Prediction in a driven-dissipative system displaying a continuous phase transition using machine learning

Chon-Kit Pun,[1] Sakib Matin,[1] W. Klein,[1,2] and Harvey Gould[1,3]

[1] *Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
[2] *Center for Computational Science, Boston University, Boston, Massachusetts 02215, USA*
[3] *Department of Physics, Clark University, Worcester, Massachusetts 01610, USA*

(Received 2 August 2019; accepted 27 December 2019; published 4 February 2020)

Prediction in complex systems at criticality is believed to be very difficult, if not impossible. Of particular interest is whether earthquakes, whose distribution follows a power-law (Gutenberg-Richter) distribution, are in principle unpredictable. We study the predictability of event sizes in the Olmai-Feder-Christensen model at different proximities to criticality using a convolutional neural network. The distribution of event sizes satisfies a power law with a cutoff for large events. We find that predictability decreases as criticality is approached and that prediction is possible only for large, nonscaling events. Our results suggest that earthquake faults that satisfy Gutenberg-Richter scaling are difficult to forecast.

DOI: 10.1103/PhysRevE.101.022102

I. INTRODUCTION

A subclass of driven-dissipative systems modeled by a two-dimensional cellular automaton has been proposed to understand the existence of power laws in many complex systems. Examples include the Bak-Tang-Wiesenfeld sandpile model [1], the Rundle-Jackson model [2], and the Olami-Feder-Christensen (OFC) model [3]; the latter two models have been used to gain insight into the nature of earthquakes. Many earthquake fault systems display a power-law event size distribution spanning many orders of magnitude. Such a power-law distribution is known as Gutenberg-Richter scaling in the seismology literature [4]. For example, the Gutenberg-Richter scaling in Southern California (1984–2000) spans about six orders of magnitude [5].

There has been substantial interest in forecasting or predicting earthquakes. However, it has been conjectured that systems at criticality are inherently unpredictable [1]. That is, events of different sizes that satisfy a scale-free distribution are due to the same physical mechanism, and thus there are no distinct precursors to distinguish one event from another [6]. The idea of unpredictability at criticality has been challenged over the years. One school of thought [7] has proposed that very large events are due to inherently different mechanisms such as self-reinforcement, synchronization [8], and nucleation [9], and they are thus in principle distinguishable from smaller events. Some support for this proposal is the use of a technique called the log-periodic-power-law fitting procedure; it has been shown to successfully predict large, non-power-law events, such as ruptures in materials [10,11] and the end of financial bubbles [12,13].

In this paper, we address the question of predictability near and at criticality by applying machine learning to the OFC model. Previous work [14] has shown that predictability in the OFC model decreases as the conservative limit (a critical point) is approached. We find results consistent with Ref. [14] and investigate the predictability of events near another critical point in the OFC model: the recently observed noise transition critical point [15]. By using a convolutional neural network (CNN), we find that the event sizes are more difficult to forecast as the critical point is approached and that only large events that do not satisfy power-law scaling can be successfully predicted.

II. CRITICALITY IN THE OFC MODEL

The Olami-Feder-Christensen (OFC) model [3] is a modified version of the spring-block model first proposed by Rundle and Jackson [2], which is a simplification of the Burridge-Knopoff model [16]. The nearest-neighbor OFC model that we will consider consists of a two-dimensional lattice of linear dimension L with each site initially assigned a random value of stress σ between the mean residual σ_R and the failure threshold σ_F. We denote the stress on each site, the stress grid, by the vector $\vec{\sigma} = (\sigma_1, \dots, \sigma_{N=L\times L})$. The system is then driven so that one site reaches σ_F, a procedure known as the zero velocity limit [2]. This site is said to fail, and its stress is reduced to $\sigma_R + \eta r$, where η is the magnitude of the noise and r is a uniform random variable in the range $[-1, 1]$. A failing site with stress σ distributes stress $(1-\lambda)(\sigma - \sigma_R - \eta r)/4$ to its four nearest-neighbor sites, where λ is the dissipation parameter. The failure of one site can trigger other sites to fail, thus creating an avalanche. The avalanche or event stops when the stress of all sites is less than σ_F. We denote the number of failing sites, or the size of the event, by s. The system is then driven again using the zero velocity limit. Periodic boundary conditions are used.

