

Predicting extreme avalanches in self-organized critical sandpiles

Anja Garber,¹ Sarah Hallerberg,² and Holger Kantz¹

¹Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Str. 38, D 01187 Dresden, Germany

²Instituto de Física de Cantabria (IFCA), CSIC–Universidad de Cantabria, E 39005 Santander, Spain

(Received 12 December 2008; revised manuscript received 18 May 2009; published 25 August 2009)

In a finite-size Abelian sandpile model, extreme avalanches are repelling each other. Taking a time series of the avalanche size and using a decision variable derived from that, we predict the occurrence of a particularly large avalanche in the next time step. The larger the magnitude of these target avalanches, the better is their predictability. The predictability which is based on a finite-size effect, is discussed as a function of the system size.

DOI: [10.1103/PhysRevE.80.026124](https://doi.org/10.1103/PhysRevE.80.026124)

PACS number(s): 05.65.+b, 05.45.Tp, 45.70.-n, 89.75.-k

Extreme events have received considerable interest over the past years [1,2] for two reasons: as natural events, their extremity is linked to large impact in particular on human society due to their mostly destructive character. From the more fundamental point of view, extreme events are large deviations due to intrinsic fluctuations of complex systems. Therefore, in the framework of complex dynamics, one would like to understand and classify possible underlying mechanisms for huge deviations and then derive statistical properties about recurrences and magnitudes. One prominent class of complex dynamics which is capable of generating arbitrarily large events without correspondingly strong external perturbations is SOC systems. Self-organized criticality (SOC) was introduced by Bak, Tang, and Wiesenfeld [3] as a dynamical mechanism to explain the abundance of power laws in nature. An open out of equilibrium system is driving itself into a critical state without fine tuning of some control parameter, by a balance of loading or stress accumulation and discharging or stress release. In the critical state, the magnitude distribution of stress release in a single event (to be properly defined) follows a power law, i.e., is scale free. Relevant characteristics for such events are not only the frequency-size distribution including a potential maximal event size, but also temporal and spatial correlations. Such correlations are the source of information for the attempt to forecast particular events, as it is routine for weather (e.g., hurricanes), and as it is urgently desired for earthquakes [4]. It has been suggested that seismicity might indeed stem from SOC-type dynamics [5,6]. In fact, the attempt to predict earthquakes, being events of strong seismic activity of the earth's crust, on the basis of past observed seismic activity [7] has inspired our work. Moreover, precursors known from earthquake prediction have already previously been applied to avalanches in sandpile models, which constitute the original SOC systems [8,9]. Avalanche predictions have also been studied in granular matter experiments [10,11]. In this paper, we show that the additional knowledge of internal system parameters such as sandpile height, clusterization or local activity previously used for avalanche prediction is not necessary. In fact, due to temporal finite-size effects in the system, predictions of similar quality can be issued based on the avalanche time series data alone, improving practical applicability.

The Abelian sandpile model [3,12] used in this paper mimics a sandpile, where sand grains randomly drop onto a

heap of sand and cause avalanches whenever the local height becomes too large. It is commonly accepted that in the thermodynamic limit, this system is critical. On a square lattice of size $L \times L$, random integer variables z_{ij} are defined which can, in the stable state, attain values $z_{ij} \in \{0, 1, 2, 3\}$. At each time step, a sand grain is added by randomly choosing a pair (i, j) and incrementing z_{ij} by 1. If any z_{ij} exceeds 3, then sand grains are redistributed,

$$z_{ij} \rightarrow z_{ij} - 4 \quad z_{i'j'} \rightarrow z_{i'j'} + 1 \quad (1)$$

for $(i', j') = (i, j \pm 1)$ and $(i', j') = (i \pm 1, j)$.

This rule is applied repetitively until all $z_{ij} \leq 3$. During such a sequence of updates which is called avalanche no new grains are added, so the avalanche occupies no extra time step. The avalanche size s_i is measured as the number of topplings, i.e., as the total number of involved system sites with their multiplicity.¹ In the thermodynamic limit $L \rightarrow \infty$ the magnitude-frequency distribution $p(s)$ of an avalanche of size s follows a power law $p(s) \propto s^{-\alpha}$ with α between -1.05 and -1.293 depending on the system size used in the simulations [13–16]. This does not only imply that for $L \rightarrow \infty$ there is nothing like a “largest possible avalanche,” but that also the mean avalanche size is infinite. In the finite-sized system, the largest possible avalanche has the size $s_{\max} = 1/6 \cdot L(L+1)(L+2)$, and the power law has a cutoff already at sizes $L^2/2$ [17]. Hence, there is a finite largest avalanche and a finite mean.

