelated in time.

Interestingly, we observe a small but statistically significant departure from i.i.d. behavior in the observed number of clusters containing only $m = 2$ points. In particular, the departure from theory in units of the estimated standard error is observed to be -2.8, which would be regarded as nearly a 3σ departure from i.i.d. behavior, while the magnitude results were not similarly effected. A possible interpretation for this observation is that there is a modest deficit due to relatively small aftershocks.

VII. LANGEVIN MODEL AND RELATION BETWEEN GAUSSIAN AND BROWNIAN NOISE

So far, we have related our random (i.i.d.) theory to results that have the appearance of being "random." Clearly,

TABLE IV. Sequence counts for global earthquake magnitude time series and the i.i.d. theory. The earthquake magnitude data are the sequence length statistics for the time series of 14 014 $m_w \geqslant 5.5$ earthquakes for the period 1 January 1977 to 31 December 2011. Also shown are estimates of the uncertainty established from the theory, i.e., the standard error, and the departure between the earthquake magnitude data and the i.i.d. theory in units of the standard error.

m	i.i.d. theory	Earthquake mag. data	Standard error	Relative error
2	1870	1882	43.24	0.277
3	1558	1539	39.47	−0.481
4	801	785	28.30	−0.565
5	312	335	17.66	1.302
6	99	104	9.95	0.503
7	27	23	5.20	−0.770
8	6	3	2.45	−1.225
9	1	1	1.00	0.000

TABLE V. Cluster counts for interoccurrence times of global earthquakes and the i.i.d. theory. The earthquake interval data are cluster length statistics for the 14 014 $m_w \geqslant 5.5$ earthquakes for the period 1 January 1977 to 31 December 2011. Also shown are estimates of the uncertainty established from the theory, i.e., the standard error, and the departure between the earthquake interoccurrence data and the i.i.d. theory in units of the standard error.

m	i.i.d. theory	Earthquake interval data	Standard error	Relative error
2	1870	1749	43.24	−2.798
3	1558	1551	39.47	−0.177
4	801	821	28.30	0.707
5	312	324	17.66	0.679
6	99	108	9.95	0.905
7	27	30	5.20	0.577
8	6	7	2.45	0.408
9	1	4	1.00	3.000

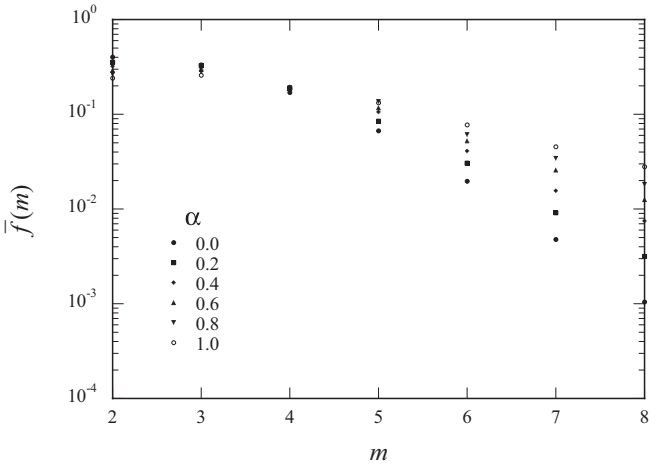

FIG. 6. Dependence of the fraction $\bar{f}(m)$ of sequences with the sequence length m on m for the maximum values of the Langevin model with $\alpha = 0.0,0.2,0.4,0.6,0.8$, and 1.0.

a Gaussian white noise time series is expected to satisfy i.i.d. statistics since adjacent values in the time series are uncorrelated. We have also shown that global seismicity both in terms of an earthquake magnitude time series and in terms of event clustering in time are well approximated by the i.i.d. theory.

We now consider an example in which there are significant deviations from i.i.d. behavior. In particular, we will consider the Langevin model given in Eq. (32). A Gaussian white noise is generated and utilized in this model. We consider time series with $\alpha = 0.0,0.2,0.4,0.8$, and 1.0. In each case, we determine the sequence lengths between maximum values for the time series that we consider. In Fig. 6, we give the fraction $\bar{f}(m)$ of the sequences that have a length m as a function of m.