The OFC model is believed to approach criticality in the conservative limit $\lambda \to 0$ [17,18]. Recently, it has been found that even for $\lambda > 0$, there exists a phase transition at a critical value of the noise $\eta_c \approx 0.07$ [15]. This phase transition is

©2020 American Physical Society

TABLE I. The correlation squared (R^2) between the event size s and the average stress $\langle\sigma\rangle$, and the correlation square between s and the spatial variance var_σ for different values of the noise η. Due to this correlation, we will normalize the stress grid before training the machine.

η	$R^2(s, \langle\sigma\rangle)$	$R^2(s, \mathrm{var}_\sigma)$	η	$R^2(s, \langle\sigma\rangle)$	$R^2(s, \mathrm{var}_\sigma)$
0.03	0.48	7.40×10^{-3}	0.07	0.22	1.56×10^{-3}
0.04	0.42	2.00×10^{-3}	0.08	0.44	8.46×10^{-2}
0.05	0.52	5.00×10^{-2}	0.09	0.23	2.48×10^{-2}
0.06	0.49	1.2×10^{-2}	0.10	0.11	2.81×10^{-5}

characterized by an event size distribution n_s of the form

$$n_s \sim s^{-\tau} \exp[-(s/s_c)^\sigma] \tag{1}$$

with

$$s_c \propto (\eta - \eta_c)^{-1/\sigma}, \tag{2}$$

and $\tau = 1.04 \pm 0.14$ and $\sigma = 0.43 \pm 0.03$. The mean cluster size χ diverges as $(\eta - \eta_c)_+^{-\gamma}$, and the connectedness length diverges as $(\eta - \eta_c)_+^{-\nu}$ [19] with $\gamma = 2.01 \pm 0.14$ and $\nu = 1.20 \pm 0.13$, consistent with the scaling relations $\gamma = (3 - \tau)/\sigma$ and $\nu = (\tau - 1)/d\sigma$, where $d = 2$ is the spatial dimension. Note that n_s satisfies power-law scaling for $s \lesssim s_c$ [15].

III. SUPERVISED MACHINE LEARNING

Our goal is to predict the event size (the number of failed sites) given the stress grid before stress has been added using the zero-velocity limit and before the onset of an event. Table I shows, at different values of the noise η, the correlation squared between the event size s and the average stress $\langle\sigma\rangle$, and the correlation square between s and the spatial variance var_σ. There is a significant correlation between s and $\langle\sigma\rangle$. To force the CNN to learn higher-order features, we first remove the correlations between the event size s and the first and second moments of the stress grid. We normalize each stress grid $\vec{\sigma}$ by its average stress and the spatial variance. That is, we rescale the stress $\sigma_{i,\mu}$ at site i for sample μ to $\tilde{\sigma}_{i,\mu} \equiv (\sigma_{i,\mu} - \langle\sigma_\mu\rangle)/\sqrt{\mathrm{var}_\mu}$, where $\langle\sigma_\mu\rangle = 1/N \sum_{i=1}^N \sigma_{i,\mu}$ is the mean stress per site of sample μ and $\mathrm{var}_\mu \equiv \sum_{i=1}^N (\sigma_{i,\mu} - \langle\sigma_\mu\rangle)^2/N$ is the spatial variance of the stress. In the following, all references to the stress will be to the rescaled stress, and we will omit the tilde symbol. We will train the CNN regressor using the rescaled stress grid $\vec{\sigma}$, sampled with quasiuniform event sizes (see the Appendix).