More importantly, the temporal succession of avalanches is also affected by the finiteness of the system. In this paper we want to study predictability of extreme events based on this fact. To do so, we generate a time series of avalanche sizes $\{s_i\}$, $i=1, \dots, N$, called the *observation series*, from Eq. (1) for $N=2 \times 10^8$ time steps (i.e., N added sand grains), after discarding a transient during which the system acquires criticality. In the thermodynamic limit, successive avalanches are independent of each other and the time series of avalanche sizes s_i is uncorrelated. Consequently, the recurrence time distribution, i.e., the distribution of time intervals in between two successive avalanches, is an exponential, as mentioned in passing in several previous publications [18–20]. It can

¹Other definitions of avalanche size are in use, e.g., the number of sand grains dissipated over the border [10].

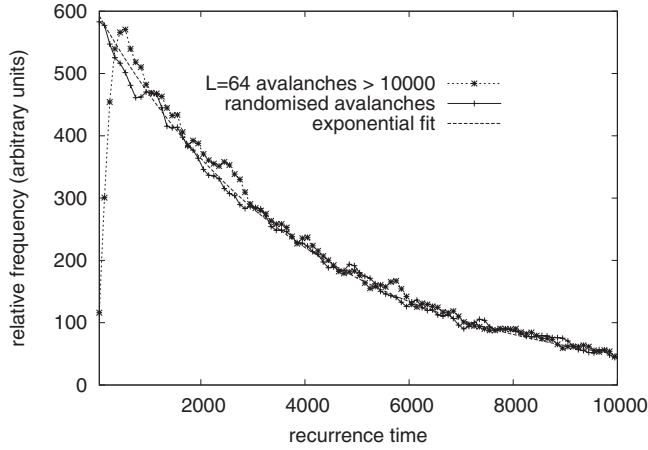


FIG. 1. Recurrence time distribution for $L=64$ and $\eta=10\,000$ for the observation series and the surrogate data. The distribution was truncated at a time of 10 000 added sand grains to enhance visibility of the depression for small times.

easily be verified by numerical simulations where the temporal properties of the thermodynamic limit are reasonably well approximated by the behavior of the small avalanches that form the majority of events and are not influenced by the system boundary. When observing the recurrence time distribution of very large avalanches in finite systems, one observes a suppression of short return times when compared to the exponential distribution of recurrences of small events. A recurrence time distribution for large avalanches is shown in Fig. 1. The figure also shows that with surrogate data obtained through a random reshuffling of the observation series $\{s_i\}$, the recurrence time distribution remains exponential as expected for temporally independent events even if only large avalanches are considered. The observed suppression therefore stems from temporal correlations between large events. In a previous publication [17] we analyzed finite-size effects in more detail and showed that this is indeed one of them: an extremely large avalanche can cause a relaxation of the system away from the critical state so that many added sand grains are needed until the system has rebuilt itself and can once again create avalanches of a similar or larger size. In this paper, we show that this rather weak repulsion of extreme avalanches in finite SOC systems can be used for their prediction, using a suitable decision variable.

For a realistic setting, we base our predictions solely on the knowledge of the time series of avalanche sizes, i.e., the observation series which was described above. From it we construct an *event series* $\{X_i\}$, $i=1, \dots, N$ such that $X_i=1$ if $s_i > \eta$ and $X_i=0$ else, i.e., a Boolean series that considers all those avalanches as events which exceed a predefined magnitude η . Only using information from the past, we want to forecast whether the next value of the event variable is 1, i.e., whether the next avalanche exceeds the magnitude η . To do so, at each time step i we construct a *decision variable*

$$y_i = \sum_{k=1}^i a^k s_{i-k}, \quad 0 < a < 1, \quad (2)$$

which is equivalent to

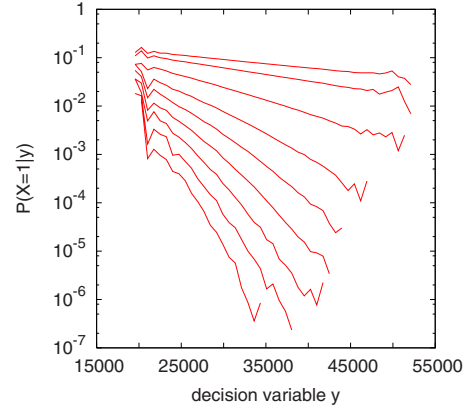


FIG. 2. (Color online) Conditional probability $P(X=1|y)$ for the system size $L=64$ as a function of the decision variable y , for events exceeding 1%, 2%, 5%, 10%, 15%, 20%, 25%, 30%, and 35% of the maximal avalanche size s_{\max} (from top to bottom).