We find a strong dependence on α with a systematic increase in longer sequence lengths with increasing values of α. The result is, as expected, since increasing α reduces short-range (high frequency) fluctuations. We quantify this result in Fig. 7, where we give the mean cluster size $\bar{m}$ as a function of the memory-related term α. We observe that $\bar{m}$ can be roughly

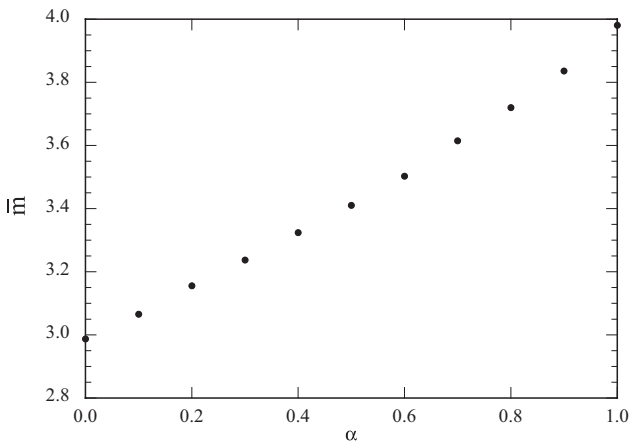

FIG. 7. Dependence of the mean cluster size $\bar{m}$ on the quantity α for the Langevin model shown in Fig. 6. We have an approximately linear increase of $\bar{m}$ from 3 to 4 as α approaches 1.

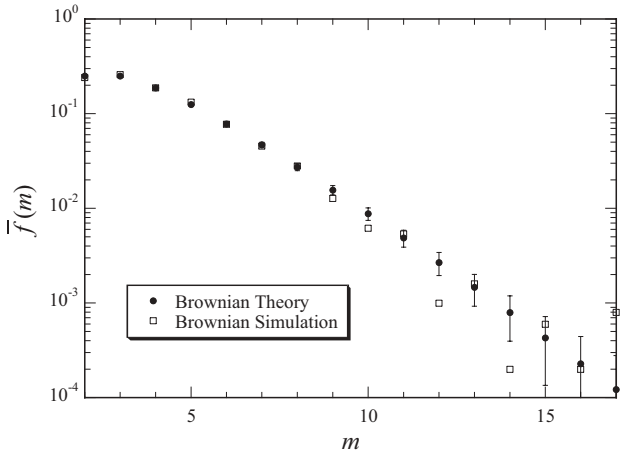

FIG. 8. Comparison between our Brownian motion theory $\hat{f}(m)$ and Brownian motion simulations $\bar{f}(m)$, including error bars, as a function of sequence length m.

approximated using linear interpolation in α between the two end-member results, i.e., an i.i.d. process ($\alpha = 0$) and Brownian motion ($\alpha = 1$). In Fig. 8, for our Brownian motion theory and Brownian motion simulations, we give the fraction $\bar{f}(m)$ of the sequences that have a length m as a function of m for $\alpha = 1$ (open squares) from Eq. (39). We also show the theoretical result $\hat{f}(m)$ for a Brownian process (solid circles) as well as an estimate for the standard error. We observe the quality of agreement between the theory and observations. Finally, in Table VI, we tabulate our comparison between our 20 000-point Brownian ($\alpha = 1$) motion simulation results and our theory.

VIII. AURORAL ELECTROJET AL INDEX

The Auroral Electrojet (AE) index was first developed by Davis and Sugiura [18] as a measure of global electrojet activity in the auroral zone and describes, in part, the Earth's magnetospheric response to solar activity. After significant computation and normalization of magnetic field data obtained

TABLE VI. Cluster counts for Brownian motion data corresponding to $\alpha = 1$ and Fig. 7.

