Autocorrelations from emergent bistability in homeostatic spiking neural networks on neuromorphic hardware

Benjamin Cramer,[1] Markus Kreft,[1] Sebastian Billaudelle,[1] Vitali Karasenko,[1] Aron Leibfried,[1] Eric Müller,[1] Philipp Spilger,[1] Johannes Weis,[1] Johannes Schemmel,[1] Miguel A. Muñoz,[2] Viola Priesemann,[3,4,*] and Johannes Zierenberg[3,†]

[1]*Kirchhoff-Institute for Physics, Im Neuenheimer Feld 227, Heidelberg University, Heidelberg, Germany*
[2]*Departamento de Electromagnetismo y Física de la Materia e Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, E-18071 Granada, Spain*
[3]*Max Planck Institute for Dynamics and Self-Organization, Am Faßberg 17, 37077 Göttingen, Germany*
[4]*Institute for the Dynamics of Complex Systems, University of Göttingen, Friedrich-Hund-Platz 1, 37077 Göttingen, Germany*

(Received 23 August 2022; accepted 8 June 2023; published 19 July 2023)

A fruitful approach towards neuromorphic computing is to mimic mechanisms of the brain in physical devices, which has led to successful replication of neuronlike dynamics and learning in the past. However, there remains a large set of neural self-organization mechanisms whose role for neuromorphic computing has yet to be explored. One such mechanism is homeostatic plasticity, which has recently been proposed to play a key role in shaping network dynamics and correlations. Here, we study—from a statistical-physics point of view—the emergent collective dynamics in a homeostatically regulated neuromorphic device that emulates a network of excitatory and inhibitory leaky integrate-and-fire neurons. Importantly, homeostatic plasticity is only active during the training stage and results in a heterogeneous weight distribution that we fix during the analysis stage. We verify the theoretical prediction that reducing the external input in a homeostatically regulated neural network increases temporal correlations, measuring autocorrelation times exceeding 500 ms, despite single-neuron timescales of only 20 ms, both in experiments on neuromorphic hardware and in computer simulations. However, unlike theoretically predicted near-critical fluctuations, we find that temporal correlations can originate from an emergent bistability. We identify this bistability as a fluctuation-induced stochastic switching between metastable active and quiescent states in the vicinity of a nonequilibrium phase transition. Our results thereby constitute a complementary mechanism for emergent autocorrelations in networks of spiking neurons with implications for future developments in neuromorphic computing.

DOI: 10.1103/PhysRevResearch.5.033035

I. INTRODUCTION

Neuromorphic computing covers a variety of brain-inspired computers, devices, and models that function fundamentally differently from common von Neumann architectures [1,2]. For instance, one can *emulate* the dynamics of neuron membrane potentials and synaptic currents in analog electronic circuits [3–7]. In general, the hardware-specific information processing and storage call for hand-in-hand development of hardware and corresponding algorithms, which can be guided by modern artificial intelligence and neuroscience likewise [8]. A complementary approach is to build customizable, large-scale neuromorphic architectures that can implement brain-inspired plasticity for self-organization, for instance, BrainScaleS [9] or Loihi [10]. These devices can exhibit diverse emergent population dynamics that depend on, among other things, model parameters, plasticity, or network architecture and that may be useful for future developments in neuromorphic computing. For instance, emergent temporal correlations imply information integration over time that can be important for understanding sequential-input-like language, i.e., integrating syllables into words, words into sentences, and sentences into meaning [11–13]. Understanding emergent timescales in neural networks is thus among the basic prerequisites for designing recurrent neural language processing models from scratch.

Recent theoretical and experimental results have emphasized the importance of homeostatic plasticity to shape the timescales of neural population dynamics [14–16]. Homeostatic plasticity is a negative feedback that adapts local neural properties to achieve a stable firing rate [17–19]. For homeostatically regulated excitable systems, one can prove analytically that lowering the input strength induces an increase in the recurrent coupling and hence increases the autocorrelation time through close-to-critical fluctuations [14,15]. These predictions are consistent with experiments on monocular deprivation, where partial reduction of input initially disrupts population activity before homeostatic plasticity tunes cortical networks back towards criticality [16].

*viola.priesemann@ds.mpg.de
†johannes.zierenberg@ds.mpg.de

Moreover, they provide a potential explanation for the experimentally observed increase in autocorrelation time along the hierarchical anatomy of the cortex [11,20–24]: Timescales are shorter in primary sensory regions and longer in higher-order cortical regions. In the context of neuromorphic computing, homeostatic plasticity was shown to serve as a guiding principle to tune neuromorphic hardware for optimal task performance [25].

However, empirical observations of large, emergent autocorrelations seem to contradict prior theoretical predictions for networks of excitatory-inhibitory (E-I) leaky integrate-and-fire (LIF) neurons. While empirical estimates of neural autocorrelation times range from $O(10\,\mathrm{ms})$ to $O(1\,\mathrm{s})$ [11,20–24,26–29], early theories and models of networks of LIF neurons in an E-I balanced state [30,31] predict almost vanishing mean correlations. Instead, more recent reassessments find conditions under which larger correlations can emerge [32–35] (see also Refs. [36,37] for overviews). Focusing on temporal correlations, recent developments in dynamic mean-field theory [34,38,39] reveal parameter ranges with larger emergent autocorrelation times. However, these autocorrelation times are still on the order of characteristic times of the membrane potential or synaptic current, which are typically $O(10\,\mathrm{ms})$, and thus distinctively below the ones observed experimentally.

In this paper, we study emergent collective dynamics and autocorrelation times in networks of excitatory and inhibitory LIF neurons emulated on the BrainScaleS neuromorphic system. This system provides large flexibility for programmable plasticity rules [9] and hence allows for homeostatic plasticity during a training phase. We verify that training with reduced external input strength induces increasing autocorrelation times in the test phase that can be more than 20 times larger than the decay time of the membrane potential of individual units. Since we are using the BrainScaleS-2 single chip, which is limited to 512 neurons, we complement our experiments with a numerical finite-size scaling analysis that reveals progressively larger autocorrelation times with increasing system size. Surprisingly, we find that in our setup, autocorrelations are not generated by close-to-critical fluctuations [14], but originate from an emergent bistability in the population firing rate. To explain this bistability, we derive a simple mean-field theory for driven excitable systems that reveals a fluctuation-induced switching between a metastable active phase and a quiescent phase, reminiscent of so-called *up and down* states in brain networks [40–43]. We finish with a discussion of how emergent bistability can affect biological and artificial neural networks, as well as other finite systems with an absorbing-to-active transition that are driven by external noise.

II. MODEL AND METHODS

To study emergent collective dynamics in homeostatic neuromorphic devices, we combine experiments on an actual neuromorphic device (BrainScaleS-2), computer simulations, and a phenomenological mean-field theory. In this section, we first describe the basic ingredients of the neuromorphic hardware under consideration (Sec. II A), formulate a mathematical model that can be implemented on this hardware (Sec. II B), build a computer simulation that reproduces the resulting dynamics (Sec. II C), and define relevant observables to study population dynamics (Sec. II D).

A. Neuromorphic hardware

BrainScaleS-2 [5,9,44] is a mixed-signal neuromorphic architecture that allows one to emulate networks of up to 512 LIF neurons [Fig. 1(b)]. The term *emulation* is used to clearly distinguish between this physical implementation, where each observable has a measurable counterpart on the neuromorphic chip, and standard software *simulations* on conventional hardware (see below). In particular, neurons are implemented as electrical circuits that emulate LIF dynamics in a time-continuous and parallel manner. The system further consists of an array of 256×512 physically implemented current-based synapses that support near-arbitrary topologies. Their dynamics emulate leaky currents and feature coupling strengths w_{ij} with a precision of 6 bits, i.e., 64 discrete values, which limits synaptic weights to integers in the range [0,63]. More technical details are provided in Appendix A. Due to the analog implementation, time constants are determined by the electrical components on the substrate and are rendered approximately a factor 1000 times faster than the ones of their biological archetype. Within this paper, all referenced timescales are converted to the equivalent biological time domain unless otherwise stated.

Homeostatic plasticity is implemented on chip by a specialized, freely programmable processor unit: the plasticity processing unit (PPU) [5]. The PPU is able to update the synaptic weights of 128 synapses in parallel. To measure local spike rates relevant for our plasticity rule (see below), we draw on dedicated circuits within each neuron that count the number of emitted spikes.

The system comes with specialized accelerators for the drawing of random numbers [45]. These facilitate an on-chip generation of Poisson-distributed input spikes as well as the efficient implementation of the stochastic homeostatic regulation without additional communication bottlenecks. The only remaining communication with the host system consists of the transfer of instructions for configuring the BrainScaleS-2 system at the beginning and the readout of the result at the end of an experiment [46,47], making the hardware implementation very fast.

The neuromorphic chip is subject to variations both in space and in time: First, the analog implementation causes temporal noise within the model dynamics, and second, the production process necessarily leads to small variations across electrical components. The latter variations can, however, be mitigated by exploiting the configurability of the BrainScaleS-2 system by resorting to calibration routines [47], thereby reducing the parameter spread across neurons (see Appendix B). The remaining variability of parameters can be quantified by their mean and standard deviation (Table I).

B. Neural network model

As a minimal model of biological spiking neurons, we consider a recurrent network of $N = 512$ (unless otherwise

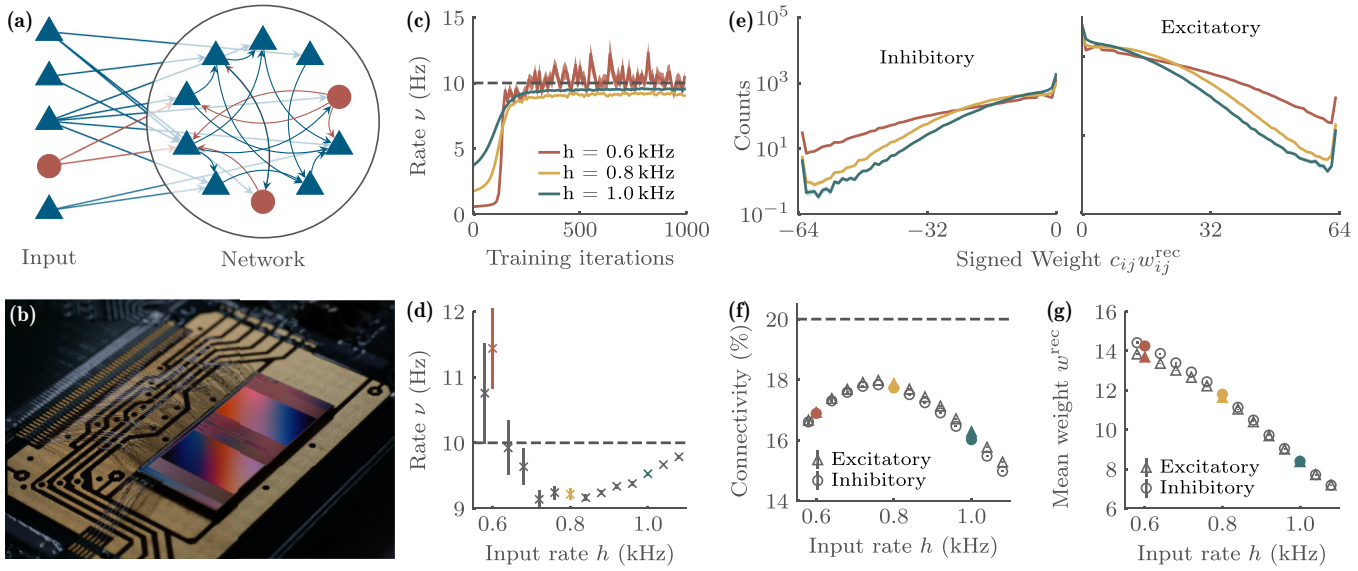

FIG. 1. Reducing the input strength to homeostatically regulated networks of E-I LIF neurons strengthens recurrent connections. (a) Illustration of a random network topology with 80% excitatory (blue triangles) and 20% inhibitory (red circles) neurons. (b) Image of the BrainScaleS-2 neuromorphic chip. Image taken from Ref. [46]. (c) Homeostatic plasticity regulates the population rate ν close to a target value (dashed line). (d) For a broad range of external input rates, ν approximates the target rate. (e) The stochastic homeostatic regulation leads to heterogeneous weight distributions for both inhibitory and excitatory synapses. The counts of excitatory weights exceed the inhibitory ones by a factor of 4 due to the imposed E-I ratio. (f) The effective connectivity, defined as the percentage of nonzero recurrent synapses ($c_{ij}w_{ij}^{\mathrm{rec}} \neq 0$), does not saturate at its maximum (dashed line) for decreasing input strengths. (g) However, the mean weight increases to compensate for a reduction of input.

stated) LIF neurons coupled to an input layer consisting of $N/2$ Poisson sources [Fig. 1(a)]. The model is built to reflect the architecture of BrainScaleS-2. Each LIF neuron integrates spikes from, on average, $K^{\mathrm{rec}} = 100$ recurrent neurons of the network and from, on average, K^{ext} external neurons of the input layer that will be varied to adjust the strength of external input. The *physical* connection between neurons i and j is randomly realized, $c_{ij} = \{-1, 0, 1\}$, and further weighted by an integer-valued coupling weight $w_{ij} \in [0, 63]$. Neurons can be either excitatory or inhibitory, which is reflected in the sign of c_{ij} for a given neuron j for all outgoing coupling synapses. Also, in analogy with cortical networks [48], 20% of the neurons in both the network and the input layer are inhibitory. While the recurrent neurons are LIF neurons, input sources generate spikes independently as a Poisson process with rate $\nu^{\mathrm{ext}} = 10\,\mathrm{Hz}$, which amounts to an average input rate per recurrent neuron of $h = \nu^{\mathrm{ext}} K^{\mathrm{ext}}$.