To assess the performance of the machine, we show in Fig. 1 the predicted event size $\hat{s}$ versus the true event size s. The top row shows the event size distribution n_s for different values of η. The bottom row shows the predicted event sizes versus the true event sizes. We see that for $\eta < \eta_c$, the machine performs impressively at predicting events that are larger than s_c [see Fig. 1(a)]. For $\eta > \eta_c$, the machine performs less impressively [see Fig. 1(c)]. At $\eta = \eta_c$, the machine fails at predicting events of all sizes [Fig. 1(b)] and yields a constant equal to the average event size s in order to minimize the error. Similar behavior is observed for $\eta \neq \eta_c$ and below s_c. Note that the "cutoff" in n_s at $\eta = \eta_c$ in Fig. 1(b) is due to the finite-size effect.

In Fig. 2 we plot the testing error,

$$\mathrm{Err}(\log \hat{s}, \log s) \equiv \sqrt{\sum_{i=1}^{M} (\log s_i - \log \hat{s}_i)^2/M}, \tag{3}$$

as a function of η. Here M is the number of samples in the testing set. The reason for using $\log s$ instead of s in the error function is because of the larger fluctuations in $(\hat{s} - s)$ for larger s and because we are interested in the relative error

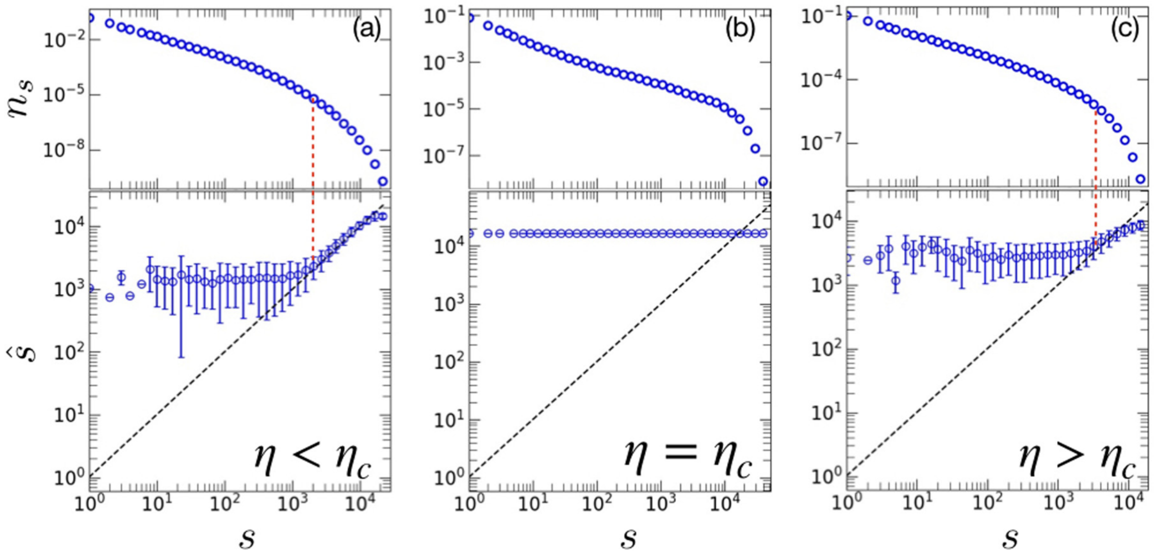

FIG. 1. The event size distribution n_s vs s (top row) and the true values of s vs the predicted $\hat{s}$ event sizes (bottom row) for (a) $\eta = 0.04 < \eta_c$, (b) $\eta = 0.07 = \eta_c$, and (c) $\eta = 0.09 > \eta_c$ (bottom row). Perfect prediction is represented by the dashed diagonal line. The vertical dotted line denotes $s = s_c$. Note that the CNN successfully predicts event sizes only for $s \gtrsim s_c$. The cutoff in n_s in (b) is a finite-size effect and does not fit the form in Eq. (1).

