

Deep learning as closure for irreversible processes: A data-driven generalized Langevin equation

Antonio Russo,¹ Miguel A. Durán-Olivencia,¹

Ioannis G. Kevrekidis,² and Serafim Kalliadasis¹

¹*Department of Chemical Engineering,*

Imperial College London, London SW7 2AZ, UK

²*Department of Applied Mathematics and Statistics,*

Johns Hopkins University Baltimore, Maryland 21218, USA

(Dated: March 25, 2019)

Abstract

The ultimate goal of physics is finding a unique equation capable of describing the evolution of any observable quantity in a self-consistent way. Within the field of statistical physics, such an equation is known as the generalized Langevin equation (GLE). Nevertheless, the formal and exact GLE is not particularly useful, since it depends on the complete history of the observable at hand, and on hidden degrees of freedom typically inaccessible from a theoretical point of view. In this work, we propose the use of deep neural networks as a new avenue for learning the intricacies of the unknowns mentioned above. By using machine learning to eliminate the unknowns from GLEs, our methodology outperforms previous approaches (in terms of efficiency and robustness) where general fitting functions were postulated. Finally, our work is tested against several prototypical examples, from a colloidal systems and particle chains immersed in a thermal bath, to climatology and financial models. In all cases, our methodology exhibits an excellent agreement with the actual dynamics of the observables under consideration.

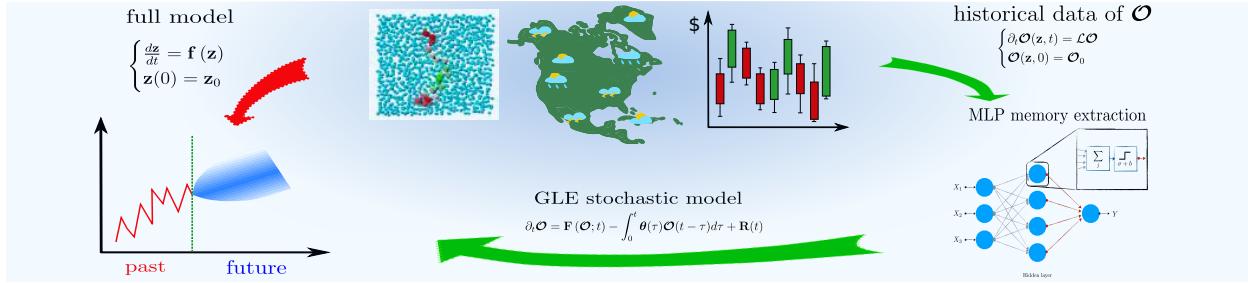


FIG. 1. Two possible approaches to simulate the time evolution of a dynamical system observable. A first one consists in solving the full deterministic dynamical system (red arrow). Despite the advantage of being exact, this approach is often not suitable either because too computationally expensive or because a model of the full system is not accessible. An alternative approach consists in building a stochastic GLE model for an observable of the dynamical system and parametrizing it through a MLP, given proper historical data (green arrows).

The mathematical description of any natural process requires a governing equation to determine the time evolution of a given set of quantities, which represent the mathematical abstraction of a given set of properties. Such quantities are known as observables. The set of observables which uniquely describe the macroscopic state of a system are typically termed as canonical observables, e.g. pressure and temperature if we refer to the thermodynamic state, or momentum and energy if we talk about the dynamical state. On the other hand, the minimum set of observables required to describe the microscopic state of a dynamical system is referred to as degrees of freedom (DoF). Statistical physics deals with the connection between macroscopic observables and DoF. Within this context, any given macro observable can be understood as a function of the system's DoF. Given the huge number of DoF a physical system typically involves, finding the exact functional form which connects the given observable and the system's DoF represents an overwhelming challenge. Going from the DoF description of the system to the observable description entails a dimensionality reduction, which allows us to describe the same phenomenon with much lesser number of variables. Such a reduction not only would provide us with the convenience of a simpler representation of the same problem, but also would give us a computationally cheaper way of describing the same phenomenon. And this is of fundamental importance, given the huge computational cost required to integrate the DoF over time in realistic scenarios, since the number of DoF is typically as large as Avogadro's number ($N_A \sim 10^{23}$). A practical way of seeing this