m	Brownian theory	$\hat{f}(m)$	Brownian simulation	$\bar{f}(m)$
2	1250	0.250	1211	0.241
3	1250	0.250	1302	0.259
4	938	0.187	940	0.187
5	625	0.125	666	0.133
6	391	0.078	389	0.077
7	234	0.047	229	0.046
8	137	0.027	141	0.028
9	78	0.016	64	0.013
10	44	0.009	31	0.006
11	24	0.005	27	0.005
12	13	0.003	5	0.001
13	7	0.001	8	0.002
14	4	0.001	1	0.000
15	2	0.000	3	0.001

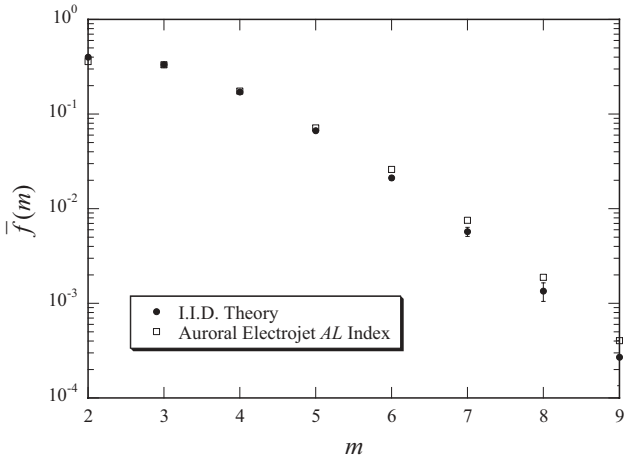

FIG. 9. Dependence of the fraction of clusters $\bar{f}(m)$ with cluster size m on m for interoccurrence times of the auroral electrojet AL index (open boxes) and for the i.i.d. theory (solid circles).

at a number of stations, the largest (upper) and smallest (lower) values are selected and are designated as the AU and AL indices, respectively. The AL index presents the lower envelope for geomagnetic activity as a function of universal time (UT). It is of particular interest, as it provides a quantitative measure of the significance of magnetospheric substorms as a mode of response of the magnetosphere to the solar wind. Thus, the AL index can be regarded as the response of a complex electromagnetic charging system (Earth's magnetospheric environment) to a stochastic energy source (the Sun and space weather). The data we present [19] cover the last two solar cycles beginning 1 January 1984 and ending 31 May 2011, and contains 44 510 events. In particular, the data correspond to the time intervals between successive substorm onsets measured to the nearest second over a time frame spanning more than 27 years. Importantly, modeling efforts employed to understand the AL index have focused on the use of high-order autoregressive methods; our model Eq. (32) is a first-order autoregressive method.

We consider a cluster in which one of the adjacent storms is nearest in time, as illustrated in Fig. 1. In Fig. 9 we give the fraction $f(m)$ of the clusters that have a size m as a function of m. Also included in the figure are the values given in Table I for the i.i.d. theory. There is a small but significant deviation between i.i.d. theory and the data. This is also illustrated in Table VII. The mean cluster size of the data is $\bar{m} = 3.0822$ compared with $\langle m \rangle = 3$ for the i.i.d. theory. This is evidence of a small degree of persistence.

TABLE VII. Cluster counts for the AL index data and the i.i.d. theory. Also shown are estimates of the uncertainty established from the theory and the departure between the AL data and the i.i.d. theory in units of the standard error.

m	i.i.d. theory	AL data	Standard error	Relative error
2	5935	5327	77.039	−7.892
3	4945	4916	70.321	−0.412
4	2543	2602	50.428	1.170
5	989	1060	31.448	2.258
6	314	387	17.720	4.120
7	85	112	9.220	2.929
8	20	28	4.472	1.789
9	4	6	2.000	1.000
10	1	1	1.000	0.000

IX. OLD FAITHFUL GEYSER ERUPTIONS

The nature of geysers has been of great interest over recorded history. The most famous of these, Old Faithful, in Yellowstone National Park was named by the Washburn expedition of 1870, who were impressed by its size and frequency. Evidently, this cone-type geyser has been erupting in nearly the same fashion throughout the recorded history of Yellowstone. Through the years, it has become one of the most studied geysers in the park. Rinehart [20] appears to have been the first to have developed a physical, albeit nonquantitative, model for its activity and its bimodal character: the longer the recurrence time between eruptions, the longer the duration of the eruption, establishing a conceptual link between the time to transport and heat water, and the extent of the eruption.