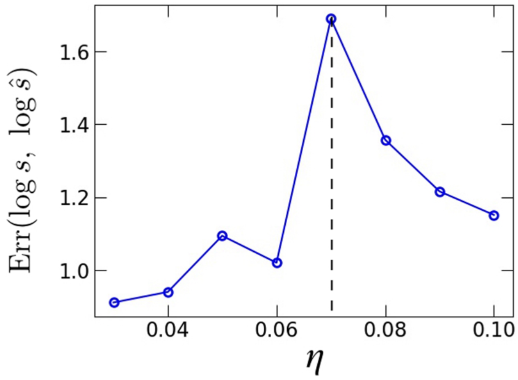

FIG. 2. The testing error of the predicted event sizes as a function of the noise η with $\mathrm{Err}(\log\hat{s},\ \log s) \equiv \sqrt{\sum_{i=1}^{M}(\log s_i - \log\hat{s}_i)^2/M}$. M is the number of samples in the testing set. The vertical dashed line indicates the location of the critical noise $\eta_c \approx 0.07$ [15]. Note the poorer predictability as the critical point (denoted by the vertical dashed line) is approached.

rather than the absolute error. The peak at η_c indicates that prediction is not possible in the OFC model at criticality using only the stress grid and the CNN.

We next look at how the values of the dissipation parameter λ affect the predictability of the system as $\lambda \to 0$. No scaling function has been found to fit the dependence of n_s on λ in the nearest-neighbor OFC model. Nevertheless, we can determine the cutoff $s_{c,\lambda}$ as the value of s for which n_s deviates from a power law (see Fig. 3). As λ decreases, the cutoff $s_{c,\lambda}$ increases. We observe that the onset of predictability is close to $s_{c,\lambda}$ and the trend persists for different values of λ.

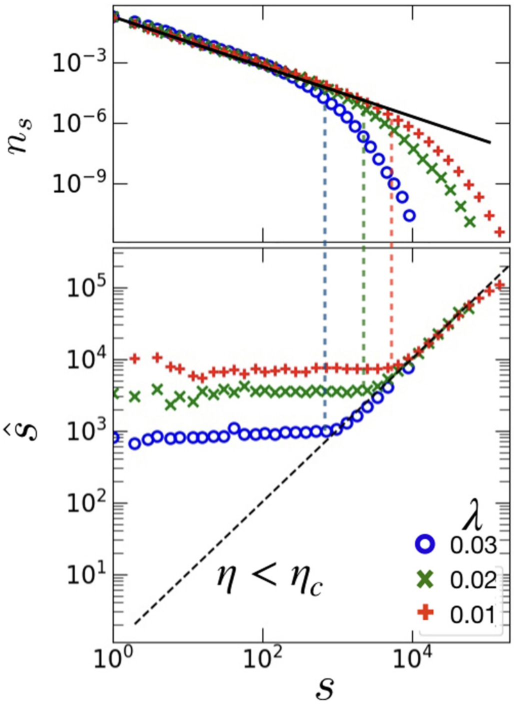

FIG. 3. Top: the event size distribution n_s for different values of λ. Bottom: the true values of s vs the predicted $\hat{s}$ event sizes for $\lambda = 0.01$ (+), 0.02 (×), and 0.03 (○). Note that the onset of predictability occurs for $s \gtrsim s_{c,\lambda}$. The vertical dashed line indicates the estimated value of the cutoff $s_{c,\lambda}$ for different values of λ.

IV. VISUALIZING THE CNN

We next explore the features that the machine has learned that allow it to successfully forecast the size of the nonscaling events, and we discuss why the critical events are difficult to forecast. We will use occlusion sensitivity analysis to identify the regions of importance of the images that are used by the CNN [20]. For example, the face of a dog is expected to contain the most relevant features in determining the type of animal. Hence, blocking the face of the dog should increase the classification error of the CNN. We implement a similar analysis by defining an occluded region in the stress grid and sweeping the occluded region across the entire image to create a map that shows the regions that are the most sensitive to the occlusion. In this way, we associate the region that gives the largest change in the predicted event size $\hat{s}$ with the region that is most useful in predicting the size of the event.