TABLE I. Model parameters. All parameters are given in the equivalent biological time domain. The errors indicate the standard deviation.

Parameter	Symbol	Value
Membrane capacitance	C^{m}	$(2.4 \pm 0.2)\,\mathrm{pF}$
Threshold potential	u^{thresh}	$(741 \pm 06)\,\mathrm{mV}$
Leak potential	u^{leak}	$(458 \pm 43)\,\mathrm{mV}$
Reset potential	u^{reset}	$(324 \pm 06)\,\mathrm{mV}$
Membrane time constant	τ^{m}	$(21.5 \pm 1.5)\,\mathrm{ms}$
Excitatory synaptic time constant	$\tau^{\mathrm{s,exc}}$	$(5.3 \pm 0.3)\,\mathrm{ms}$
Inhibitory synaptic time constant	$\tau^{\mathrm{s,inh}}$	$(5.4 \pm 0.2)\,\mathrm{ms}$
Synaptic delay	τ^{d}	$(1.0 \pm 0.1)\,\mathrm{ms}$
Refractory period	τ^{ref}	2.0 ms
Excitatory weight scaling factor	γ^{exc}	$(0.57 \pm 0.10)\,\mathrm{nA}$
Inhibitory weight scaling factor	γ^{inh}	$(0.67 \pm 0.10)\,\mathrm{nA}$
Number of recurrent synapses per neuron	K^{rec}	100
Number of neurons	N	512
Input weight	w^{in}	17
Learning rate	λ	0.468 75
Target rate	ν^{*}	10 Hz
Update probability	p	2.3%
Number of updates	n	1000
External rate	ν^{ext}	10 Hz
Static experiment duration	T	100 s

The dynamics of a recurrent LIF neuron i is modeled by a leaky membrane potential $u_i(t)$ given by

$$\tau_i^{\mathrm{m}} \dot{u}_i(t) = u_i^{\mathrm{leak}} - u_i(t) + R_i I_i(t), \tag{1}$$

where $\tau_i^{\mathrm{m}} = C_i^{\mathrm{m}} R_i$ is the membrane time constant with the membrane capacitance C_i^{m} as well as the resistance R_i, and u_i^{leak} is the leak potential. Similarly, $I_i(t)$ denotes a leaky synaptic current which is described by

$$\tau_i^{\mathrm{s}} \dot{I}_i(t) = -I_i(t) + \gamma_i \sum_j c_{ij} w_{ij} \sum_k \delta\left(t - t_j^k - \tau^{\mathrm{d}}\right), \tag{2}$$

where τ_i^{s} is the synaptic time constant, γ_i is a scale factor, w_{ij} are dimensionless coupling weights between neurons i and j (which covers recurrent and external presynaptic neurons), and $\sum_k \delta(t - t_j^k - \tau^{\mathrm{d}})$ is the spike train of neuron j that arrives at neuron i with past spike times t_j^k and synaptic time delay τ^{d}. Spikes are generated once a neuron's membrane potential crosses a threshold, i.e., $u_i(t) > u_i^{\mathrm{thres}}$, after which the membrane potential is reset to u_i^{reset}, where it remains for

the duration of the refractory period τ^{ref}. Explicit parameters were motivated by neurophysiology but are subject to hardware constraints (cf. Table I).

While external coupling weights are fixed to $w_{ij}^{\text{in}} = 17$, recurrent couplings are initialized at $w_{ij}^{\text{rec}} = 0$ and subject to homeostatic plasticity during a training phase to regulate the single-neuron firing rate around a target of $\nu^* = 10\,\text{Hz}$. More specifically, we implement homeostatic plasticity as an iterative, stochastic update of all realized ($c_{ij} \neq 0$) recurrent weights w_{ij}^{rec}. Each iteration consists of a 5s time window for which we record the firing rate ν_i of each individual neuron. In between iterations, we stochastically update each recurrent weight w_{ij}^{rec} independently with probability p by an amount

$$\Delta w_{ij} = \lambda(\nu^* - \nu_j), \tag{3}$$

which depends only on the local information of the postsynaptic neuron i and where λp sets the timescale of the homeostasis. While our results in the main text are obtained with probabilistic weight changes in order to overcome artifacts from the limited precision of w_{ij} (see Appendix C for the effect of p and λ), we obtain similar results when instead updating each weight by Δw_{ij} plus integer rounding noise (see Supplemental Material [49]). To preserve the effective sign of excitatory and inhibitory weights, w_{ij} are restricted to positive values and saturate at zero. Besides this, the proposed simple update scheme does not distinguish between excitatory and inhibitory couplings. After the homeostatic update, the network dynamics are evolved for about 1 s in order to allow the network activity to re-equilibrate before assessing ν_i for the next update. Importantly, we only employ homeostatic plasticity during the training stage of our experiment. All correlation analyses are evaluated on spike data from a testing phase (typically $T = 100\,\text{s}$) with fixed synaptic weights.

C. Computer simulation

For comparison and finite-size scaling analysis, we use additional computer simulations where we employ the PYTHON simulation package BRIAN2 [50]. This package generates from the differential equations (1) and (2) a discrete-time Euler integration scheme together with full control over all system parameters. We use these simulations to (i) cross-validate the results from the neuromorphic chip (see Supplemental Material) and (ii) analyze how changing system sizes, $N = \{256, 512, 768, 1024\}$, beyond the hardware-limiting constraints, affect our conclusion. The integration time step is set to $\Delta t = 50\,\mu\text{s}$ to approach the time-continuous nature of the BrainScaleS-2 system. To closely mimic the emulated networks, we draw the individual neuron parameters from Gaussian distributions specified by the measured parameter variability of the neuromorphic chip (Table I). In addition, independent temporal noise with standard deviation $\sigma\sqrt{2 * \tau_i^{\text{m}}}$, with $\sigma = 2\,\text{mV}$, is added to Eq. (1).

D. Observables

The main observables we consider are derived from the *instantaneous population firing rate* $\nu(t)$, defined as the number of network spikes within a time bin Δt

$$\nu(t) = \frac{1}{N\Delta t}\sum_{i=1}^{N}\sum_{k=1}^{S_i}\int_t^{t+\Delta t}\delta\big(t - t_i^k\big), \tag{4}$$

where t_i^k are the spike times of neuron i, S_i is the number of spikes emitted by neuron i, and $\Delta t = 5\,\text{ms}$.

From a time series $\nu(t)$, we calculate the stationary *autocorrelation function*

$$C(t') = \frac{\text{Cov}[\nu(t+t')\nu(t)]}{\text{Var}[\nu(t)]}, \tag{5}$$

where t' are multiples of Δt. From the decay of the autocorrelation function it is possible to derive the timescale(s) of temporal correlations. In our case, we found the autocorrelation function to be described by a single exponential decay, $C(t') = e^{-t'/\tau_{\text{AC}}}$, and extracted a single autocorrelation time τ_{AC} by fit routines.

To estimate statistical errors, we average over 50 independent experiments.

III. RESULTS

A. Homeostatically regulated neuromorphic hardware compensates lack of external input by strengthening recurrent connections

To begin, we verify that the experimental setup—the neuromorphic chip with homeostatic regulation during development—reaches a stationary dynamical state with firing rates sufficiently close to the target rate. Starting from the initial condition of zero recurrent weights ($w_{ij}^{\text{rec}} = 0$), we observe for our chosen parameters a transient relaxation behavior that reaches a stationary firing rate after about 200 update iterations, independent of the external input rate h [Fig. 1(c)]. Note that for this representation, the firing rate is evaluated over an interval of $T = 100\,\text{s}$ between iterations, and further averaged over 50 network realizations. One can see that for larger values of h (blue curve) the relaxation is smoother than for lower values of h (red curve). The stationary firing rates are close to the target rate $\nu^* = 10\,\text{Hz}$ [Fig. 1(d)]; however, there is a systematic h dependence that presumably originates from the firing rate being a nonlinear function of the mean coupling, $\nu(\langle w\rangle)$, as observed in mean-field calculations of E-I networks [34]. Since we find consistent results for reference computer simulations (see Supplemental Material), we conclude that the experimental setup reliably self-organizes into a stationary dynamics state with neuron firing rates sufficiently close to the target rate.

We next investigate how homeostatically regulated E-I networks compensate a reduction of external input with a strengthening of recurrent connections [Figs. 1(e)–1(g)]. In particular, we find that the histograms of both inhibitory as well as excitatory recurrent coupling weights become flatter with decreasing h, indicating strong heterogeneity [Fig. 1(e)]. Interestingly, the effective connectivity—the fraction of all physical K^{rec} recurrent weights that are not zero—does not reach its maximum theoretical value of $K^{\text{rec}}/N = 100/512 \approx 20\%$ [Fig. 1(f)]. Instead, it even decreases for low h, which is likely a consequence of the strong variability of rates between iterations [cf. Fig. 1(c)] that results in large weight changes

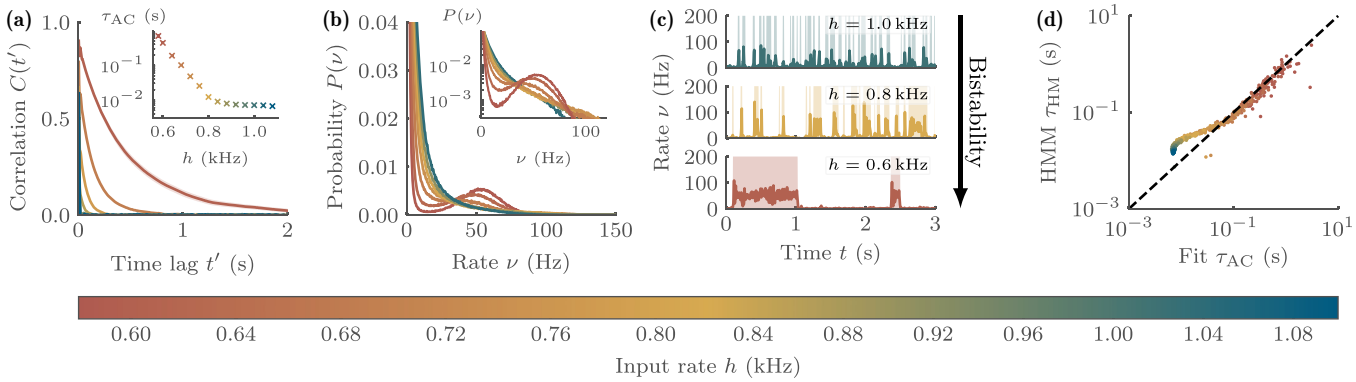

FIG. 2. Reducing the input strength increases autocorrelation of the network rate through emergent bistability. (a) For low input rates h, the population activity exhibits exponentially shaped autocorrelations $C(t')$ with autocorrelation times τ_{AC} significantly exceeding the largest single-neuron timescale. (b) In this regime, the distribution $P(\nu)$ of the population rate ν shows a bimodal trend. (c) The associated phases of high and low ν can be fitted by a two-state HMM. (d) Based on the transition rates of this HMM, an equivalent timescale τ_{HM} can be estimated which coincides with τ_{AC} for low h.

for the given plasticity rule and does not affect our main conclusions (see Supplemental Material for a milder plasticity rule).