National Park Service (NPS) geologist Ralph Hutchinson, among others, collected substantial amounts of data which were then analyzed statistically by Azzalini and Bowman [21], who validated the bimodality in a quantitative way. Intuitively, this can be seen as an outcome of a uniform rate of water supply to the geyser system. As a result, eruption predictions can be made using a regression formula based on the duration of an eruption. NPS rangers predict eruption times within ± 10 minutes 90% of the time, but it is not possible to predict more than one eruption in advance. Dowden *et al.* [22] developed a hydrodynamic model for eruptions that yielded an estimate for geyser power output. Hutchinson *et al.* [23] reported on pressure and temperature as a function of time measured from 1983 to 1993 using probes at 22 m depth and employed a video camera to characterize the conduit geometry. Hutchinson died during an avalanche just before the submission of the manuscript and the National Park Service no longer employs a geologist to monitor Old Faithful's activity.

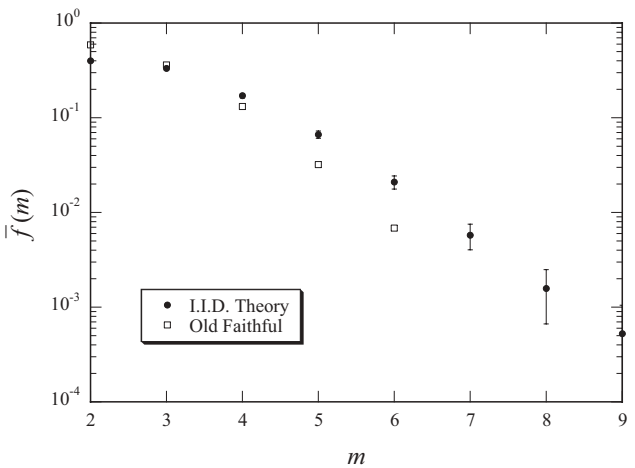

FIG. 10. Dependence of the fraction of clusters $\bar{f}(m)$ with cluster size m on m for the recurrence times of eruption of Old Faithful geyser (open boxes) and for the i.i.d. theory (solid circles).

TABLE VIII. Cluster counts for Old Faithful geyser data for the recurrence time for the 5692 events in 2010 and the i.i.d. theory. Also shown are estimates of the uncertainty established from the theory and the departure between the geyser data and the i.i.d. theory in units of the standard error.

m	i.i.d. theory	Geyser data	Standard error	Relative error
2	760	1119	27.57	13.02
3	633	689	25.16	2.23
4	326	251	18.06	−4.15
5	127	61	11.27	−5.86
6	40	13	6.32	−4.27
7	11	0	3.32	−3.32
8	3	0	1.73	−1.73
9	1	0	1.00	−1.00

The Geyser Observation and Study Association (GOSA), an incorporated nonprofit scientific and educational corporation, was founded some years before Hutchinson's untimely death and has monitored Old Faithful geyser electronically using data loggers since 2000. Since then, its recurrence intervals have varied from 44 to 125 min with an average of about 90–92 min. This device measures the runoff water temperature at a point about 20 m from the vent toward the west. The sensor picks up preplay and the eruption start. The data were collected by Ralph Taylor, a park volunteer working under an NPS research permit, and by personnel working for the Geology Department of the Yellowstone Center for Resources. Taylor assembled the eruption times and NPS the temperature data. The digitized data and other information can be obtained online at Ref. [24] we employed the data set for our analysis.

We consider a cluster in which one of the adjacent eruptions is the nearest in time as illustrated in Fig. 1. In Fig. 10, we give the fraction $\bar{f}(m)$ of the clusters that have a size m as a function of m. Also included in the figure are the values given in Table I for the i.i.d. theory. There is clearly a strong deviation from i.i.d. behavior. The mean cluster size is $\bar{m} = 2.6685$ compared with the value $\langle m \rangle = 3$ for the i.i.d. theory. We take this low value of $\bar{m}$ to be evidence of antipersistence in the behavior of the geyser time series. Extreme antipersistence would be a sequence of long-short-long-short. . .. In this case, we would have $\bar{m} = 2$. The observation that the mean cluster size is shorter than the value $\langle m \rangle = 3$ found for an i.i.d. process is direct evidence of antipersistence in the time series. The degree to which a departure from i.i.d. behavior is present in the data is clear from Table VIII.