is by comparing the micro- and macroscopic descriptions of a sea wave. The microscopic description would require us to integrate over time all the water particles' positions and velocities over time, while the same phenomenon can be described with the simplest wave equation [1, 2]. Unfortunately, postulating a dynamical law for a given observable dynamics is not as simple as in the case of a wave (neither so regular). This forces us to try and figure out the connection between the DoF and the observable dynamics by using a rigorous approach.

Fortunately, however, there exists a mathematical formalism which permits us to find the formal structure of the observable dynamics by starting with the DoF deterministic dynamics, without having to know exactly the functional dependence of the observable on the DoF. Among other names, it is known as the projection-operator technique (POT) [3–6]. Despite not yielding a closed governing equation (given the limitation of not knowing exactly the functional dependence of the observable on the DoF), in some cases it produces a successful model after applying convenient simplifications. The first success of the POTs goes back to their introduction by Mori and Zwanzig to formally derive the dynamics of a Brownian particle, previously described only phenomenologically by Langevin. Once a projection operator \mathcal{P} is defined, Mori-Zwanzig formalism allows to derive the dynamic evolution of the observables in the form of GLEs. GLEs have a stochastic form that includes a first term accounting for the Markovian contribution, a second one constituting the memory of the system (non-Markovian term) and a last one, usually interpreted as noise. The non-Markovian term is in the form of a time convolution involving, in general, complex functions of the original systems. However, in many relevant cases[7–10], the memory term can be expressed simply as the convolution between the observable and a tensor function $\boldsymbol{\theta}(t)$, known as memory kernel. Thus, given a microscopic dynamical system described by a vector of DoF $\mathbf{z} \in \mathbb{R}^n$ with time evolution $\partial_t \mathbf{z} = \mathbf{f}(\mathbf{z})$, the corresponding GLEs has the following form:

$$\partial_t \mathcal{O}(t) = \mathcal{P}\mathcal{L}\mathcal{O} - \int_0^t \boldsymbol{\theta}(\tau)\mathcal{O}(t-\tau)d\tau + \mathbf{R}(t). \quad (1)$$

The vector $\mathbf{R}(t)$, being orthogonal to \mathcal{O} , is interpreted as a stochastic term, with correlation given by the fluctuation dissipation theorem $\langle \mathbf{R}(t), \mathbf{R}(t') \rangle = \boldsymbol{\theta}(t-t')\langle \mathcal{O}, \mathcal{O} \rangle$, where the notation $\langle \mathbf{A}, \mathbf{B} \rangle$ indicates the inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \int \rho(\mathbf{z})\mathbf{A}(\mathbf{z})\mathbf{B}^*(\mathbf{z}) d\mathbf{z}$, with $\rho(\mathbf{z})$ being a normalized pdf defined in the phase space of the original system, and \mathbf{B}^* the conjugate

transpose of \mathbf{B} . The term \mathcal{PLO} depends only on the current system configuration and, in some cases, corresponds to the mean force term [7–10] (see Appendix for more details). Several approaches have been developed to compute the potential of mean force of a system, including adaptive biasing forces [11] and umbrella sampling [12]. The non-Markovian term depends on the previous evolution of the system and is characterized by the memory kernel function $\theta(t)$, which also unequivocally determines the characteristics of the noise term $\mathbf{R}(t)$ through the fluctuation dissipation theorem. As a consequence, a proper approximation of $\theta(t)$ is required to preserve the main features of the original high dimensional system into the reduced one. However, the memory kernel depends on both the full set of DoF and the whole history of the complex system, hence making the problem often intractable.