Since we see in Fig. 1 that the CNN can only predict events whose sizes are in the exponential cutoff region, we choose the sample from the exponential cutoff region for visualization (for details, see the Appendix). In Fig. 4 we visualize three randomly chosen samples for which $s > s_c$ for (a) $\eta < \eta_c$, (b) $\eta = \eta_c$, and (c) $\eta > \eta_c$. In the top row we show the failure maps, which correspond to the number of times that a site has failed. In the second row we show the sensitivity maps from the occlusion sensitivity analysis. Since we choose events of size $s > s_c$, an occlusion that yields a decrease in the predicted event size $\hat{s}$ implies a worse prediction. For $\eta \neq \eta_c$, the region that gives the largest decrease in $\hat{s}$ if occluded coincides with the failure region. We call the region with the largest decrease in $\hat{s}$ if occluded the sensitive region. In the third row, we plot the local average stress map. To determine this map, the local average stress of a site is computed by averaging the stress of sites within a square of linear dimension $b = 10$, centered at that site. Note that the region of high local stress overlaps with the failure and the sensitive regions. This consistency is reasonable because regions with high local stress have a greater probability of initiating and sustaining a larger event.

Among the 32 channels in the third layer of the CNN, we chose the channel that is visually the most similar to the structure of the failure region. We interpret the channel as the high level feature learned by the CNN. We plot the channels in the bottom row of Fig. 4. From these channels, we see that the machine has learned the connection between the high local stress and failure regions.

To understand why prediction is difficult at $\eta = \eta_c$, we look at the failure maps in the top row of Fig. 4. We see that the failure regions become more diffuse at $\eta = \eta_c$ compared to the more compact failure regions away from η_c. Although the local average stress map and the failure map remain qualitatively similar, the stress gradient between the high local stress region and the surrounding background is much smaller at $\eta = \eta_c$.

More quantitatively, we define a high stress region as a collection of nearest-neighbor sites whose local average stress is above the cutoff σ_c [21]. We measure the radius of gyration

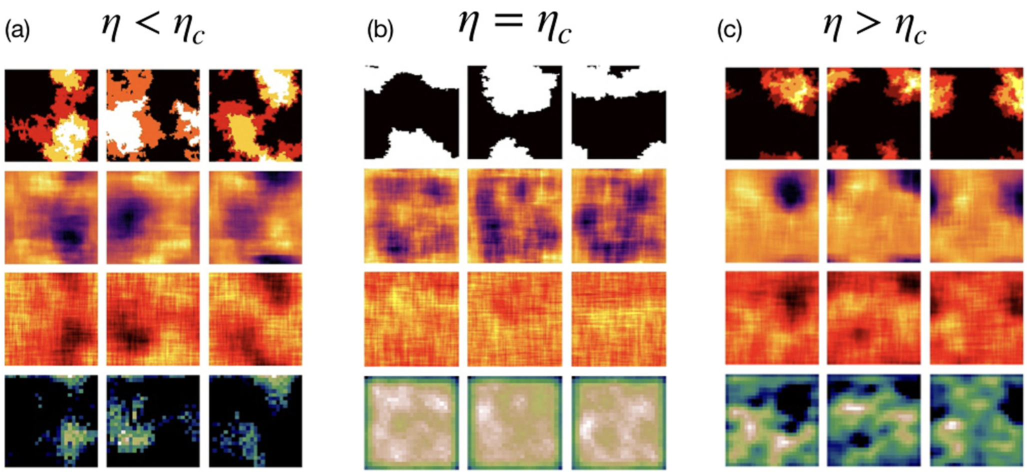

FIG. 4. Top row: the number of times that a site has failed (failure map). Brighter colors represent more failures. Second row: the sensitivity map from the occlusion sensitivity analysis. Darker regions are more sensitive to the occlusion of that region. Third row: local average stress map. Darker colors represent higher stress. Bottom row: several channels (features) chosen from the third layer in the CNN. Note that the four rows are structurally similar for (a) $\eta = 0.04 < \eta_c$ and (c) $\eta = 0.09 > \eta_c$. The three samples for each value of η are randomly chosen from the tail region of the event size distribution from the testing set (see the Appendix).