X. STANDARD & POOR'S 500 DATA FROM JANUARY 1928 TO DECEMBER 2011

Many extremely long financial time series are available. As a typical example, we will consider the daily closing prices of the Standard and Poor's 500 stock index for the period 1928 to 2011. Following standard practice [25], we utilize the daily returns R_i defined by

$$R_i = \frac{Y_i - Y_{i-1}}{Y_{i-1}}, \tag{42}$$

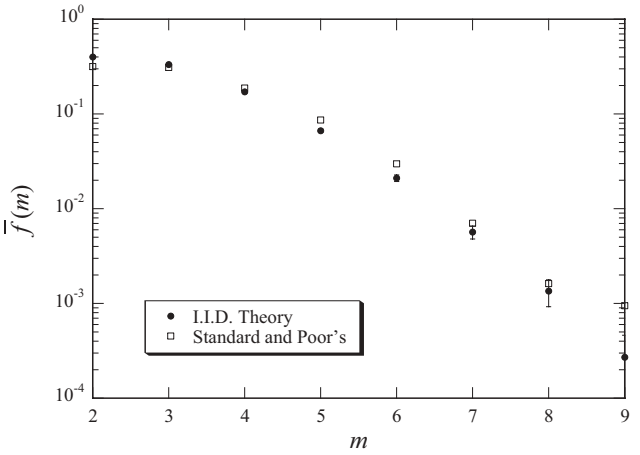

FIG. 11. Dependence of the fraction of sequences $\bar{f}(m)$ with sequence length m on m for the time series of daily returns of the Standard and Poor's 500 stock index (open squares) and the i.i.d. theory (solid circles).

where Y_i is the closing price on day i and Y_{i-1} is the closing price on the previous day; R_i is the fractional daily gain or loss. The daily values of R_i constitute a time series. Again, a maximum value is defined to be an R_i that is larger than both R_{i-1} and R_{i+1}. We determine the sequence lengths of days between these maximum values of R_i in our time series. The data were obtained from two sources. For the period 1928–1993, we employed Ley [26], while the 1994–2011 data was from Ref. [27].

In Fig. 11, we give the fraction $\bar{f}(m)$ of sequence lengths m as a function of m. Also included in the figure are the values for the i.i.d. theory from Table I as well as the expected standard error. The mean sequence length for the financial data is $\bar{m} = 3.1858$ compared with the value $\langle m \rangle = 3$ for the i.i.d. theory. We also present in Table IX our results quantifying the departure of the financial from an i.i.d. process. From a comparison of the financial value of $\bar{m} = 3.1858$ with results for the Langevin model given in Fig. 7, we see a correspondence with $\alpha \approx 0.25$, an indication of mild persistence.

TABLE IX. Sequence counts for the 22 147 Standard and Poor's 500 daily returns for the period 1928–2011 and the i.i.d. theory. Also shown are estimates of the uncertainty established from the i.i.d. theory, i.e., the standard error, and the departure between the Standard and Poor's data and the i.i.d. theory in units of the standard error.

m	i.i.d. theory	S&P data	Standard error	Relative error
2	2956	2344	54.37	−11.26
3	2464	2293	49.64	−3.44
4	1267	1388	35.60	3.40
5	493	639	22.20	6.58
6	156	220	12.49	5.12
7	42	52	6.48	1.54
8	10	12	3.16	0.63
9	2	7	1.41	3.54

XI. DISCUSSION

Our studies of events in time, depicted as points on a time line, considered the cluster statistics of nearest-neighbor events in time. Our analysis was for random i.i.d. events which, by definition, are uncorrelated in time. We focused on developing a theory for the presence of such clusters, which we tested against a variety of standard distribution functions. An analytic expression for the distribution of cluster sizes $f(m)$ was given in Eq. (20) and numerically listed in Table I. The mean cluster size was found to be $\langle m \rangle = 3$.