More important is the observation that, as shown in Fig. 1(g), the mean coupling weights w^{rec} increase almost linearly with decreasing input rate. A fit of the form

$$\langle w^{rec}\rangle(h) = \alpha - \beta h, \tag{6}$$

where $\langle\cdot\rangle$ stands for the average across synaptic connections over either excitatory or inhibitory populations, yields $\alpha \approx 22.75$ and $\beta \approx 14.23$ for excitatory weights or $\alpha \approx 26.1$ and $\beta \approx 16.7$ for inhibitory weights. Hence a reduction in input rate clearly strengthens the recurrent connections in homeostatically regulated E-I networks consistent with the theoretical consideration that the loss of external input needs to be compensated by recurrent activity generation in order to maintain a constant firing rate [14].

In addition to supporting general theoretical arguments, our setup allows us to investigate how our homeostatic self-organization affects the interplay between excitatory and inhibitory neurons. In fact, it is quite surprising that the mean coupling weights for excitatory and inhibitory weights are so similar [Fig. 1(g)], i.e., $\langle w^{rec}_{inh}\rangle \approx \langle w^{rec}_{exc}\rangle$, given that each neuron receives four times more input from excitatory than from inhibitory neurons. Naively, this implies strong excitation dominance in contrast to the expected inhibition dominance required for asynchronous irregular activity [30,51] to reproduce experimental single-neuron statistics [52–55]. This outcome can, however, be explained by our symmetric plasticity rule that does not distinguish between excitatory and inhibitory synapses and thereby fosters solutions with $\langle w^{rec}_{inh}\rangle \approx \langle w^{rec}_{exc}\rangle$. For small networks with homogeneous weights (see Appendix D), the condition $w^{rec}_{inh} \approx w^{rec}_{exc}$ turns out to be in the vicinity of a phase transition between regular (high firing) and irregular (low firing) dynamics. Reducing h makes this transition more abrupt and closer to $w^{rec}_{inh} = w^{rec}_{exc}$, implying that homeostatic plasticity regulates E-I networks towards a regular-to-irregular transition when decreasing the external input rate.

B. Homeostatically regulated neuromorphic hardware with low external input generates large autocorrelation times through emergent bistability

Next, we verify the theoretical prediction [14] that a homeostatically regulated system exhibits an increased autocorrelation to compensate for a decreasing external input (Fig. 2). For this, we consider a network after homeostatic development with fixed weights and evaluate the autocorrelation function of the population firing rate $\nu(t)$ over an interval of $T = 100$ s. Indeed, the autocorrelation functions, $C(t')$, show increasingly long tails with decreasing input rate h [Fig. 2(a)]. Moreover, they are well described by exponential decays, $C(t') = e^{-t'/\tau_{AC}}$, with increasing autocorrelation times τ_{AC} for decreasing h [Fig. 2(a) inset]. While this general trend has been reported for much smaller neuromorphic systems before [25], the inset of Fig. 2(a) reveals the emergence of two distinct regimes. For $h > 0.8$ kHz, we find autocorrelation times to saturate with increasing h, suggesting that the uncorrelated Poisson input successfully decorrelates already weakly correlated activity, giving rise to an *input-driven regime*. In contrast, for $h < 0.8$ kHz, we find an apparent divergence of τ_{AC} with decreasing h, such that this regime is characterized by dominant recurrent activation compensating for the lacking input, which results in increasing autocorrelation times for decreasing h, giving rise to a *recurrent-driven regime*.

Surprisingly, we observe that the autocorrelations originate from a bistable population rate [Figs. 2(b)–2(d)]. Specifically, the distribution $P(\nu)$ changes from unimodal for higher input strengths to bimodal for lower input strengths [Fig. 2(b)]. The latter suggests that the population rate starts to alternate between two distinct states. Indeed, close inspection of the time evolution of $\nu(t)$ reveals that for decreasing input strength the population rate switches between a low-rate state and a high-rate state [Fig. 2(c)], resembling up and down states in cortical networks [40–43]. Such a switching behavior is reminiscent of a Markov jump process between states of high and low firing rates [56], specifically a two-state hidden Markov model (HMM) [57]. We fitted a two-state HMM to the stationary population rate (discretized in steps of Δt) and obtained a 2×2 Markov matrix. Since the rate is stationary, the first eigenvalue is 1, i.e., $\lambda_1 = 1$, and the second eigenvalue

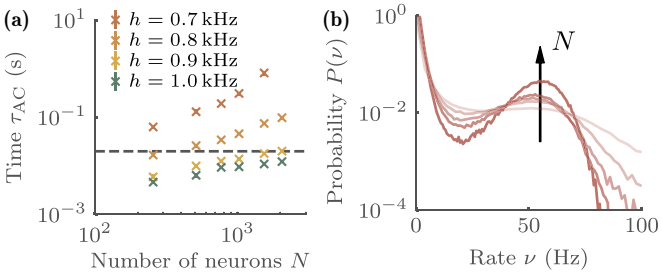

FIG. 3. Finite-size scaling of homeostatically regulated E-I networks with LIF neurons from computer simulations. (a) Autocorrelation time τ_{AC} as a function of system size N for different external input rates h. One can see a faster-than-power-law growth for $h < 0.8$ kHz, while τ_{AC} seems to saturate on the order of the dominant single-neuron timescale (dashed line) for $h > 0.8$ kHz. (b) Distributions of the population firing rate in windows of 5 ms for $h = 0.7$ kHz show that the bimodal shape remains for increasing N. The barrier in between high- and low-firing states grows with N.

λ_2 determines how quickly perturbations decay back to the stationary solution. This is related to the autocorrelation time as $\tau_{HM} = -\Delta t / \ln(\lambda_2)$. Indeed, the autocorrelation time of the HMM correlates with the autocorrelation time measured from the population rate for small input strengths, where the population rate becomes bistable [Fig. 2(d)]. This is fundamentally different from the *a priori* expected close-to-critical fluctuations [14], which would lead to scale-free avalanches [58] for small h that we do not observe (see Appendix E). We thus conclude that the emergent bistability is the underlying mechanism of the large autocorrelation times observed in the population dynamics of homeostatically regulated E-I networks of LIF neurons.

C. Computer simulations reveal an increasing dynamical barrier of emergent bistability with system size

Since our neuromorphic hardware only supports networks with up to 512 LIF neurons, we use computer simulations to verify the experimental results for increasing network sizes. In brief, we parametrize the simulations to match the experimental setup and use the BRIAN2 PYTHON package to solve the model (for details, see Sec. II). Indeed, we can reproduce the experimental results with our software implementation: We observe comparable bistable activity with similar autocorrelation functions (see Supplemental Material). However, while computer simulations in principle allow us to study any system size, they are much less efficient than the neuromorphic emulation. It is worth noting that for our application, i.e., homeostatically regulating a network of $N = 512$ LIF neurons for 6000 s, the computer simulation on an Intel Xeon E5-2630v4 (roughly 100 000 s at about 50 W) takes $O(10^4)$ more time and $O(10^7)$ more energy compared with the corresponding emulation on BrainScaleS-2 (about 6 s at a power budget of 100 mW).

Having established that the computer simulation reproduces the experimental results, we can study how the measured autocorrelation time τ_{AC} depends on the network size N [Fig. 3(a)]. Due to the large computational efforts, we focus on four representative input strengths: a low input strength ($h = 0.7$ kHz) where we observe bistable activity in the experiment, two medium input strengths ($h = 0.8$ kHz and $h = 0.9$ kHz) near the onset of bistability, and a high input strength ($h = 1.0$ kHz) where the network does not exhibit bistability. Only for $h = 0.7$ kHz do we observe an exponential increase in autocorrelation time with system size. Instead, at $h = 0.8$ kHz the autocorrelation time appears to grow as a power law, while for even larger values of h the τ_{AC} start to saturate on the order of the dominant single-neuron timescale (dashed line). Our numerical results further corroborate the classification into two distinct regimes: A recurrent-driven regime for low input strength with large emergent autocorrelations and an input-driven regime for high input strength with vanishing autocorrelations.

To further investigate the origin of the emergent autocorrelations, we study the shape of the probability distribution of local population rates $\nu(t)$ as a function of network size [Fig. 3(b)]. We observe that for low input strength, the bimodal distribution becomes more pronounced with increasing suppression of intermediate population-rate values. One can relate the suppression of intermediate rates to a *dynamical barrier* by interpreting the time course of the instantaneous population rates as a trajectory of the dynamical system in the potential $V(\nu) = -\ln P(\nu)$. This barrier would be analogous to the activation energy in an Arrhenius-type equation, i.e., $r \propto e^{-\Delta V/T}$, such that for a given level of fluctuation T the rate r to transition between low- and high-firing-rate regimes is lowered for increasing barriers ΔV. Since the height of this dynamical barrier increases with N, this explains the increasing autocorrelation time with system size.

D. Mean-field theory of emergent bistability from fluctuation-induced switching between metastable active and quiescent states

To qualitatively explain how bistability can emerge in a recurrent network with heterogeneous weights, we construct a simple mean-field theory based on the time evolution of a fraction of active neurons at a given time t, $\rho(t)$, which can be considered a proxy of the population rate $\nu(t)$ up to some factor. Let us consider a general mean-field ansatz

$$\dot{\rho}(t) = -\tau_{MF}\rho(t) + h(1-\rho(t)) + (1-\rho(t)) \times [\omega_1\rho(t) + \omega_2\rho^2(t) + \cdots], \quad (7)$$

where the first term describes the spontaneous decay of activity in the absence of inputs with some characteristic timescale τ_{MF}, the term proportional to h represents external input that can only activate inactive neurons [hence the $(1-\rho)$ factor], and the last term represents the gain function that describes recurrent activations, expanded in a power series of the activity. Here, the coefficients of expansion $(\omega_1, \omega_2, \ldots)$ are an effective representation of the full coupling matrix w_{ij}^{rec} (with ω_1 being proportional to the mean synaptic strength). The mean-field equation can be rewritten in a more compact form by grouping-up terms with different powers of the activity,

$$\dot{\rho}(t) = h - a\rho(t) - b\rho^2(t) + \cdots, \quad (8)$$

where $a = \tau_{MF} + h - \omega_1$ and $b = \omega_1 - \omega_2 > 0$ to ensure stability. It is important to notice that this mean-field

equation assumes infinitely large network sizes, $N \to \infty$, for which additional noise terms vanish.