R_g of the largest high stress region in each sample, and we define the density of the high stress region $\tilde{\phi}$ as the sum of the local average stress within the area of radius R_g divided by πR_g^2. The density of the high stress region $\tilde{\phi}$ decreases as η_c is approached (see Fig. 5). The smaller density difference makes it more difficult for the machine to obtain the appropriate cutoff for the high stress region, thus making prediction more difficult. Multiple failures occur when sites fail more than once and are more prominent for very large events for $\eta \neq \eta_c$, which is why the machine underestimates the event sizes of very large events [see Figs. 1(a) and 1(c)].

V. DISCUSSION

Since we have normalized the stress grid by the average stress before training the machine, the first and second moments of the stress grid do not contain information that can be used by the CNN to predict event sizes. The fact that the machine learns the association between the region of high local average stress and the event size means that the machine has learned the optimal cutoff that separates the high stress regions from the low stress regions. This task becomes increasingly difficult as the critical point is approached because the high stress region becomes less distinguishable from its background.

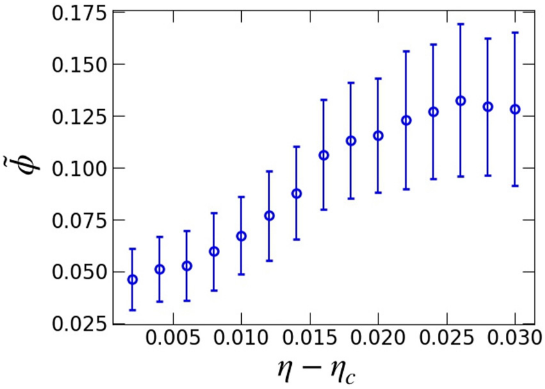

FIG. 5. The density of the high stress region $\tilde{\phi}$ vs $\eta_c - \eta_c$. We hypothesize that the decrease in predictability at η_c is due to the decrease in the density of the high stress region because the CNN has a more difficult time identifying the high stress region.

We have found evidence that events whose size distribution satisfies a power law lack distinguishable features that allow the machine to predict their size. This lack of distinguishable features is related to the difficulty of distinguishing between the fluctuations and the background at critical points [22]. For the large nonscaling events, there exists features that allow the machine to successfully predict the event sizes. Similar conclusions are found for the dissipation [17] transition. Our results suggest that large nonscaling events are qualitatively different from the smaller scaling events. This conclusion agrees with the conjecture [6] that prediction is not possible at a true critical point, where there is no deviation from a power law for large events.

It is known that small, large, and very large events in the long-range OFC model are due to different mechanisms, namely fluctuations about the spinodal critical point, failed nucleation, and arrested nucleation events, respectively [9,23,24]. These different mechanisms suggest that very large events are in principle distinguishable from other events. The caveat is that all three types of events follow a power law, albeit with different exponents. It would be interesting to see if a machine can learn the difference between the different scaling events. It is important to note that the failed and arrested nucleation events, despite the fact that they satisfy a power-law distribution, do not exhibit the same diffusive nature as the smaller events (spinodal fluctuations) on the Gutenberg-Richter scaling plot. This difference appears to be what the CNN picks up.

Real earthquake forecast also utilizes temporal information, which we have neglected in this work. An example of utilizing temporal information to predict "earthquakes" in a laboratory setting can be found in Ref. [25].

The connection between the OFC model and real earthquake faults is complicated. The complications involve not only the values of the exponents such as τ and σ, but also the relation between the noise-induced critical point in the model and the possible existence of such a critical point in real faults. The stress transfer range in real earthquake faults is governed by the long-range elastic force. For OFC models with long-range stress transfer, the values of σ and τ in the conservative limit $\lambda \to 0$ are different from those in the nearest-neighbor model [15]. We have investigated the noise transition in the next- and next-next-nearest-neighbor OFC model, and we observed qualitatively the same behavior; that is, the machine is unable to predict the size of the events that satisfy a power law. We also found that as the stress transfer range increases, the value of the critical noise, η_c, approaches zero. The values of τ and σ in this limit will be determined in future work.

In real earthquake faults, the geometry of the fault changes the exponents τ and σ in ways that are not well understood. The results in Ref. [18] indicate that there is a relation between the spinodal critical point in the OFC model and scaling in real faults, but the relation is not conclusive.