We went on to explore time series of quantities emerging from biological and physical processes. We noted the observation made by Kac [3] that, when plotted, the average peak-to-peak sequence interval would be three events. We believed that there was a connection between the statistics of clusters and that of peak-to-peak sequences. We found that connection when we studied sequences of interoccurrence times as though they were time series. This equivalence is illustrated in Fig. 1. The interoccurrence time between two neighboring clusters is, by definition, larger than the two neighboring time intervals. In terms of the equivalent time series, this interval is a local maximum in that it is larger than the two adjacent values. As can be seen in Fig. 1, the distribution of sequences is identical to that for the equivalent time series. Thus, Eq. (20) is applicable to the mean sequence length between local maxima, which is $\langle m \rangle = 3$, a result previously demonstrated by Newman [2].

The equivalence discussed above allows us to apply sequence statistics for local maxima in any time series. Our analysis is valid for our i.i.d. time series (i.e., without correlations) for any statistical distribution. To exhibit the validity of our analysis, we have determined the sequence statistics for a Gaussian white noise. The excellent agreement between the simulations and the analytic results are illustrated in Fig. 3 and Table III. The mean cluster length for the simulations is $\bar{m} = 2.9916$ compared with the theoretical value $\langle m \rangle = 3$.

We have also studied the sequence statistics for sequence lengths between local maxima in Brownian motion. Once again, an analytical expression was obtained for the statistical distribution of the sequence lengths, and this is given in Eq. (39). The mean sequence length was found to be $\langle m \rangle = 4$. We have generated Brownian motion as the running sum of Gaussian white noises and obtained the sequence statistics for the local maximum values of our Brownian motion. The excellent agreement between our simulation and the analytic results is illustrated in Fig. 8 and Table V. The mean cluster length is $\bar{m} = 3.9809$ compared with the theoretical value $\langle m \rangle = 4$.

In order to study the dependencies of sequences statistics on correlations, we utilize the recursive form of the Langevin equation, which is equivalent to a first-order AR process, given in Eq. (32). The mean cluster size $\bar{m}$ is given as a function of the Langevin parameter α in Fig. 7. The transition from a white noise [$\alpha = 0$, $\langle m \rangle = 3$] to a Brownian motion [$\alpha = 1$, $\langle m \rangle = 4$] is clearly illustrated.

In order to illustrate the application of sequence and cluster statistics, we consider several sets of observations. We, first, have compared our results with the global occurrence of large earthquakes. We consider both the time series of magnitudes and the time intervals between occurrences. For the sequence statistics of peak earthquake magnitudes but slightly less so for the nearest-neighbor cluster statistics in time, we find excellent agreement with the i.i.d. theory. We speculate that the small departure observed in the interoccurrence time statistics is due to aftershocks. Nevertheless, the mean sequence size for large earthquakes is $\bar{m} = 3.006$, and the mean cluster size for the temporal occurrence of large earthquakes is $\bar{m} = 2.996$. We suggest that this exemplifies the adage that "bad things come in threes."

We next considered the auroral electrojet (AL) index. The data corresponds to the time intervals between successive substorm onsets. The cluster statistics of nearest-neighbor events are given in Fig. 9 and Table VII. The mean cluster length of the data is $\bar{m} = 3.0822$ compared with $\langle m \rangle = 3$ for the theory. This is evidence for a small degree of persistence.

We also studied the time intervals between the eruptions of Old Faithful geyser in Yellowstone National Park. The cluster statistics of nearest-neighbor eruptions in time are given in Fig. 10 and Table VIII. The mean cluster length of the data is $\bar{m} = 2.6685$ compared with the i.i.d. value $\langle m \rangle = 3$. This is indicative of antipersistence. There is a statistically significant tendency for a long interval to be followed by a short interval and a short interval to be followed by a long interval, more so than strictly "random" behavior would dictate.