$$y_{i+1} = ay_i + as_i \quad (3)$$

and thus to a linear filter also known as AR(1) process. y_i is independent of s_k for $k \geq i$, hence when predicting the value of X_i on the basis of the value of y_i , causality is respected. The quality of the predictions based on y_i depends on the time scale expressed by the parameter a in Eq. (2). A suitable choice, independently of η , is $a = \exp\{-1/T(L)\}$ with $T(L=64)=225$, $T(L=128)=450$, $T(L=256)=900$.

A necessary condition for a decision variable to support predictability is that the conditional probability $P(X=1|y)$, i.e., the probability for an event to happen given a specific value of the decision variable y , has at least one significant maximum in y . A numerical estimate of $P(X=1|y)$ obtained by binning and counting of pairs (X_i, y_i) from our simulated time series data is shown in Fig. 2 on a semilogarithmic scale for different values of the minimal event magnitude η and a system size $L=64$. Since the average $\langle P(X=1|y) \rangle_y$ is the event rate, these probabilities are smaller the larger η is. The conditional probabilities can be determined only on the observed range of y (which appears shortened for larger thresholds η in Fig. 2 due to finite sampling and omitting of data for those values of y where too few avalanches occurred in total), but are not subject to any boundary condition such as being zero outside this range. The numerical results are consistent with an exponential distribution $P(X=1|y) \propto \exp[-b(\eta)y]$, where $b(\eta) \approx c\eta/s_{\max}$ is an η -dependent decay rate where the numerical constant c depends on L ($c \approx 0.0002$ for $L=64$). Therefore, all curves have a trivial maximum at the smallest observed y value, i.e., when only small avalanches have occurred during the past period considered in the decision variable. This reflects the temporal repulsion of extreme events underlying our predictions. As expected, the smaller the minimal event magnitude η , the less pronounced the maximum is, since the weak correlations are a finite-size effect which diminishes if smaller events are considered. For even smaller η , this conditional probability approaches a constant, which is also obtained for an analysis of the surrogate data and which implies that the sequence of avalanche sizes has no predictive power, and predictions cannot be

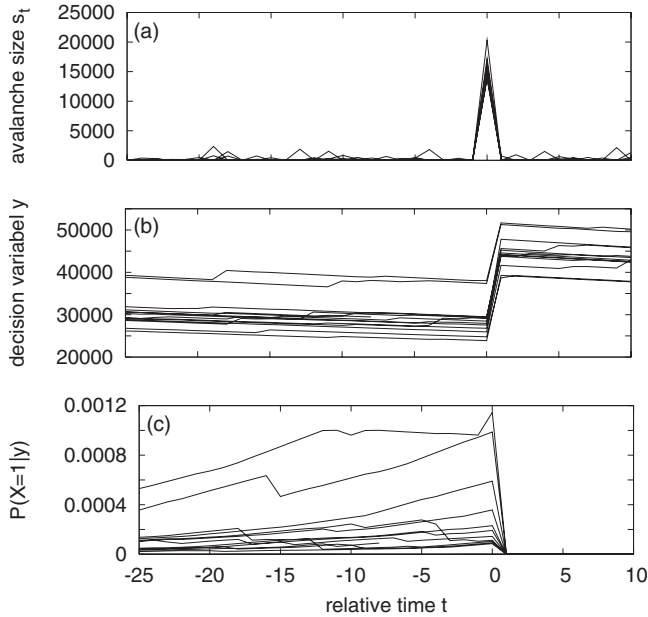


FIG. 3. Illustration of probabilistic predictions (superposition of ten examples): at relative time 0, an extreme event takes place (top panel: avalanche series s_t). The decision variable before the event drops to small values (medium panel), whereas the predicted probability for the event to happen is high (bottom panel).

made better than chance. This will be verified by an analysis of the prediction quality below.