To describe finite networks, one needs to introduce an additional stochastic term to the mean-field equation (8) that accounts for demographic fluctuations. Demographic fluctuations are characteristic of systems with an absorbing or quiescent state [59], where fluctuations of the total number of active units around some mean $N\rho(t)$ are expected to have a standard deviation that scales with $\sqrt{N\rho(t)}$ as a consequence of the central-limit theorem. For the fraction of active nodes in a finite network, we then obtain to leading order in system size

$$\dot{\rho}(t) = h - a\rho(t) - b\rho^2(t) + \sqrt{\rho(t)/N}\eta(t), \tag{9}$$

where $\eta(t)$ is Gaussian white noise with zero mean and variance σ^2. This (Ito-)Langevin equation can be expressed as a Fokker-Planck equation, with the steady-state solution [60] (see also Supplemental Material)

$$P(\rho) = \mathcal{N} \exp\left\{-\frac{2N}{\sigma^2}V(\rho)\right\}, \tag{10}$$

a normalization constant $\mathcal{N}$, and the potential

$$V(\rho) = \left(\frac{\sigma^2}{2N} - h\right)\ln\rho + a\rho + \frac{b}{2}\rho^2. \tag{11}$$

This potential $V(\rho)$ either can have a single (formally diverging) minimum at $\rho = 0$ (unimodal activity distribution) or can have two local minima (bistable activity distribution). The condition for extrema of the potential V implies that a bistable solution occurs when $a^2 - 4b(\frac{\sigma^2}{2N} - h) > 0$. With the additional conditions for a positive density, i.e., $\rho > 0$, as well as a positive slope at $\rho = 0$, i.e., $\rho^2\frac{d^2V}{d\rho^2}(0) = (\frac{\sigma^2}{2N} - h) > 0$, we expect to observe bistable dynamics for

$$a < -2\sqrt{b\left(\frac{\sigma^2}{2N} - h\right)} < 0. \tag{12}$$

To incorporate the effect of training recurrent weights with homeostatic regulation, we recall our empirically obtained anticorrelation, $\langle w\rangle = \alpha - \beta h$, upon homeostatic training [Fig. 1(g)]. In our mean-field theory, Eq. (8), we assume this to dominantly affect $a = \tau_{\mathrm{MF}} + h - \omega_1 \approx \tau_{\mathrm{MF}} - \alpha + h(1+\beta)$ and make the common assumption that $b = \omega_1 - \omega_2$ is constant up to higher-order effects. Inserting a into Eq. (12), we find that—for suitable parameters—lowering h can indeed induce a transition from a unimodal to a bimodal potential (Fig. 4).

The h-dependent transition from unimodal to bimodal can be visualized by numerically evaluating the mean-field model [Fig. 4(b)]. The numerical integration of Eq. (9) is straightforward [61] but needs special care to avoid running into the domain of negative numbers due to numerical imprecisions (see Appendix F). The resulting trajectories show typical demographic fluctuations for higher inputs and bistable activity for lower input. Since the involved parameters are not easily related in an explicit way to the experiment, this theoretical result is a qualitative explanation of the observed effect, and all parameters are in arbitrary units.

Our mean-field theory implies that emergent bistable population activity can be rationalized as a fluctuation-induced switching between a metastable active phase and a quiescent phase. For a system with an absorbing-to-active nonequilibrium phase transition for vanishing input, we find that finite-size fluctuations are responsible for a metastable active state (high rate) and external fluctuations lead to a metastable quiescent state (low rate). To transit from one state to another, the system needs to overcome a dynamical barrier, where the transition from high to low rate requires demographic noise, whereas the transition from low to high rate requires external noise.

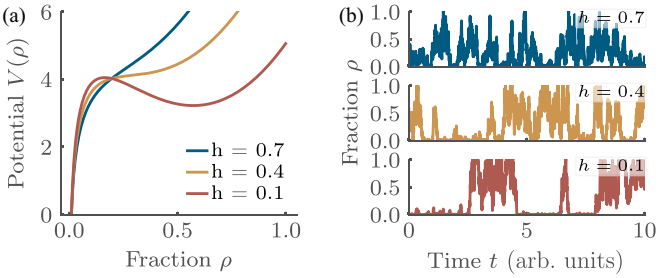

FIG. 4. Mean-field theory of emergent bistability upon reducing input to homeostatically regulated recurrent network. Our mean-field theory describes the temporal evolution of the fraction of active neurons, ρ, with metastable solutions given by the minimum of the potential, Eq. (11). (a) For suitable parameters ($\tau_{\mathrm{MF}} = 10$, $\alpha = 30$, $\beta = 15$, $b = 25$, $\sigma = 50$, $N = 512$), the potential exhibits a single minimum for large h but two minima for small h. (b) Numeric evaluation of the corresponding stochastic mean-field equation ($\Delta t = 10^{-7}$) shows fluctuating dynamics for large h and emergent bistability for low h.

IV. DISCUSSION

In summary, we showed that networks of E-I LIF neurons with homeostatic regulation during training can self-organize for low external input into a dynamical regime with stochastic switching between states of high population firing rate (up state) and states of low population firing rate (down state). Stochastic switching is the result of an emergent bistability, where a dynamical barrier between two metastable states (up and down) can be crossed due to fluctuations: finite-size activity fluctuations to cross from up to down and external-noise fluctuations to cross from down to up. The crossing rate decreases with the barrier height, similar to classical nucleation rates decreasing for a larger free-energy barrier [62–64], and we showed numerically that the barrier height increases with system size. Finally, a reduced crossing rate implies an increased autocorrelation time, which we demonstrated for both neuromorphic hardware and numerical simulations. Our findings of large emergent autocorrelation times in networks of spiking neurons complement recent observations in networks regulated by spike-timing-dependent plasticity [25] or trained to perform working-memory tasks [65].

Importantly, the stochastically switching population activity that we observe does not require an active adaptation mechanism: The emergent bistability was recorded in the testing phase after turning off homeostatic plasticity. While plasticity was necessary to shape the weight distribution, it is not relevant for the stochastic switching between metastable-

active (high rate) and metastable-absorbing (low rate) states. Our basic mechanism of a dynamical barrier that separates two metastable fixed points is consistent with previous observations of perturbation-induced state switching in spiking neural networks [66–68]. Here, we do not require additional, strong external perturbations, because we can control the height of the dynamical barrier and thereby observe the state switching induced by the available, weak external input during finite-time recordings. Importantly, the stochastic switching observed here is different from adaptation-based mechanisms, such as adaptation currents [69], or depletion of synaptic utility [70,71].

It is interesting to discuss more in depth the connection with the mechanism of self-organized bistability (SOB) [72,73], which has recently been shown to be relevant for collective brain dynamics [74]. This is a mechanism akin to self-organized criticality (SOC) [73,75], where the system self-organizes by means of a feedback loop between the level of activity (overall firing rate) and the control parameter (the mean synaptic weight) to the edge of a phase transition. In the case of SOB this is a discontinuous phase transition. In the present case, there is a similar feedback loop during training. However, rather than acting on a global control parameter, this feedback acts differentially for each synaptic weight, thereby generating a broad weight distribution, which, for low external input, tunes the system to a bistable state at the edge of the transition between high and low firing rates. Due to this bistability, in combination with external drive and finite-size fluctuations, we observe stochastic switching in the test phase (no homeostatic plasticity).

The here-identified mechanism of a stochastic state switching thereby presents an alternative perspective on so-called up and down states. Up and down states are defined on the level of a single-neuron membrane potential that switches between states with higher membrane potential, resulting in spiking responses, and those with lower membrane potential [40]. While some of the aforementioned models utilize adaptation mechanisms to generate up and down states [70,72,74], we here develop an alternative explanation: If neurons homeostatically regulate their firing rates, a decreasing external input can result in emergent autocorrelations with potential functional benefits [14,25] until a point where bistability can emerge on the population level. As a result of such emergent bistability, single neurons would switch between states of high and low synaptic input, which could in turn cause up and down states on the level of their membrane potentials. This explanation is consistent with experimental observations of up and down states in striatal neurons [40,41], which focused on single-cell bistability, and in cortical slices [42] and neuronal cultures [76], where up and down states were argued to be a collective (network) effect, and agrees with observations of bistability in networks of LIF neurons trained to store spatiotemporal patterns [77]. Our results thereby provide paths to test whether experimental observations of up and down states are connected to emergent bistability from homeostatic regulation.

Our simple mean-field theory implies that fluctuation-induced stochastic switching could be a very general effect in driven, finite systems with absorbing states. Examples of such systems include collective dynamics in epidemic spread [78], neural networks [58,79,80], ecosystems [81,82], and ultracold Rydberg atomic gases [83]; catalytic reactions on surfaces [84]; calcium dynamics in living cells [85]; or turbulence in liquid crystals [86,87] and active nematics [88]. Indeed, for some of these systems, stochastic switching has been observed, e.g., as switching behavior in disease models [89] or rate models of neural activity [90], for cellular automata with long-range interactions [91], for CO oxidation [92–94], or as phase separation in active matter [95,96]. Future work could include generalizing our results to systems that can be described by scalar fields, which is a common situation in nonequilibrium statistical physics, and investigating under what conditions one can observe (or avoid) stochastic switching on a macroscopic scale.

To ensure reproducibility, the code has been made freely available [97].

ACKNOWLEDGMENTS

This work has received funding from the European Union Sixth Framework Programme (FP6/2002-2006) under Grant Agreement No. 15879 (FACETS), the European Union Seventh Framework Programme (FP7/2007-2013) under Grant Agreements No. 604102 (HBP), No. 269921 (BrainScaleS), and No. 243914 (Brain-i-Nets), the Horizon 2020 Framework Programme (H2020/2014-2020) under Grant Agreements No. 720270, No. 785907, and No. 945539 (HBP), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2181/1-390900948 (the Heidelberg STRUCTURES Excellence Cluster), the Helmholtz Association Initiative and Networking Fund [Advanced Computing Architectures (ACA)] under Project No. SO-092, and the Manfred Stärk Foundation. M.A.M. acknowledges support from the Spanish Ministry and Agencia Estatal de Investigación (AEI) through Project I+D+i (Reference No. PID2020-113681GB-I00), financed by MICIN/AEI/10.13039/501100011033 and FEDER "A way to make Europe," as well as the Consejería de Conocimiento, Investigación Universidad, Junta de Andalucía, and European Regional Development Fund, Project No. P20-00173, for financial support. V.P. and J.Z. were supported by the Max Planck Society. J.Z. received financial support from the Joachim Herz Stiftung and the Plan Propio de Investigación y Transferencia de la Universidad de Granada. The authors acknowledge support from the state of Baden-Württemberg through bwHPC and the DFG through Grant No. INST 39/963-1 FUGG (bwForCluster NEMO).

APPENDIX A: HARDWARE DETAILS

The connectivity on BrainScaleS-2 is physically represented by two arrays of synapses, each with a set of 256×256 synapses. Input spikes s_i^j enter these arrays from the left via synapse drivers and are forwarded to a whole row of synapses. Each synapse within this row locally filters incoming events, weights them according to a 6-bit weight, and eventually transmits them to its home neuron. The neurons are arranged in an additional row below the array of synapses. Emitted neuronal spikes are injected back into the array via a flexible on-chip spike router.

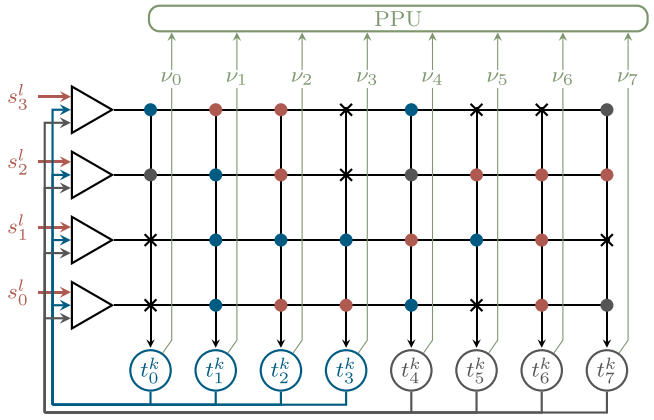

FIG. 5. The connectivity on BrainScaleS-2 is physically represented by two arrays of synapses. The routing capabilities of BrainScaleS-2 are utilized such that both arrays can be treated as a larger virtual one. Input events s_i^l enter this array together with recurrent events t_i^k (blue and gray) from the left via synapse drivers (triangles). The latter forward the events to a whole row of synapses (circles). Each synapse locally filters incoming events and transmits either input events (red) or recurrent events (blue and gray) to its home neuron. Sparsity is implemented by silencing out synapses (black crosses). Homeostatic regulation is carried out by the on-chip PPU by accessing neuronal firing rates ν_i to update synaptic weights in a row-wise parallel manner.

In this paper, we exploit the routing capabilities of the BrainScaleS-2 system to unify both synapse arrays to a virtual array of size 256×512 (Fig. 5). The event filtering within each synapse located between synapse driver i and neuron j is used to transmit either the input events of spike source i or the recurrent events of neuron i or $i + 256$, respectively. We map our networks by configuring a random set of on average K^{rec} synapses per column of synapses to transmit recurrent events. In addition, on average K^{in} randomly chosen synapses relay the input spike trains. All remaining synapses are configured to transmit no events.