The final example that we considered was related to the time series of daily closing prices of the Standard and Poor's 500 stock indexes. In order to eliminate the influence of long-term trends, we considered the daily returns defined in Eq. (42). The prices can be approximated as Brownian motion so the differences (returns) can be approximated as a white noise. The comparison of the sequence statistics of the returns with the i.i.d. (white noise) theory given in Fig. 11 and Table IX indicate a small positive correlation. For the data, $\bar{m} = 3.1455$ compared with the i.i.d. value $\langle m \rangle = 3$.

In this paper, we formulated the problem of describing events in time through their interoccurrence times and the emergence of clusters based on nearest-neighbor associations. Further, we formulated the problem of describing time series of events characterized by measured amplitudes which, when plotted, produced sequences of peak-to-peak events. We showed that both problems were mathematically identical insofar as their description of the number of events per cluster and per sequence. Further, we calculated in closed form the associated distribution function for the number of events in a cluster or in a sequence, demonstrating that the mean value was 3. We also consider the Brownian motion problem corresponding to a running sum of time series value, obtained its associated distribution function, and demonstrated that its mean value was 4. Finally, we applied these analytic results in the exploration to a variety of problems. Since biological time series for feral animal populations tend to be relatively small and subject to great uncertainty, we focused on physical and economic time series, exploring problems ranging from measures of natural hazards to financial indices. We found, in many cases, remarkable agreement with the theory we

formulated based on i.i.d. processes and, when there were significant departures from the theory, we could apply the added insight to a deeper understanding of the underlying problem.

ACKNOWLEDGMENTS

We thank David Campbell, Robert McPherron, Margaret Kivelson, Joe Rudnick, Philip Sharp, and John Rundle for useful discussions.

[1] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. (John Wiley & Sons, New York, 1968), Vol. I.
[2] W. Newman, Chaos (submitted).
[3] L. Cole, J. Wildlife Manage. **15**, 233 (1951).
[4] http://www3.imperial.ac.uk/cpb.
[5] M. Gilpin, Am. Nat. **107**, 727 (1973).
[6] F. Ayala, M. Gilpin, and J. Ehrenfeld, Theor. Population Biol. **4**, 331 (1973).
[7] S. Chandrasekhar, Rev. Mod. Phys. **15**, 1 (1943).
[8] C. Kittel, *Elementary Statistical Physics* (Wiley, New York, 1958).
[9] M. Tata, Z. Wahrscheinlich Leit. **12**, 9 (1969).
[10] N. Glick, Am. Math. Mon. **85**, 2 (1978).
[11] V. Nezvorov, Theory Probab. Appl. **32**, 201 (1986).
[12] B. Arnold, N. Balakrishnan, and H. Nagaraja, *Records* (John Wiley & Sons, New York, 1998).
[13] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th ed. (Wiley, Hoboken, NJ, 2008).
[14] D. L. Turcotte, *Fractals and Chaos in Geology and Geophysics*, 2nd ed. (Cambridge University Press, Cambridge, UK, 1997).
[15] http://www.globalcmt.org/.
[16] A. Dziewonski, T. Chou, and J. Woodhouse, J. Geophys. Res. **86**, 2825 (1981).
[17] G. Ekstrom, A. Dziewonski, N. Matgernovskaya, and M. Nettles, Phys. Earth Planet. Int. **148**, 327 (2005).
[18] T. Davis and M. Sugiura, J. Geophys. Res. **71**, 785 (1966).
[19] T.-S. Hsu and R. L. McPherron, Adv. Space Res. (in press, 2012), doi: 10.1016/j.asr.2012.06.034.
[20] J. Rinehart, J. Geophys. Res. **74**, 566 (1969).
[21] A. Azzalini and A. Bowman, Appl. Statist. **39**, 357 (1990).
[22] J. Dowden, P. Kapadia, G. Brown, and H. Rymer, J. Geophys. Res. **96**, 18059 (1991).
[23] R. Hutchinson, J. Westphal, and S. Kieffer, Geology **25**, 875 (1997).
[24] http://www.geyserstudy.org/.
[25] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, UK, 2000).
[26] E. Ley, Am. Statistician **50**, 311 (1996).
[27] http://finance.yahoo.com.