The actual prediction is made by applying a threshold to predicted probabilities of an event to follow. A probabilistic predictor based on y_i is a map $y_i \mapsto \hat{p}_i$, $\hat{p}_i \in [0, 1]$. As a consequence of the Neyman-Pearson Lemma [21] and as discussed in [22], the optimal such predictor is $\hat{p}_i = P(X=1|y_i)$, i.e., using the conditional probability of event and decision variable of the stochastic process under study. A simple threshold converts this probabilistic prediction into a deterministic prediction $y_i \mapsto \hat{X}_i$ via

$$\hat{X}_i = \begin{cases} 1: & \text{if } P(X=1|y_i) > p_c, \\ 0: & \text{else,} \end{cases} \quad (4)$$

where $\hat{X}_i \in \{0, 1\}$ is the prediction of the event variable. The value of the parameter p_c is freely adjustable to determine the total alarm rate. Comparing \hat{X}_i and X_i allows one to verify or falsify one's prediction. This map optimizes the prediction for a given decision variable y with respect to the receiver operating characteristics (ROC) analysis which will be discussed below. Since for the SOC model chosen in this paper the probability distribution $P(X=1|y)$ is *a priori* unknown, we estimate it by construction of a histogram from the first half of the avalanche time series data (training set) (see Fig. 2). The decision variable y_i is thereafter calculated at each time step of the second half of the avalanche time series (test set—predictions are issued only after the first 10 000 time steps of the test set to ensure that no transient effects distort the prediction). This keeps the predictions strictly out of sample.

In Fig. 3 we show an example of how the decision vari-

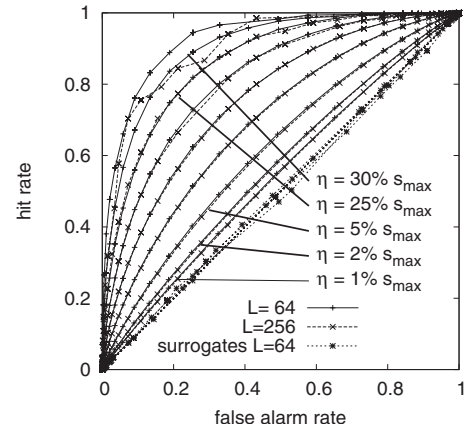


FIG. 4. ROC plots for the prediction of extreme avalanches for different minimal event magnitudes η , comparing $L=64$ and $L=256$ and the surrogate data for $L=64$.

able y and the predicted probability for an event to happen in the next time step, $\hat{p} = P(X=1|y)$, evolve in the last 25 time steps before and immediately after ten exemplarily chosen large events. As can be seen, the decision variable y overall decreases as the contribution of the last large event is multiplied with an increasingly small prefactor a^k as it lies farther and farther in the past. Once the event has taken place, the decision variable y then jumps to much larger values through the addition of a large avalanche size and thereby suppresses the prediction of large probabilities immediately after the event has taken place. This is reflected in the predicted probability of an extreme event to happen, which jumps to zero once the event has taken place. Before the event, it overall increases and can be quite large for some time, so that for suitable values of p_c a sequence of alarms $\hat{X}_i=1$ is given. As can also be seen from Fig. 3, the conditional probability $\hat{p} = P(X=1|y)$ is subject to significant fluctuations giving rise to prediction errors, which shall be studied in the following.

Since extreme events are rare, the scoring of the prediction skill should not depend explicitly on the rate of events. For binary predictions $\hat{X}_i \in \{0, 1\}$ the prediction of events becomes a classification task, with two types of errors: missing an event, and giving a false alarm. We therefore choose the receiver operating characteristics (ROC) [23] as a method to analyze prediction quality. This is a plot of the hit rate versus the false alarm rate, as a function of the total rate of alarms, which here is tuned by the threshold p_c . For $p_c \rightarrow \max_y P(X=1|y)$, the condition for $\hat{X}_i=1$ is almost never satisfied, so that both hit rate and false alarm rate are zero. For very small values of p_c , both rates tend to unity. Only if in between the hit rate exceeds the false alarm rate, the predictor is useful. A random prediction of \hat{X}_i would generate equal hit and alarm rate, i.e., the diagonal in the ROC plot is the benchmark.