On BrainScaleS-2, the effect of each synapse, i.e., excitatory or inhibitory, is determined by the synapse drivers and is therefore a row-wise property within the synapse array. For our networks, we program 20% of the synapse drivers (randomly selected) to be inhibitory.

The homeostatic plasticity is implemented on chip by drawing on both PPUs [5]. To that end, the number of emitted spikes is accessed and loaded into the single instruction multiple data (SIMD) vector units of the PPUs for subsequent weight update calculations. Each vector unit allows us to update a half row of synaptic weights (128 synapses) in parallel. Calculations are performed with a precision of 8 bits in a fractional-saturation arithmetic. The random numbers required to implement stochastic weight updates are directly drawn in parallel on chip via dedicated accelerators.

APPENDIX B: CALIBRATION AND PARAMETRIZATION

The fabrication-induced device variations of the analog neuromorphic substrate are mitigated by calibration routines. Here, we utilize bisection methods to adjust the neurosynaptic parameters inferred from recorded traces to desired targets (Fig. 6). Afterwards, the resulting LIF neuron parameters are measured, and their means as well as standard deviations are used for the parametrization of the equivalent software models. To align the impact of a single spike on all downstream neurons on hardware and in software, we characterize the postsynaptic potential (PSP) height as a function of the configured weight value w_{ij}. In more detail, we obtained the

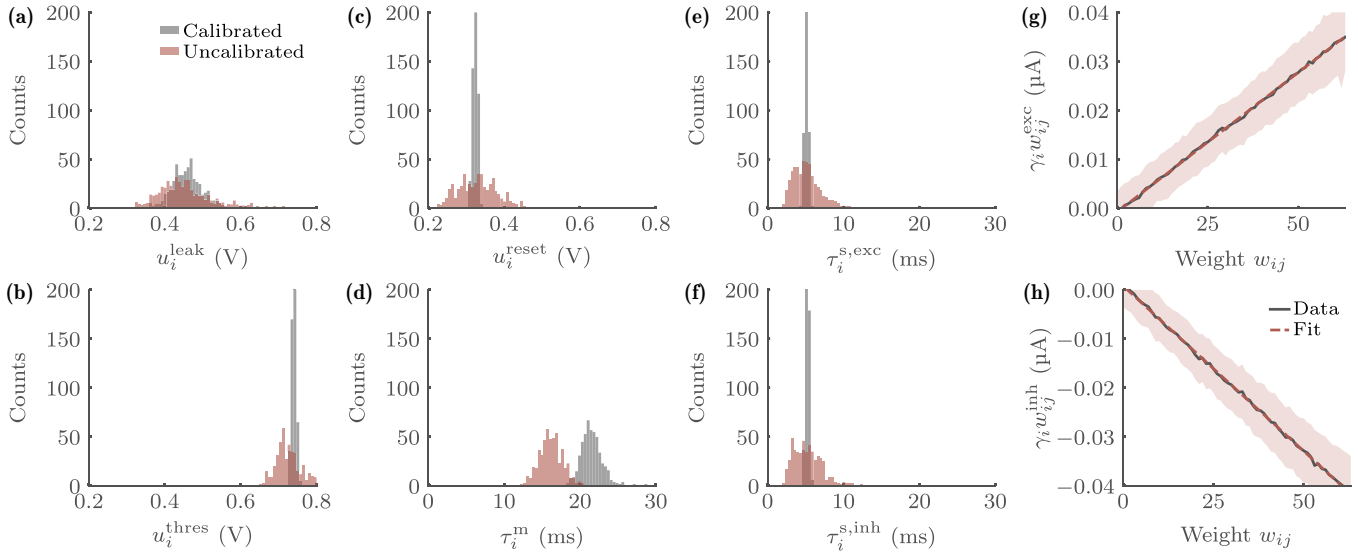

FIG. 6. Parameter distributions on the BrainScaleS-2 system. Calibration routines allow us to reduce the parameter spread between circuit instances by drawing on the configurability of BrainScaleS-2. The calibration targets for (a) the leak potential u_i^{leak} and (b) the threshold potential u_i^{thres} are chosen such that their distance is as high as possible. (c) The target for the reset potential is chosen to be slightly below u_i^{leak}. (d) The membrane time constants τ_i^{m} are calibrated to be larger than (e) the excitatory synaptic time constants $\tau_i^{\text{s,exc}}$ and (f) the inhibitory synaptic time constants $\tau_i^{\text{s,inh}}$. The latter are calibrated to coincide. Linear fits to measurements of the PSP height for various weight values w_{ij} allow us to estimate (g) the excitatory weight scaling factors γ_i^{exc} as well as (h) the inhibitory weight scaling factor γ_i^{inh}.

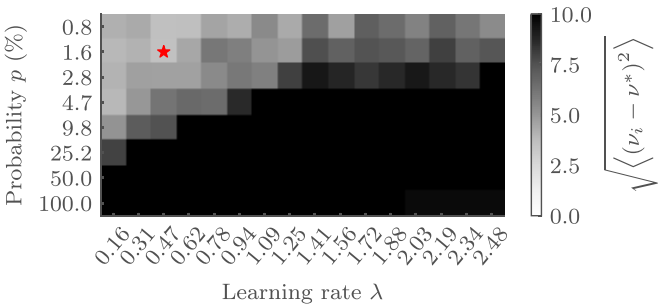

FIG. 7. Parametrization of the homeostatic regulation. The homeostatic regulation is parametrized by the update acceptance probability p as well as the learning rate λ. Shown is the variance of the firing rate of each neuron ν_i with respect to the target rate ν^*, $\sqrt{\langle(\nu_i - \nu^*)^2\rangle}$, averaged over 100 experiments for an input rate of $h = 0.6\,\mathrm{kHz}$. The configuration used within the experiments presented in the main text is highlighted by the red star.

weight scaling factor γ_i in Eq. (2) by fitting the ideal solution of Eq. (1)

$$u_i(t) = u^{\mathrm{leak}} + \frac{\tau^{\mathrm{m}} \cdot \tau^{\mathrm{s}} \cdot \gamma_i w_{ij}}{C^{\mathrm{m}}(\tau^{\mathrm{s}} - \tau^{\mathrm{m}})} \Theta\left(t - t_j^0\right) \times \left[\exp\left(-\frac{t - t_j^0}{\tau^{\mathrm{s}}}\right) - \exp\left(-\frac{t - t_j^0}{\tau^{\mathrm{m}}}\right)\right], \tag{B1}$$

to recordings of each neuron's membrane potential $u_i(t)$ in response to a single stimulating event t_j^0 relayed over a single synapse with weight w_{ij}. To ensure stable fits, we fix all fit parameters to the calibration target values except for the leak potential u^{leak} and our estimate of $\gamma_i w_{ij}$ that we call y. From the linear fit $y = \gamma_i w_{ij}$, we then obtain an estimate of γ_i for excitatory and inhibitory weights, respectively [Figs. 6(g) and 6(h)]. All estimated parameters are summarized in Table I.

APPENDIX C: PARAMETRIZATION OF THE HOMEOSTATIC REGULATION

The homeostatic regulation as given by Eq. (3) comes with two independent parameters: the learning rate λ as well as the update acceptance probability p. We obtained optimal parameters by performing a grid search for $h = 0.6\,\mathrm{kHz}$ and assessing the variance of the firing rate of all neurons ν_i with respect to the target rate ν^*, i.e., $\sqrt{\langle(\nu_i - \nu^*)^2\rangle}$. For a broad range of parameters, most of the LIF neurons emit spikes at a rate resembling the target rate (Fig. 7). Only for high values of λ and high values of p does the firing rate systematically deviate due to the integer arithmetic used for weight update calculations on the neuromorphic system.

Most notably, we also used the determined optimal parameters within our software simulations. This pursued strategy renders extensive parameter sweeps in software superfluous and moreover showcases the benefits of the accelerated analog emulation of neurosynaptic dynamics due to the referenced efficiency in terms of speed and power consumption.

APPENDIX D: PHASE DIAGRAMS OF NETWORKS WITH HOMOGENEOUS AND STATIC WEIGHTS

To understand why we can observe fluctuating or bistable dynamics in networks with homeostatically regulated weights despite apparent excitation dominance (cf. Figs. 1 and 2), we study here the phase diagram of comparable networks with homogeneous and static weights. Due to small fluctuations in the transition point for different realizations of small networks, we focus on a single network realization and split the measurement into 200 blocks of length 30 s. To ensure spiking activity even for low input strengths h, we initially increased h for 5 s and subsequently let the networks equilibrate for another 5 s.

We first perform a full sweep over the w_{exc}-w_{inh} plane on both a BrainScaleS-2 emulation and a corresponding software simulation for three exemplary input strengths [Figs. 8(a) and 8(b)]. While the overall trends of the firing rates in emulation and simulation are quite comparable, the transition from low to high firing rates is clearly shifted. We attribute these remaining differences to (i) the fact that our simulations do not capture correlations in the variability of parameters, but instead implement uncorrelated Gaussian noise, and (ii) additional saturation effects within the analog circuits of BrainScaleS-2.

When we consider as a proxy for the transition between high-firing phase and low-firing phase the line where $\nu \approx 10$, we notice that this transition occurs for $w^{\mathrm{exc}} \approx (w^{\mathrm{inh}} + o)/s$ with $o > 0$ being an h-dependent offset [Figs. 8(c) and 8(d)]. Hence this transition does not occur for a fixed inhibition-excitation ratio, $g = w^{\mathrm{inh}}/w^{\mathrm{exc}} \approx s - o/w^{\mathrm{exc}}$. Instead, g depends nontrivially on the excitatory coupling as well as on the parameters s and o, which further depend on the input rate h and the specific choice of input coupling (see Sec. II).

When we now interpret our symmetric homeostatic rule to only allow identical couplings $w_{\mathrm{exc}} = w_{\mathrm{inh}}$ [Figs. 8(c) and 8(d), black dashed line], then homeostatic plasticity should adjust the weights to the intersection between the transition lines and the unit line. While this is strongly simplified, it approximately recovers the range of resulting mean weights that we find for the homeostatically regulated neuromorphic chip (Fig. 1) and simulations (see Supplemental Material).