In previous studies, several approaches have been proposed to parametrize GLEs. Analytical forms can be only obtained for specific systems, such as a particle in a harmonic oscillator heat bath[13], while numerical techniques are necessary for more complex systems characterized by non-linear interactions. For instance, in Ref. [14] the authors adopt a perturbation scheme, which is yet “too complex for general use”. Despite its accuracy, the algorithm developed in Ref. [15] to parametrize GLEs involves sampling of the full original system, thus, becoming computationally prohibitive for large systems. Another procedure involving large matrix computations and Krylov sub-space approximations is shown in Ref. [10]. In Ref. [16], an iterative approach is used to compute a discrete approximation of $\theta(t)$ from the system autocorrelation functions. In both Refs [17, 18], the researchers propose to extract the memory kernel by Laplace transforming the correlation functions computed from some historical data of the observables. However, this strategy exhibits serious limitations when the available data on the observables are affected by fluctuations, as shown in what follows.

In this work, we present a novel data-driven approach, which makes use of a multilayer perceptron (MLP) to reach an optimal parameterization of the memory kernel. The MLP is provided with proper historical data of the observable of interest obtained either from simulations or databases. Hence, the MLP algorithm executes an optimization procedure to evaluate an approximation of the memory kernel with a degree of accuracy depending on the number of neurons in the hidden layer. Compared to previous approaches, our approximation through MLP shows enhanced accuracy and robustness, especially when the available data are limited or affected by significant fluctuations. In the presented procedure, the memory kernel is extracted in the form of a multi-exponential functions, thus enabling us to derive

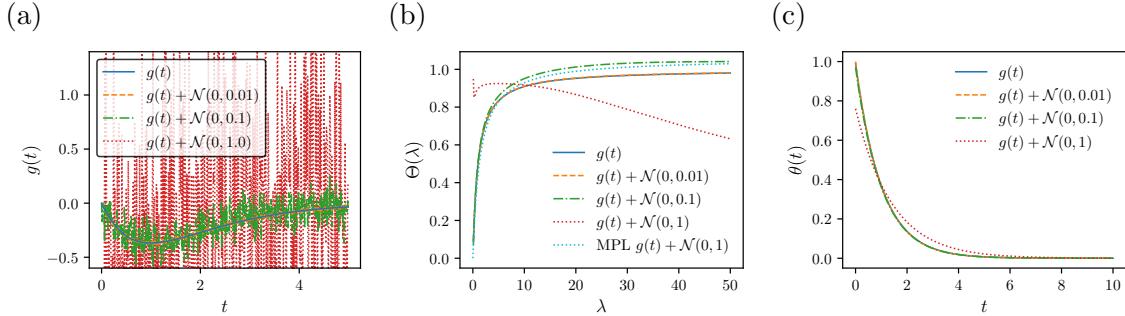


FIG. 2. Convolution function $g(t)$ affected by random noise with varying amplitudes (a). Comparison between the memory kernel θ computed in the Laplace space (b) and with our MLP-based method (c), for $g(t)$ affected by random noises. For comparison purpose, in (b) we also report the Laplace transform of the memory kernel obtain with our MLP for the strongest noise.

a tractable stochastic integration algorithm of the non-Markovian process characterized by a time-correlated noise. The universal approximation theorem[19, 20] guarantees a wide applicability of the presented methodology. In fact, we test the method in some relevant case studies from chemistry, biology, climatology and finance.

I. THEORY AND METHOD

A. Correlations equation

Let us consider a system in equilibrium condition, such that the historical data of some observables of the system can be considered a realization of a stationary process. If we take the aforementioned inner product of the GLE with $\mathcal{O}(0)$, one obtains the correlation equation:

$$\mathbf{g}(t) = - \int_0^t \boldsymbol{\theta}(t-\tau) \mathbf{h}(\tau) d\tau, \quad (2)$$