We note that the OFC model is a cellular automaton with no dynamics and no real friction force. The effect of a more realistic dynamics and a velocity weakened friction was studied in the long-range Burridge-Knopoff model [26]. It was found that the dynamics of the long-range Burridge-Knopoff model is much richer than that in the OFC model. It would be of much interest to apply machine learning to the long-range Burridge-Knopoff model.

Our main purpose has been to investigate the role of a critical point, indicated by the existence of scaling, on the possibility of forecasting with the aid of machine learning. This possibility is of great interest to the earthquake community because several investigators have conjectured that if the statistical distribution of earthquakes is generated by a critical point, or self-organized criticality, forecasting would not be possible. This work adds some evidence that this conjecture is correct. The addition of other characteristics that would lead to a stronger connection to real earthquake faults is something we are pursuing. However, our main result—that forecasting in the vicinity of a critical point, even with the assistance of machine learning, does not appear to be possible—has implications beyond the area of earthquake forecasting. One such area that we are actively pursuing is whether we can forecast the occurrence of nucleation events in metastable states near the spinodal critical point.

ACKNOWLEDGMENTS

We would like to thank Huang Shan, Pankaj Mehta, and Timothy Khouw of the Department of Physics at Boston University for useful discussions.

APPENDIX: SAMPLING METHOD AND CNN ARCHITECTURE

After discarding the transient (10^6 plate updates), we run the OFC model for an additional 10^7 plate updates and record the event sizes and the random number seed. We then construct the event size distribution and randomly choose five samples from each bin of the distribution (or the number of samples in that bin if there are fewer than five samples) and record the time of events in each bin. We then rerun the simulation using the same random number seed and save the stress grids at the recorded times. This procedure ensures that the number of samples in each bin remains the same for different values of the noise. This sampling method is more desirable than random sampling because the data sampled in this way are more "balanced," that is, the machine does not overlearn samples of any particular event size. We divide the data into a training set (63% of the data), a validation set (7% of the data), and a testing set (30% of the data). The validation set is used to early-stop the training process to prevent overfitting. The total amount of data varies from 20 500 to 164 951 for different values of the noise. We find that the performance of the machine remains qualitatively the same if we use the same amount of data for different values of the noise, but the performance of the machine still depends on the noise.

The architecture of the CNN consists of eight alternating layers of convolutional layers and maxpooling layers (four layers each). The depths of the convolutional layers are 8, 16, 32, and 64, each with a filter of size 5×5. We used zero padding on the boundaries to ensure the same size after each convolution. The output of the last maxpooling layer is connected to a fully connected neural network with one hidden layer of 25 nodes. All layers use relu (rectified linear unit) as the activation function except for the last layer, which uses a linear activation function. Dropout [27] with dropout rate $= 0.1$ is applied to the layer immediately before the fully connected layer. The particular choice of activation functions and structures is standard in the machine learning literature. A thorough discussion of the advantages of this particular CNN architecture can be found in Ref. [28].

The samples shown in Fig. 4 are drawn from the testing set, that is, the last 30% of the data. Specifically, the samples in Fig. 4 are drawn from populations of size (a) 10 997, (b) 49 485, and (c) 6150.

[1] P. Bak, C. Tang, and K. Wiesenfeld, Self-Organized Criticality: An Explanation of the $1/f$ Noise, Phys. Rev. Lett. **59**, 381 (1987).

[2] J. B. Rundle and D. D. Jackson, Numerical simulation of earthquake sequences, Bull. Seismol. Soc. Am. **67**, 1363 (1977).

[3] Z. Olami, H. J. S. Feder, and K. Christensen, Self-Organized Criticality in a Continuous, Nonconservative Cellular Automaton Modeling Earthquakes, Phys. Rev. Lett. **68**, 1244 (1992).

[4] B. Gutenberg and C. F. Richter, Frequency of earthquakes in California, Bull. Seismol. Soc. Am. **34**, 185 (1944).