In Fig. 4 we report ROC curves for predictions using Eq. (4) on avalanche sequences for system sizes $L=64$ and $L=256$ as well as surrogate data for $L=64$ and events exceeding different magnitudes η . The ROC plot shows two striking results: The larger the magnitude of the target events, limited by η from below, the better is the predictability. The

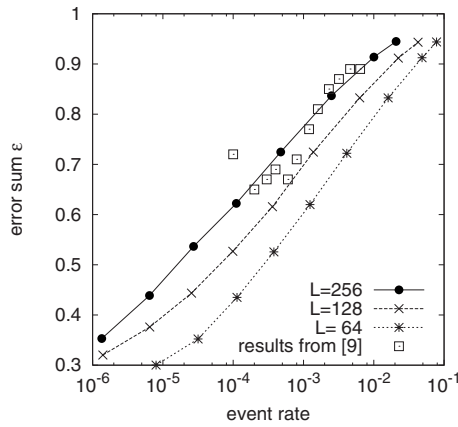


FIG. 5. The error sums ϵ for our predictions in a representation which permits the comparison to the results in [9]. The symbols on every line represent the nine different η values of Fig. 1.

predictability for given η does not depend on the system size L , if η is defined as a fraction of the maximal avalanche size $s_{\max}(L)$, showing that predictability is indeed based on a finite-size effect since the relative frequency of the events is not the influencing factor. Choosing $\eta \propto s_{\max}$, each two curves for $L=64$ and $L=256$ are indeed almost identical except for statistical fluctuations for large η . In contrast, as expected, the surrogate data show no predictability regardless of the chosen threshold η , always lying along the diagonal in the ROC plot, except for statistical fluctuations.

Different prediction algorithms for extreme avalanches in the BTW Abelian sandpile model were already implemented previously. The most notable analysis of the predictability on a sandpile was carried out by Shapoval and Shnirman [9]. Instead of the time series of avalanche sizes $\{s_i\}$, in [9] two different internal variables of the sandpile (average height, size of critical clusters) are used as inputs (decision variables). Anomalous values of these variables, so-called precursors, indicate that an avalanche is likely to follow. If these monitored variables are found to be close to the previously identified precursor, an alarm is given. Shapoval and Shnirman evaluate these predictions using the normalized error sum ϵ which is defined as the sum of the rate of unpredicted avalanches and the total alarm rate. For predictions that are no worse than chance, $\epsilon \in [0, 1]$ and the quality of prediction improves the smaller ϵ is.

The error sum ϵ is a possible representative of a whole ROC curve. For small event rates it is closely related to its closest approach to the ideal point (0,1), known as Kolmogorov-Smirnov distance. For comparison with [9], we therefore issued predictions using for every pair (L, η) the value of p_c in Eq. (4), for which the ROC curve is closest to the ideal point (0,1). We report the error sums ϵ in Fig. 5 for different lower event magnitudes η (expressed by the corresponding event rate) for three different system sizes $L=64$,

$L=128$, and $L=256$. The symbols show results taken from Fig. 2 in [9] for a system size of $L=256$ and an implementation of the average height precursor which resulted in the smallest prediction error sum. As can be seen, the prediction quality compares rather well for equal system size. However, we attain this level of predictability using only past avalanche size data, no knowledge of other internal variables is required.

Figure 5 seems to indicate that smaller systems have a higher predictability, which would be in contrast to the ROC curves of Fig. 4. However, the symbols plotted on the curves in Fig. 5 correspond to specific fractions of the maximal avalanche size between 1% and 35% (from large to small event rates; for $L=256$ too few events exist for $\eta=0.35s_{\max}$ so this point was left out). Except for statistical fluctuations for small event rates which mirror the ROC curves, the points corresponding to the same fraction also show the same error sum, i.e., they form triples of identical ϵ values. Therefore, this confirms that predictability is determined by the avalanche size in relation to its maximal possible value $s_{\max}(L)$ and not by their rates of occurrence. If, however, one wishes to predict the largest x percent of events in a system, then indeed predictability *decreases* with increasing system size, since the rate of events exceeding a fixed fraction of $s_{\max}(L)$ decreases approximately with L^{-2} .