where we introduced the matrices $\mathbf{g}(t) = \langle \partial_t \mathcal{O}(t) - \mathcal{P}\mathcal{L}\mathcal{O}(t), \mathcal{O}(0) \rangle$ and $\mathbf{h}(t) = \langle \mathcal{O}(t), \mathcal{O}(0) \rangle$, and we took advantage of the orthogonality between the random force and the initial value of the observable to set $\langle \mathbf{R}(t), \mathcal{O}(0) \rangle \sim \mathbf{0}$ (see Appendix). Since the inner product $\langle \mathbf{A}, \mathbf{B} \rangle$ corresponds to the ensemble average of the matrix product \mathbf{AB}^* , for ergodic systems $\mathbf{g}(t)$ and $\mathbf{h}(t)$ are evaluated from data by means of time averages. In many scenarios, such as one-dimensional systems or systems of spherical particles, $\boldsymbol{\theta}(t)$, $\mathbf{g}(t)$ and $\mathbf{h}(t)$ are diagonal

matrices, i.e. $\boldsymbol{\theta}(t) = \theta(t)\mathbf{1}$, $\mathbf{g}(t) = g(t)\mathbf{1}$ and $\mathbf{h}(t) = h(t)\mathbf{1}$. In such cases, hereinafter, we will denote the scalar functions simply as $\theta(t)$, $g(t)$ and $h(t)$.

B. Memory kernel in the Laplace space

Recently, Ref. [18] proposed a way to compute the memory kernel using the properties of Laplace transform defined as $\mathcal{L}_p(\mathbf{A}(t)) = \int_0^\infty \mathbf{A}(t)e^{-t/\lambda}$. In fact, it can be easily shown that (2) in the Laplace space takes the simple form $\mathcal{L}_p(\boldsymbol{\theta}(t)) = \mathcal{L}_p(\mathbf{g}(t))[\mathcal{L}_p(\mathbf{h}(t))]^{-1}$. However, despite its simplicity, this approach is not suitable in case of limited data, which produce correlations affected by random noise. As an example, let us consider a Gaussian error $\boldsymbol{\epsilon}(t)$ affecting only the function $\mathbf{g}(t)$. Then, the error acting on the Laplace transform of the kernel $\boldsymbol{\Theta}(\lambda) = \mathcal{L}_p(\boldsymbol{\theta}(t))$, is defined as $\Delta\boldsymbol{\Theta}(\lambda) = \tilde{\boldsymbol{\Theta}}(\lambda) - \boldsymbol{\Theta}(\lambda) = \mathcal{L}_p(\boldsymbol{\epsilon}(t))[\mathcal{L}_p(\mathbf{h}(t))]^{-1}$. If we assume $\boldsymbol{\epsilon}(t)$ to be the sum of non-systematic local errors, i.e. $\boldsymbol{\epsilon}(t) = \sum_i \boldsymbol{\epsilon}_i(t) = \sum_i \mathbf{k}_i \delta(t - t_i)$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$, then the total error acting on the memory kernel in the Laplace space becomes $\Delta\boldsymbol{\Theta}(\lambda) = \sum_i \mathbf{k}_i e^{-t_i/\lambda} [\mathcal{L}_p(\mathbf{h}(t))]^{-1}$. This argument shows that local random errors in the real space turn into non-local contributions in the Laplace space. Such a propagation can lead to significant inaccuracies, that sometimes compromise the memory kernel approximation. As an example, let us consider the simple case, that can be analytically solved, with $h(t) = e^{-t}$, and the noisy $g(t) = -te^{-t}$, as reported in Fig. 2(a). Fig. 2(b) shows that the exact memory kernel computed in the Laplace space (as in Ref. [18]) diverges from the analytical solution when noise becomes significantly large. To overcome this issue, we propose to adopt a MLP-based procedure that gives an optimal approximation of $\boldsymbol{\theta}$ in the real space. Our method is robust and allows to reproduces accurately the expected function even when the strongest noise affects the data, as shown in Fig. 2(c).