To characterize the dynamical phases of high and low firing rates, we focus on the special cut plane of $w^{\mathrm{exc}} = w^{\mathrm{in}} = 17$ on the BrainScaleS-2 system. We record the mean neuron firing rate, the integrated autocorrelation time, the network Fano factor, and the coefficient of variation (CV) of interspike intervals as a function of the inhibition dominance, $g = w^{\mathrm{inh}}/w^{\mathrm{exc}}$. The integrated autocorrelation time is estimated from integrating the autocorrelation function $C(t')$; cf. Eq. (5). We follow the common convention [98] to define $\tau_{\mathrm{int}} = \Delta t[\frac{1}{2} + \sum_{l=1}^{l_{\max}} C(l)]$, where $l_{\max}$ is self-consistently obtained as the first l for which $l > 6\,\tau_{\mathrm{int}}(l)$. This reliably estimates the scale of temporal correlations for fully sampled systems and did not become unstable due to the typical oscillations in the autocorrelation function observed for networks of LIF neurons. Since the communication bottleneck of the hardware constrained long samples for high firing rates, we partitioned each recording into $L = 200$ chunks of size $T = 30\,\mathrm{s}$ and estimated for each

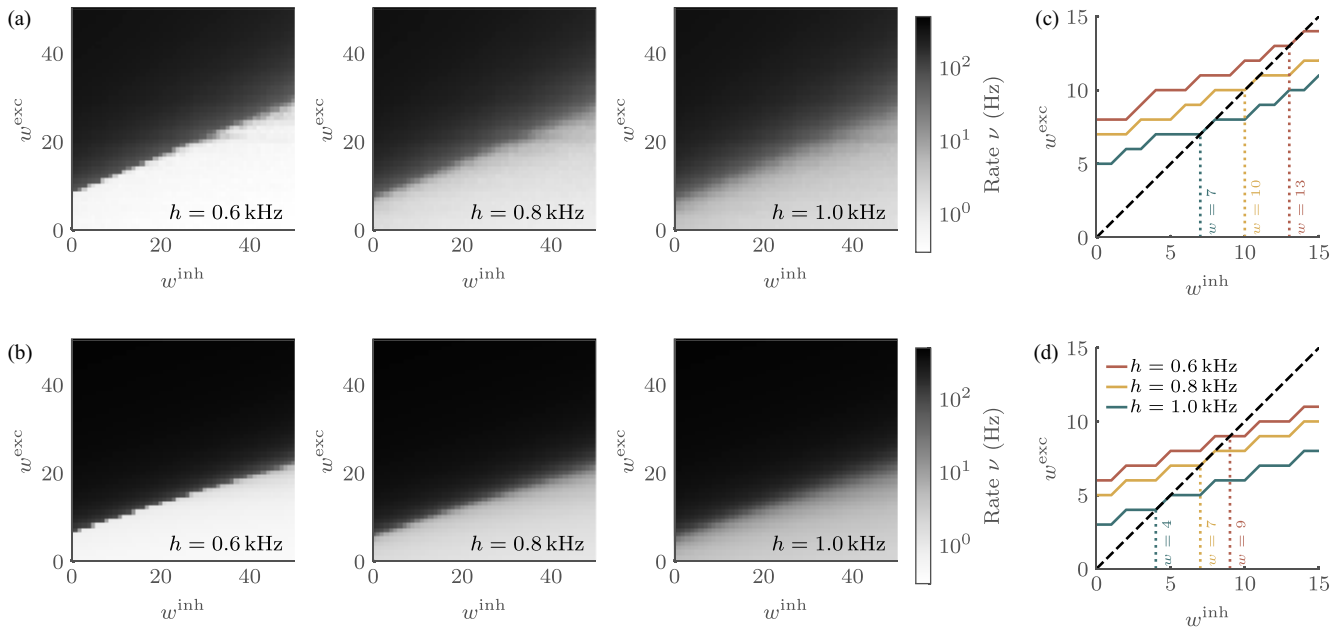

FIG. 8. Phase diagrams of networks with homogeneous and static weights. (a) The firing rates ν for three exemplary input rates h are comparable between hardware (a) and software (b) implementations. However, there is a small shift of the transitions from high firing rates ν to intermediate firing rates between both. For each value of the configured inhibitory weight w^{inh}, the firing rate is closest to 10 Hz for a coinciding excitatory weight $w^{\mathrm{exc}}_{\mathrm{cross}}$ for both emulation (c) and simulation (d).

chunk the moments as averages, i.e., $\overline{\nu(t)}_l = \frac{1}{T}\sum_t \nu(t)$ and $\overline{\nu(t)\nu(t+t')}_l = \frac{1}{T}\sum_t \nu(t)\nu(t+t')$. To avoid finite-data biases [99], we then first obtained the best estimates of the mean $\overline{\nu(t)} = \frac{1}{L}\sum_l \overline{\nu(t)}_l$ and analogously of the correlation term, to then estimate the covariance as $\mathrm{Cov}[\nu(t+t')\nu(t)] = \overline{\nu(t)\nu(t+t')} - \overline{\nu(t)}^2$. Similarly, we estimate the network Fano factor of the population rate as the ratio between variance and mean, i.e., $F = (\overline{\nu^2(t)} - \overline{\nu(t)}^2)/\overline{\nu(t)}$, and the CV as the average across neurons, i.e., $\mathrm{CV} = \frac{1}{N}\sum_i \sqrt{\overline{\delta t^2}_i - \overline{\delta t}_i^2}/\overline{\delta t}_i$ with interspike intervals δt_i^j of neuron i.

We find that for the considered setup with an E-I input layer, the transition from high firing rates to low firing rates is reminiscent of a regular-to-irregular transition [Fig. 9(a)]. For the special choice $w^{\mathrm{exc}} = w^{\mathrm{in}}$, the transition occurs at $g \approx 1$ for $h \to 0$, where the dynamic phase in the inhibition-dominated regime appears absorbing despite nonvanishing input due to the small system size. In the vicinity of the h-dependent transition, we observe peaks in the autocorrelation time [Fig. 9(b)], which we expect to vanish due to the absorbing state in the limit of $h \to 0$ and $N \to \infty$ [100]. We find that the network Fano factor, estimated from the population activity with $\Delta t = 5$ ms [Fig. 9(c)], is zero in the regular phase and low in the irregular phase, separated again by a peak that shifts towards $g = 1$ with decreasing h and becomes narrower. Last, we observe the average coefficient of variation of single-neuron interspike intervals to change from $\mathrm{CV} \approx 0$ for $g < 1$, indicating regular spiking, to $\mathrm{CV} \approx 1$ above the transition, indicating irregular spiking [Fig. 9(d)].

To illustrate the dynamic phases of regular and irregular activity, we show distributions of population rates [Fig. 9(e)] as well as spike raster plots and the time evolution of the population rate [Fig. 9(f)]. For $g = 17/17 = 1$ we find all h in a stable active state. For $g = 20/17 \approx 1.2$, $h = 0.3$ kHz is already in the quiescent state, while $h = 0.5$ kHz shows strong variance between high rate and low rates that hinder estimation of autocorrelation times, but all other h remain mostly in the stable active state. For $g = 26/17 \approx 1.5$, we observe the highest autocorrelations for $h = 0.7$ kHz due to strong fluctuation-driven excursions into the high-firing-rate regime. Further increasing g also causes the other h to fall into low-firing-rate states, where for small h the state appears absorbing with practically no population activity following upon the few external perturbations. Note that single points of the phase diagrams cannot be directly compared with the results after homeostatic regulation, which results in heterogeneous weight distributions [cf. Fig. 1(e)], because we here fix $w^{\mathrm{exc}} = w^{\mathrm{in}}$.

APPENDIX E: AVALANCHE ANALYSIS REVEALS

To verify that the autocorrelations we observe are not a result of close-to-critical fluctuations, we investigate the distribution of avalanche sizes. For (self-organized) critical systems, one would expect avalanche sizes s to be scale free [75,101], i.e., an avalanche-size distribution $p(s)$ that can be described by a power law.

Here, we follow the convention to estimate avalanches from a time-discrete firing rate $\nu(t)$ [58,102]. To constrain the temporal bin size to causal activity propagation, we estimate the spike delay from the solution of the LIF equation, Eq. (B1). More specifically, the peak of the excitatory postsynaptic potential (EPSP) is an estimate of the maximal time until a certain spike can causally induce a threshold crossing. We obtain the peak time of the EPSP from the condition $du/dt = 0$ in Eq. (B1), which together with the spike delay

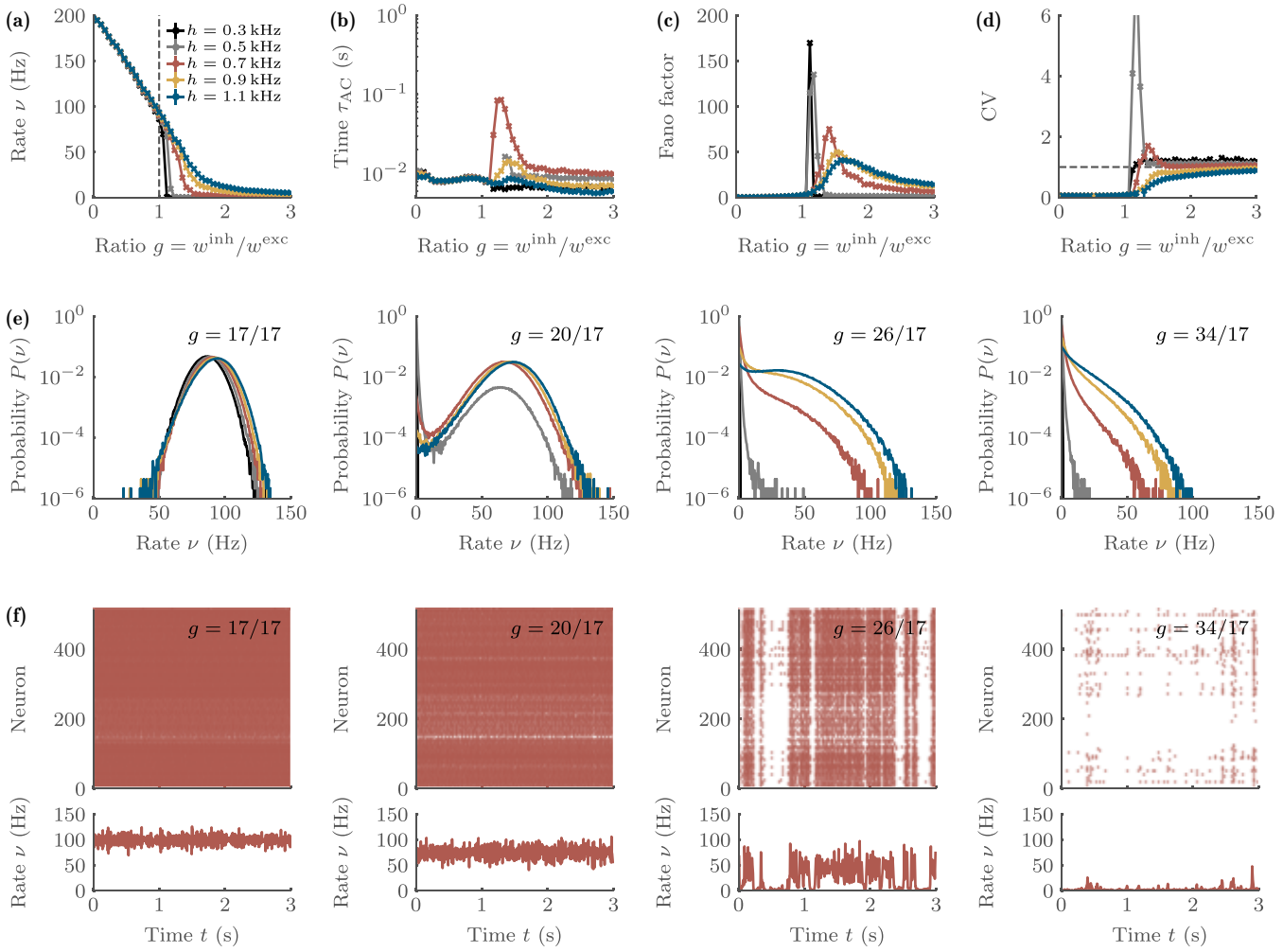

FIG. 9. Dynamics of networks with homogeneous and static weights. (a) The transitions from high firing rates ν to intermediate firing rates occurs in the vicinity of $g \approx 1$ for decreasing h. (b) The autocorrelation time τ_{AC} and (c) the Fano factor are estimated based on the population activity obtained with a bin size of 5 ms and show peaks around the finite-size transition. (d) The average coefficient of variation of single-neuron interspike intervals suggests regular spiking for small g and irregular spiking for large g. (e) Distributions of population rates shown in (a) for slices of different g show the transition from regular firing at high rates to irregular firing at low rates. (f) Example snapshots of spike raster plots and population rate for $h = 0.7$ kHz for different g.

yields

$$\tau^{tot} = \tau^{d} + \frac{\tau^{m}\tau^{s}\ln\left(\frac{\tau^{m}}{\tau^{s}}\right)}{\tau^{m} - \tau^{s}}. \tag{E1}$$

Since τ^{tot} sets an upper estimate of the causal delay, we here set the bin size for avalanche detection to $\Delta t = \tau^{tot}/2 = 5$ ms, in agreement with our previous time discretization. An avalanche is then defined as the number of spikes in consecutive nonempty bins in $\nu(t)$, for which we measured $L = 1000$ chunks of size $T = 100$ s.