We have presented predictions of large avalanches based on the joint distribution of events and a decision variable which is solely derived from past avalanche data. Previous work employed additional internal system parameters as decision variables [8,9], which reduce applicability in practical prediction tasks, where often internal system states are not measurable. This work presents three relevant results: first, particularly large events in a close to SOC system can be predicted on the basis of past observations. Second, the predictive power stems from temporal correlations which are pure finite-size effects, i.e., in the infinite system size limit predictability disappears as all avalanches become independent of each other. Third, under variation of system size predictability persists, if the magnitude η used to define extreme events is scaled linearly in the maximal possible avalanche size. However, the rate of occurrence of such large events decreases dramatically with increasing system size. As a consequence of the relevance of finite-size effects, events are the better predictable the larger they are. Considering that natural systems which are candidates for showing SOC-like behavior are finite, our results suggest to search for finite-size effects in time series recordings such as seismic activity or fracture. However, current literature indicates that both large earthquakes [24] and large events in an earthquake model of SOC [25] are clustered rather than evenly spread so the temporal correlations in their occurrence seem to stem from mechanisms different from the finite-size effects in the Abelian sandpile model.

- [1] *The Science of Disasters. Climate Disruptions, Heart Attacks, and Market Crashes*, edited by A. Bunde, J. Kropp, and H. J. Schellnhuber (Springer, Berlin, 2002).
- [2] *Extreme Events in Nature and Society*, edited by S. A. Albeverio, V. Jentsch, and H. Kantz (Springer, Berlin, 2006).
- [3] P. Bak, C. Tang, and K. Wiesenfeld, Phys. Rev. Lett. **59**, 381 (1987); Phys. Rev. A **38**, 364 (1988).
- [4] See the online debate in www.nature.com/nature/debates/earthquake.
- [5] Z. Olami, Hans Jacob S. Feder, and K. Christensen, Phys. Rev. Lett. **68**, 1244 (1992).
- [6] P. Bak, K. Christensen, L. Danon, and T. Scanlon, Phys. Rev. Lett. **88**, 178501 (2002).
- [7] V. I. Keilis-Borok and V. G. Kossobokov, Phys. Earth Planet. Inter. **61**, 73 (1990).
- [8] S. L. Pepke and J. M. Carlson, Phys. Rev. E **50**, 236 (1994).
- [9] A. B. Shapoval and M. G. Shnirman, Int. Jour. Mod. Phys. C **17**, 1777 (2006).
- [10] J. Rosendahl, M. Vekic, and J. E. Rutledge, Phys. Rev. Lett. **73**, 537 (1994); I. Sammis and J. M. Carlson, *ibid.* **74**, 4357 (1995); J. Rosendahl, M. Vekic, and J. E. Rutledge, *ibid.* **74**, 4358 (1995).
- [11] O. Ramos, E. Altshuler, and K. J. Måløy, Phys. Rev. Lett. **102**, 078701 (2009).
- [12] D. Dhar, Phys. Rev. Lett. **64**, 1613 (1990).
- [13] S. Lübeck and K. D. Usadel, Phys. Rev. E **55**, 4095 (1997).
- [14] L. P. Kadanoff, S. R. Nagel, L. Wu, and S. M. Zhou, Phys. Rev. A **39**, 6524 (1989).
- [15] S. S. Manna, Physica A **179**, 249 (1991).
- [16] E. Milshtein, O. Biham, and S. Solomon, Phys. Rev. E **58**, 303 (1998).
- [17] A. Garber and H. Kantz, Eur. Phys. J. B **67**, 437 (2009).
- [18] G. Boffetta, V. Carbone, P. Giuliani, P. Veltri, and A. Vulpiani, Phys. Rev. Lett. **83**, 4662 (1999).
- [19] M. Paczuski, S. Boettcher, and M. Baiesi, Phys. Rev. Lett. **95**, 181102 (2005).
- [20] A. Corral, Phys. Rev. Lett. **95**, 159801 (2005).
- [21] J. Neyman and E. S. Pearson, Philos. Trans. R. Soc. London **231**, 289 (1933).
- [22] S. Hallerberg, J. Bröcker, and H. Kantz, *Nonlinear Time Series Analysis In The Geosciences*, Lecture Notes in Earth Sciences (Springer, Berlin, 2008).
- [23] J. P. Egan, *Signal Detection Theory and ROC analysis* (Academic, New York, 1975).
- [24] Y. Kagan and D. Jackson, Geophys. J. Int. **104**, 117 (1991).
- [25] K. Christensen and Z. Olami, J. Geophys. Res. **97**, 8729 (1992).