C. Memory kernel extraction through MLP

Among the different possible neural network structures, MLPs have gained popularity because of their versatility and capabilities in non-linear function approximations [21]. MLPs consist of at least three layers, known as input, hidden and output layers, each one including several nodes (Fig. 1). The transformation of the dataset at each node is determined by an activation function σ . The network learning process consists in an optimization algorithm

aiming to find the weights $w_{j,i}^l$ and the bias b_j^l that minimize a cost (or error) function C computed at the output of the MLP. Every repetition of this algorithm is called epoch, and denoted as e , and the whole procedure is commonly known as learning process.

In the present work, we adopt a three layer MLP with a single input and a single output function. The hidden layer has an arbitrary number of neurons, determining the degree of accuracy of the memory kernel approximation. The universal approximation theorem guarantees that such a structure of the network is able to approximate any continuous function defined on a compact subset of \mathbb{R}^d [19, 20]. As regards the activation functions, we adopt in the hidden layer $\sigma(z) = \int_0^t h(t-\tau)e^{z(\tau)}d\tau$, with $h(t)$ being known a priori, while $\sigma(z) = z$ at the output layer. The learning algorithm adopted is the resilient back-propagation algorithm with an adaptive learning rate η [22]. Provided the MLP with the two matrices $\mathbf{g}(t)$ and $\mathbf{h}(t)$, the memory kernel is then extracted in the form of an exponential series, namely as $\boldsymbol{\theta}(t) \sim \sum_{k=1}^{N_n} \mathbf{A}_k e^{\mathbf{B}_k(t)}$, where N_n is the number of nodes in the hidden layer, \mathbf{A}_k are matrices of real numbers and \mathbf{B}_k are matrices with all real negative coefficients.

D. GLE time integration

The integration of the GLE dynamics is a not-trivial task for two reasons: first, the convolution integral depends on the full history of the observable, and second, the stochastic term is correlated in time. Several approaches have been proposed to face these issues based on the introduction of a set of auxiliary variables, i.e. Refs [18, 23, 24]. In the present work, we take advantage of the exponential structure of $\boldsymbol{\theta}(t)$ to implement an integration algorithm. The history-dependent convolution term is written as a sum of the additional variables $\mathbf{Z}_k(t)$, each defined as $\mathbf{Z}_k(t) = \int_0^t \mathbf{A}_k e^{\mathbf{B}_k(t-\tau)} \mathcal{O}(\tau) d\tau$, so that their evolution equation can be expressed as $\dot{\mathbf{Z}}_k(t) = \mathbf{B}_k \mathbf{Z}_k(t) - \mathbf{A}_k \mathcal{O}(t)$. The noise $\mathbf{R}(t)$ has to be generated with proper time correlations in order to satisfy the fluctuation-dissipation theorem. The introduction of an additional set of auxiliary variables $\boldsymbol{\xi}_k(t)$ allows us to express it as function of a standard white noise processes. In details, the noise term is decomposed as $\mathbf{R}(t) = \sum_{k=1}^{Nn} \mathbf{R}_k(t) = \sum_{k=1}^{Nn} \mathbf{b}_k \boldsymbol{\xi}(t)$, so that the corresponding evolution reads $\dot{\mathbf{R}}_k(t) = \mathbf{B}_k \mathbf{R}_k(t) + \mathbf{b}_k \boldsymbol{\xi}(t)$, where $\boldsymbol{\xi}(t)$ is a white noise with zero mean and time correlation $\langle \boldsymbol{\xi}(t) \boldsymbol{\xi}(s) \rangle = 2 \langle \mathcal{O}, \mathcal{O} \rangle \delta(t-s)$, while the coefficients \mathbf{b}_k can be computed numerically (for details see Appendix). As a result, after defining the variables $\mathbf{S}_k(t) = -\mathbf{Z}_k(t) + \mathbf{R}_k(t)$, the GLE can be rewritten in