The resulting avalanche-size distributions do not show power-law behavior as expected close to a nonequilibrium phase transition (Fig. 10, data points). In fact, we can compare the tails of the avalanche-size distribution with expectations from a two-state HMM (dashed curves). For this, we assume that for the HMM the state durations are exponentially distributed (which we confirmed for low h, not shown). Introducing a lifetime T_+ and a conditional average rate ν_+

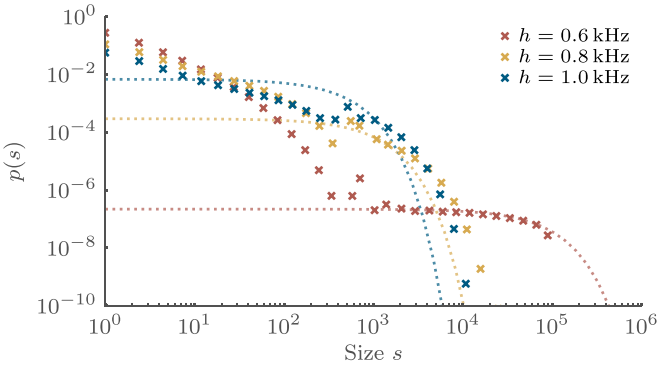

FIG. 10. The tail of the avalanche-size distribution from homeostatically regulated neuromorphic networks can be explained by the timescales of a Hidden Markov Model (HMM). Empirical avalanche-size distributions for different h in a log-binned representation do not show a power-law shape. In contrast, the cutoff scale of their tails coincides with estimates from the corresponding HMM.

for the high-firing state, we can approximate large avalanche sizes as $s = \nu_+ T_+$ with $p(s) \propto p(T_+ = s/\nu_+)$. Since we do not have an unbiased estimate of the fraction of avalanches in the low-firing rate, we constrain the amplitude to align the tails of corresponding distributions. The visual match between distribution tails indicates that in all cases we observe an exponential decay with a cutoff scale consistent with that of a HMM. We thus conclude that the empirical avalanche-size distributions show no sign of scale-free avalanches.

APPENDIX F: SIMULATION OF MEAN-FIELD MODEL

To simulate the time evolution of the mean-field model, one needs to take special care to avoid negative densities from numerical imprecisions that would render the multiplicative noise imaginary [61]. In short, the steps involve first evaluating an exact solution of the noise and linear terms and then an Euler integration of the remaining quadratic term. The precise mean-field equation we solve is

$$\dot{\rho}(t) = h - (\tau_{\mathrm{MF}} - \alpha + h[1+\beta])\rho(t) + \sigma\sqrt{\rho(t)/N}\eta(t) - b\rho^2(t). \tag{F1}$$

This equation is decomposed into the linear term plus noise (first line), for which one can obtain an analytical solution, and the higher-order term (second line), which can be trivially integrated.

For the square-root noise plus linear term, i.e., $\dot{\rho}(t) = h + a\rho + \tilde{\sigma}\sqrt{\rho}\eta$, starting from $\rho(t) = \rho_0$ one knows that the solution of the Fokker-Planck equation for time $t + \Delta t$ is [103]

$$P(\rho, t+\Delta t) = \lambda e^{\lambda(\rho_0\omega+\rho)}\left[\frac{\rho}{\rho_0\omega}\right]^{\mu/2} I_\mu(2\lambda\sqrt{\rho_0\rho\omega}), \tag{F2}$$

where I_μ is the modified Bessel function of order μ, $\omega = e^{a\Delta t}$, $\lambda = 2a/\sigma^2(\omega - 1)$, and $\mu = -1 + 2h/\sigma^2$. Using a Taylor-series expansion of the Bessel function, it was shown in Ref. [61] that rewriting $P(\rho, t+\Delta t)$ implies that the density after Δt can be simply drawn from the mixture

$$\rho^* = \mathrm{Gamma}[\mu + 1 + \mathrm{Poisson}[\lambda\rho_0\omega]]/\lambda. \tag{F3}$$

We thus evolve $\rho(t)$ in discrete time steps Δt in two steps [61]: First, we generate from $\rho_0 = \rho(t)$ the stochastic solution ρ^* using Eq. (F3). Second, we integrate the remaining term as $\rho(t+\Delta t) = \rho^*/(1 + \rho^* b\Delta t)$.

For the example we show in Fig. 4, we used $\Delta t = 10^{-7}$, $\tau = 10$, $\alpha = 19$, $\beta = 10$, $b = 12$, $\sigma = 40$, and $N = 512$. For further details we refer the reader to the available code [97].

[1] C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, and J. S. Plank, A survey of neuromorphic computing and neural networks in hardware, arXiv:1705.06963 [cs.NE].

[2] S. Furber, Large-scale neuromorphic computing systems, J. Neural Eng. **13**, 051001 (2016).

[3] J. Schemmel, D. Bruderle, K. Meier, and B. Ostendorf, Modeling synaptic plasticity within networks of highly accelerated I&F neurons, in *2007 IEEE International Symposium on Circuits and Systems ISCAS* (IEEE, Piscataway, NJ, 2007), pp. 3367–3370.

[4] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, A wafer-scale neuromorphic hardware system for large-scale neural modeling, in *2010 IEEE International Symposium on Circuits and Systems ISCAS* (IEEE, Piscataway, NJ, 2010), pp. 1947–1950.

[5] S. Friedmann, J. Schemmel, A. Grübl, A. Hartel, M. Hock, and K. Meier, Demonstrating hybrid learning in a flexible neuromorphic hardware system, IEEE Trans. Biomed. Circuits Syst. **11**, 128 (2017).

[6] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs), IEEE Trans. Biomed. Circuits Syst. **12**, 106 (2018).

[7] B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations, Proc. IEEE **102**, 699 (2014).

[8] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, Physics for neuromorphic computing, Nat. Rev. Phys. **2**, 499 (2020).

[9] C. Pehle, S. Billaudelle, B. Cramer, J. Kaiser, K. Schreiber, Y. Stradmann, J. Weis, A. Leibfried, E. Müller, and J. Schemmel, The BrainScaleS-2 accelerated neuromorphic system with hybrid plasticity, Front. Neurosci. **16**, 795876 (2022).

[10] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C.-K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng *et al.*, Loihi: A neuromorphic manycore processor with on-chip learning, IEEE Micro **38**, 82 (2018).

[11] U. Hasson, J. Chen, and C. J. Honey, Hierarchical process memory: Memory as an integral component of information processing, Trends Cogn. Sci. **19**, 304 (2015).

[12] L. Rudelt, D. G. Marx, M. Wibral, and V. Priesemann, Embedding optimization reveals long-lasting history dependence in neural spiking activity, PLoS Comput. Biol. **17**, e1008927 (2021).

[13] B. Cramer, Y. Stradmann, J. Schemmel, and F. Zenke, The Heidelberg spiking data sets for the systematic evaluation of spiking neural networks, IEEE Trans. Neural Netw. Learning Syst. **33**, 2744 (2022).

[14] J. Zierenberg, J. Wilting, and V. Priesemann, Homeostatic Plasticity and External Input Shape Neural Network Dynamics, Phys. Rev. X **8**, 031018 (2018).

[15] F. Y. K. Kossio, S. Goedeke, B. van den Akker, B. Ibarz, and R.-M. Memmesheimer, Growing Critical: Self-Organized Criticality in a Developing Neural System, Phys. Rev. Lett. **121**, 058301 (2018).

[16] Z. Ma, G. G. Turrigiano, R. Wessel, and K. B. Hengen, Cortical circuit dynamics are homeostatically tuned to criticality in vivo, Neuron **104**, 655 (2019).

[17] G. G. Turrigiano, K. R. Leslie, N. S. Desai, L. C. Rutherford, and S. B. Nelson, Activity-dependent scaling of quantal amplitude in neocortical neurons, Nature (London) **391**, 892 (1998).

[18] G. G. Turrigiano and S. B. Nelson, Homeostatic plasticity in the developing nervous system, Nat. Rev. Neurosci. **5**, 97 (2004).

[19] G. Turrigiano, Homeostatic synaptic plasticity: Local and global mechanisms for stabilizing neuronal function, Cold Spring Harb. Perspect. Biol. **4**, a005736 (2012).

[20] J. D. Murray, A. Bernacchia, D. J. Freedman, R. Romo, J. D. Wallis, X. Cai, C. Padoa-Schioppa, T. Pasternak, H. Seo, D. Lee, and X.-J. Wang, A hierarchy of intrinsic timescales across primate cortex, Nat. Neurosci. **17**, 1661 (2014).

[21] R. V. Raut, A. Z. Snyder, and M. E. Raichle, Hierarchical dynamics as a macroscopic organizing principle of the human brain, Proc. Natl. Acad. Sci. USA **117**, 20890 (2020).

[22] M. Spitmaan, H. Seo, D. Lee, and A. Soltani, Multiple timescales of neural dynamics and integration of task-relevant signals across cortex, Proc. Natl. Acad. Sci. USA **117**, 22522 (2020).

[23] R. Gao, R. L. van den Brink, T. Pfeffer, and B. Voytek, Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture, eLife **9**, e61277 (2020).

[24] J. H. Siegle, X. Jia, S. Durand, S. Gale, C. Bennett, N. Graddis, G. Heller, T. K. Ramirez, H. Choi, J. A. Luviano, P. A. Groblewski, R. Ahmed, A. Arkhipov, A. Bernard, Y. N. Billeh, D. Brown, M. A. Buice, N. Cain, S. Caldejon, L. Casal *et al.*, Survey of spiking in the mouse visual system reveals functional hierarchy, Nature (London) **592**, 86 (2021).

[25] B. Cramer, D. Stöckel, M. Kreft, M. Wibral, J. Schemmel, K. Meier, and V. Priesemann, Control of criticality and computation in spiking neuromorphic networks with plasticity, Nat. Commun. **11**, 2853 (2020).

[26] D. F. Wasmuht, E. Spaak, T. J. Buschman, E. K. Miller, and M. G. Stokes, Intrinsic neuronal dynamics predict distinct functional roles during working memory, Nat. Commun. **9**, 3499 (2018).

[27] S. E. Cavanagh, J. P. Towers, J. D. Wallis, L. T. Hunt, and S. W. Kennerley, Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex, Nat. Commun. **9**, 3498 (2018).

[28] J. Wilting, J. Dehning, J. Pinheiro Neto, L. Rudelt, M. Wibral, J. Zierenberg, and V. Priesemann, Operating in a reverberating regime Eenables rapid tuning of network states to task requirements, Front. Syst. Neurosci. **12**, 55 (2018).

[29] J. Wilting and V. Priesemann, Between perfectly critical and fully irregular: A reverberating model captures and predicts cortical spike propagation, Cereb. Cortex **29**, 2759 (2019).

[30] C. van Vreeswijk and H. Sompolinsky, Chaos in neuronal networks with balanced excitatory and inhibitory activity, Science **274**, 1724 (1996).

[31] A. Renart, J. de la Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris, The asynchronous state in cortical circuits, Science **327**, 587 (2010).

[32] O. Harish and D. Hansel, Asynchronous rate chaos in spiking neuronal circuits, PLoS Comput. Biol. **11**, e1004266 (2015).

[33] R. Rosenbaum, M. A. Smith, A. Kohn, J. E. Rubin, and B. Doiron, The spatial structure of correlated neuronal variability, Nat. Neurosci. **20**, 107 (2017).

[34] F. Mastrogiuseppe and S. Ostojic, Intrinsically-generated fluctuating activity in excitatory-inhibitory networks, PLoS Comput. Biol. **13**, e1005498 (2017).

[35] C. Baker, C. Ebsch, I. Lampl, and R. Rosenbaum, Correlated states in balanced neuronal networks, Phys. Rev. E **99**, 052414 (2019).

[36] P. E. Latham, Correlations demystified, Nat. Neurosci. **20**, 6 (2017).

[37] D. Dahmen, M. Layer, L. Deutz, P. A. Dąbrowska, N. Voges, M. von Papen, T. Brochier, A. Riehle, M. Diesmann, S. Grün, and M. Helias, Global organization of neuronal activity only requires unstructured local connectivity, eLife **11**, e68422 (2022).

[38] S. Ostojic, Two types of asynchronous activity in networks of excitatory and inhibitory spiking neurons, Nat. Neurosci. **17**, 594 (2014).