[5] P. Bak, K. Christensen, L. Danon, and T. Scanlon, Unified Scaling Law for Earthquakes, Phys. Rev. Lett. **88**, 178501 (2002).

[6] P. Bak and K. Chen, The physics of fractals, Physica D **38**, 5 (1989).

[7] J. Laherrere and D. Sornette, Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales, Eur. Phys. J. B **2**, 525 (1998).

[8] D. Sornette, Dragon-kings, black swans and the prediction of crises, Int. J. Terraspace Sci. Eng. **2**, 1 (2009).

[9] W. Klein, M. Anghel, C. Ferguson, J. B. Rundle, and J. Sá Martins, Statistical analysis of a model for earthquake faults with long-range stress transfer, in *Geocomplexity and the Physics of Earthquakes*, edited by J. B. Rundle, D. L. Turcotte, and W. Klein, Geophysical Monograph Series Vol. 120 (American Geophysical Union, 2000), pp. 43–71.

[10] J.-C. Anifrani, C. Le Floc'h, D. Sornette, and B. Souillard, Universal log-periodic correction to renormalization group scaling for rupture stress prediction from acoustic emissions, J. Phys. I **5**, 631 (1995).

[11] A. Johansen and D. Sornette, Critical ruptures, Eur. Phys. J. B **18**, 163 (2000).

[12] J. A. Feigenbaum and P. G. O. Freund, Discrete scale invariance in stock markets before crashes, Int. J. Mod. Phys. B **10**, 3737 (1996).

[13] D. Sornette, A. Johansen, and J.-P. Bouchaud, Stock market crashes, precursors and replicas, J. Phys. I **6**, 167 (1996).

[14] S. L. Pepke and J. M. Carlson, Predictability of self-organizing systems, Phys. Rev. E **50**, 236 (1994).

[15] S. Matin, C.-K. Pun, H. Gould, and W. Klein, Effective ergodicity breaking phase transition in a driven-dissipative system, Phys. Rev. E **101**, 022103 (2020).

[16] R. Burridge and L. Knopoff, Model and theoretical seismicity, Bull. Seismol. Soc. Am. **57**, 341 (1967).

[17] K. Christensen and Z. Olami, Scaling, phase transitions, and nonuniversality in a self-organized critical cellular-automaton model, Phys. Rev. A **46**, 1829 (1992).

[18] C. A. Serino, K. F. Tiampo, and W. Klein, New Approach to Gutenberg-Richter Scaling, Phys. Rev. Lett. **106**, 108501 (2011).

[19] D. Stauffer, Scaling theory of percolation clusters, Phys. Rep. **54**, 1 (1979).

[20] M. D. Zeiler and R. Fergus, Visualizing and understanding convolutional networks, *European Conference on Computer Vision* (Springer, Switzerland, 2014), pp. 818–833.

[21] The local average stress follows a normal distribution with mean $\langle\tilde{\sigma}\rangle$ and variance $\widetilde{\text{var}}$. We define the cutoff σ_c to be $\sigma_c = \langle\tilde{\sigma}\rangle + \sqrt{\widetilde{\text{var}}}$.

[22] A. Coniglio and W. Klein, Clusters and Ising critical droplets: a renormalisation group approach, J. Phys. A **13**, 2775 (1980).

[23] M. Anghel, W. Klein, J. B. Rundle, and J. S. S'a Martins, Scaling in a cellular automaton model of earthquake faults, arXiv:cond-mat/0002459.

[24] J. Xia, C. A. Serino, M. Anghel, J. Tobochnik, H. Gould, W. Klein, and J. B. Rundle (unpublished).

[25] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, and P. A. Johnson, Machine Learning Predicts Laboratory Earthquakes, Geophys. Res. Lett. **44**, 9276 (2017).

[26] J. Xia, H. Gould, W. Klein, and J. B. Rundle, Simulation of the Burridge-Knopoff Model of Earthquakes with Variable Range Stress Transfer, Phys. Rev. Lett. **95**, 248501 (2005).

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).

[28] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. **810**, 1 (2019).