[39] A. van Meegen and S. J. van Albada, Microscopic theory of intrinsic timescales in spiking neural networks, Phys. Rev. Res. **3**, 043077 (2021).

[40] C. Wilson, Up and down states, Scholarpedia **3**, 1410 (2008).

[41] E. A. Stern, A. E. Kincaid, and C. J. Wilson, Spontaneous subthreshold membrane potential fluctuations and action potential variability of rat corticostriatal and striatal neurons in vivo, J. Neurophysiol. **77**, 1697 (1997).

[42] R. Cossart, D. Aronov, and R. Yuste, Attractor dynamics of network UP states in the neocortex, Nature (London) **423**, 283 (2003).

[43] J. Hidalgo, L. F. Seoane, J. M. Cortés, and M. A. Muñoz, Stochastic amplification of fluctuations in cortical up-states, PLoS One **7**, e40710 (2012).

[44] J. Schemmel, S. Billaudelle, P. Dauer, and J. Weis, Accelerated analog neuromorphic computing, in *Analog Circuits for Machine Learning, Current/Voltage/Temperature Sensors, and High-Speed Communication: Advances in Analog Circuit Design 2021*, edited by P. Harpe, K. A. Makinwa, and A. Baschirotto (Springer International, Cham, Switzerland, 2022), pp. 83–102.

[45] J. Schemmel, S. Billaudelle, P. Dauer, and J. Weis, Accelerated analog neuromorphic computing, in *Analog Circuits for Machine Learning, Current/Voltage/Temperature Sensors, and High-speed Communication*, edited by P. Harpe, K. A. Makinwa, and A. Baschirotto (Springer, Cham, 2022).

[46] E. Müller, C. Mauch, P. Spilger, O. J. Breitwieser, J. Klähn, D. Stöckel, T. Wunderlich, and J. Schemmel, Extending BrainScaleS OS for BrainScaleS-2, arXiv:2003.13750.

[47] E. Müller, E. Arnold, O. Breitwieser, M. Czierlinski, A. Emmel, J. Kaiser, C. Mauch, S. Schmitt, P. Spilger, R. Stock, Y. Stradmann, J. Weis, A. Baumbach, S. Billaudelle, B. Cramer, F. Ebert, J. Göltz, J. Ilmberger, V. Karasenko, M. Kleider *et al.*, A scalable approach to modeling on accelerated neuromorphic hardware, Front. Neurosci. **16** (2022).

[48] R. J. Douglas and K. A. C. Martin, Inhibition in cortical circuits, Curr. Biol. **19**, R398 (2009).

[49] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.5.033035 for details on the validation of the analog emulation; a comparison to a different homeostatic update rule; and a detailed calculation of the mean-field solution.

[50] D. Goodman and R. Brette, The Brian simulator, Front. Neurosci. **3**, 192 (2009).

[51] N. Brunel, Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons, J. Comput. Neurosci. **8**, 183 (2000).

[52] B. D. Burns and A. C. Webb, The spontaneous activity of neurones in the cat's cerebral cortex, Proc. R. Soc. London, Ser. B **194**, 211 (1976).

[53] W. R. Softky and C. Koch, The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs, J. Neurosci. **13**, 334 (1993).

[54] C. F. Stevens and A. M. Zador, Input synchrony and the irregular firing of cortical neurons, Nat. Neurosci. **1**, 210 (1998).

[55] R. B. Stein, E. R. Gossen, and K. E. Jones, Neuronal variability: Noise or part of the signal? Nat. Rev. Neurosci. **6**, 389 (2005).

[56] O. C. Ibe, Introduction to Markov processes, in *Markov Processes for Stochastic Modeling*, 2nd ed., edited by O. C. Ibe (Elsevier, Oxford, 2013), Chap. 3, pp. 49–57.

[57] L. Rabiner and B. Juang, An introduction to hidden Markov models, IEEE ASSP Mag. **3**, 4 (1986).

[58] J. M. Beggs and D. Plenz, Neuronal avalanches in neocortical circuits, J. Neurosci. **23**, 11167 (2003).

[59] M. Henkel, H. Hinrichsen, and S. Lübeck, *Absorbing Phase Transitions* (Springer, Dordrecht, 2008).

[60] M. A. Muñoz, Nature of different types of absorbing states, Phys. Rev. E **57**, 1377 (1998).

[61] I. Dornic, H. Chaté, and M. A. Muñoz, Integration of Langevin Equations with Multiplicative Noise and the Viability of Field Theories for Absorbing Phase Transitions, Phys. Rev. Lett. **94**, 100601 (2005).

[62] J. Feder, K. Russell, J. Lothe, and G. Pound, Homogeneous nucleation and growth of droplets in vapours, Adv. Phys. **15**, 111 (1966).

[63] D. Kashchiev, *Nucleation* (Elsevier, New York, 2000).

[64] J. Zierenberg, P. Schierz, and W. Janke, Canonical free-energy barrier of particle and polymer cluster formation, Nat. Commun. **8**, 14546 (2017).

[65] R. Kim and T. J. Sejnowski, Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks, Nat. Neurosci. **24**, 129 (2021).

[66] N. Brunel and X.-J. Wang, Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition, J. Comput. Neurosci. **11**, 63 (2001).

[67] A. Renart, R. Moreno-Bote, X.-J. Wang, and N. Parga, Mean-driven and fluctuation-driven persistent activity in recurrent networks, Neural Comput. **19**, 1 (2007).

[68] E. M. Tartaglia and N. Brunel, Bistability and up/down state alternations in inhibition-dominated randomly connected networks of LIF neurons, Sci. Rep. **7**, 11916 (2017).

[69] N. Parga and L. Abbott, Network model of spontaneous activity exhibiting synchronous transitions between up and down states, Front Neurosci. **1**, 57 (2007).

[70] D. Millman, S. Mihalas, A. Kirkwood, and E. Niebur, Self-organized criticality occurs in non-conservative neuronal networks during 'up' states, Nat. Phys. **6**, 801 (2010).

[71] J. A. Bonachela, S. de Franciscis, J. J. Torres, and M. A. Muñoz, Self-organization without conservation: Are neuronal avalanches generically critical? J. Stat. Mech. (2010) P02015.

[72] S. di Santo, R. Burioni, A. Vezzani, and M. A. Muñoz, Self-Organized Bistability Associated with First-Order Phase Transitions, Phys. Rev. Lett. **116**, 240601 (2016).

[73] V. Buendía, S. di Santo, J. A. Bonachela, and M. A. Muñoz, Feedback mechanisms for self-organization to the edge of a phase transition, Front. Phys. **8**, 333 (2020).

[74] V. Buendía, S. di Santo, P. Villegas, R. Burioni, and M. A. Muñoz, Self-organized bistability and its possible relevance for brain dynamics, Phys. Rev. Res. **2**, 013318 (2020).

[75] P. Bak, C. Tang, and K. Wiesenfeld, Self-organized criticality, Phys. Rev. A **38**, 364 (1988).

[76] R. Vardi, A. Goldental, S. Sardi, A. Sheinin, and I. Kanter, Simultaneous multi-patch-clamp and extracellular-array recordings: Single neuron reflects network activity, Sci. Rep. **6**, 36228 (2016).

[77] S. Scarpetta, I. Apicella, L. Minati, and A. de Candia, Hysteresis, neural avalanches, and critical behavior near a first-order transition of a spiking neural network, Phys. Rev. E **97**, 062305 (2018).

[78] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. **87**, 925 (2015).

[79] D. R. Chialvo, Emergent complex neural dynamics, Nat. Phys. **6**, 744 (2010).

[80] J. Wilting and V. Priesemann, 25 years of criticality in neuroscience—established results, open controversies, novel concepts, Curr. Opin. Neurobiol.Comput. Neurosci.**58**, 105 (2019).

[81] M. Scheffer, *Critical Transitions in Nature and Society* (Princeton University Press, Princeton, NJ, 2009).

[82] P. V. Martín, J. A. Bonachela, S. A. Levin, and M. A. Muñoz, Eluding catastrophic shifts, Proc. Natl. Acad. Sci. USA **112**, E1828 (2015).

[83] S. Helmrich, A. Arias, G. Lochead, T. M. Wintermantel, M. Buchhold, S. Diehl, and S. Whitlock, Signatures of self-organized criticality in an ultracold atomic gas, Nature (London) **577**, 481 (2020).

[84] M. Ehsasi, M. Matloch, O. Frank, J. H. Block, K. Christmann, F. S. Rys, and W. Hirschwald, Steady and nonsteady rates of reaction in a heterogeneously catalyzed reaction: Oxidation of CO on platinum, experiments and simulations, J. Chem. Phys. **91**, 4949 (1989).

[85] M. Bär, M. Falcke, H. Levine, and L. S. Tsimring, Discrete Stochastic Modeling of Calcium Channel Dynamics, Phys. Rev. Lett. **84**, 5664 (2000).

[86] K. A. Takeuchi, M. Kuroda, H. Chaté, and M. Sano, Directed Percolation Criticality in Turbulent Liquid Crystals, Phys. Rev. Lett. **99**, 234503 (2007).

[87] K. A. Takeuchi, M. Kuroda, H. Chaté, and M. Sano, Experimental realization of directed percolation criticality in turbulent liquid crystals, Phys. Rev. E **80**, 051116 (2009).

[88] A. Doostmohammadi, T. N. Shendruk, K. Thijssen, and J. M. Yeomans, Onset of meso-scale turbulence in active nematics, Nat. Commun. **8**, 15326 (2017).

[89] L. Böttcher, J. Nagler, and H. J. Herrmann, Critical Behaviors in Contagion Dynamics, Phys. Rev. Lett. **118**, 088301 (2017).

[90] A. van Meegen, T. Kühn, and M. Helias, Large-Deviation Approach to Random Recurrent Neuronal Networks: Parameter Inference and Fluctuation-Induced Transitions, Phys. Rev. Lett. **127**, 158302 (2021).

[91] A. Pizzi, A. Nunnenkamp, and J. Knolle, Bistability and time crystals in long-ranged directed percolation, Nat. Commun. **12**, 1061 (2021).

[92] G. Ertl, Oscillatory kinetics and spatio-temporal self-organization in reactions at solid surfaces, Science **254**, 1750 (1991).

[93] Y. Suchorski, S. M. Kozlov, I. Bespalov, M. Datler, D. Vogel, Z. Budinska, K. M. Neyman, and G. Rupprechter, The role of metal/oxide interfaces for long-range metal particle activation during CO oxidation, Nat. Mater. **17**, 519 (2018).

[94] H. Wang, T. Shen, S. Duan, Z. Chen, and X. Xu, Bistability for CO oxidation: An understanding from extended phenomenological kinetics simulations, ACS Catal. **9**, 11116 (2019).

[95] D. Martin, H. Chaté, C. Nardini, A. Solon, J. Tailleur, and F. Van Wijland, Fluctuation-Induced Phase Separation in Metric and Topological Models of Collective Motion, Phys. Rev. Lett. **126**, 148001 (2021).

[96] L. Di Carlo and M. Scandolo, Evidence of fluctuation-induced first-order phase transition in active matter, New J. Phys. **24**, 123032 (2022).

[97] https://github.com/Priesemann-Group/neuromorphic-bistability.

[98] W. Janke, Statistical analysis of simulations: Data correlations and error estimation, in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms: Winter School, 25 February - 1 March 2002, Rolduc Conference Centre, Kerkrade, The Netherlands; Lecture Notes*, NIC Series No. 10, edited by J. Grotendorst (NIC, Jülich, 2002), pp. 423–445.

[99] F. H. C. Marriott and J. A. Pope, Bias in the estimation of autocorrelations, Biometrika **41**, 390 (1954).

[100] J. Zierenberg, V. Buendía, B. Cramer, V. Priesemann, and M. A. Muñoz (unpublished).

[101] G. Pruessner, *Self-Organised Criticality: Theory, Models and Characterisation* (Cambridge University Press, Cambridge, 2012).

[102] R. Zeraati, V. Priesemann, and A. Levina, Self-organization toward criticality by synaptic plasticity, Front. Phys. **9**, 619661 (2021).

[103] W. Feller, Two singular diffusion problems, Ann. Math. **54**, 173 (1951).