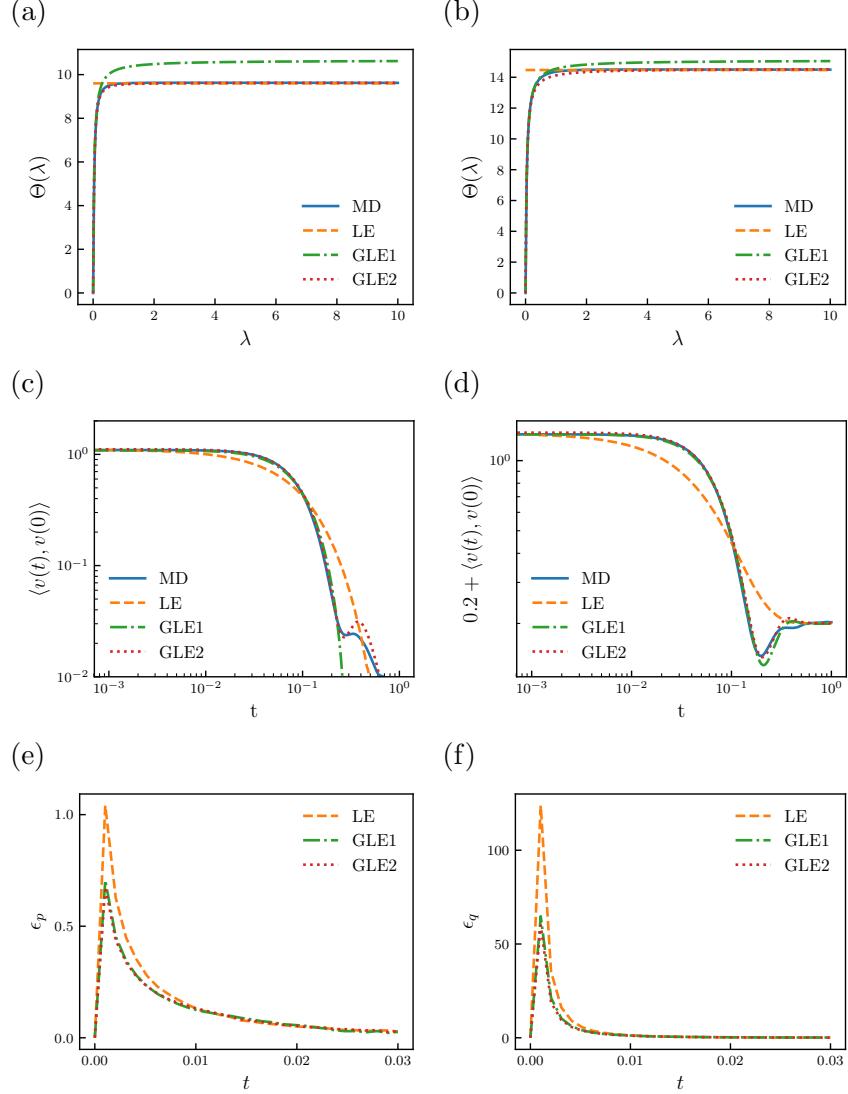


FIG. 3. Memory kernels computed with the MLP are compared against the one obtained directly from MD simulation in the Laplace space, for (a) LDL and (b) HDL cases. Velocity correlation functions computed from MD, LE and GLE dynamics over 10^4 trajectories for both (c) LDL and (d) HDL cases are also reported. GLE1 and GLE2 refer to the memory kernel approximations obtained respectively with 1 and 2 neurons in the hidden layer. In (e-f) we report the mean square differences $\epsilon_p(t)$ and $\epsilon_q(t)$ between the pdfs of the reduced systems (GLE and LE) and the exact pdf of the full system (MD) as function of the relaxation time.

form of extended dynamics as:

$$\begin{cases} \dot{\mathcal{O}} = \mathbf{F}(\mathcal{O}(t)) + \sum_{k=1}^{N_n} \mathbf{S}_k(t) \\ \dot{\mathbf{S}}_k(t) = \mathbf{B}_k \mathbf{S}_k(t) - \mathbf{A}_k \mathcal{O}(t) + \mathbf{b}_k \boldsymbol{\xi}(t), \end{cases} \quad (3)$$

with $\mathbf{F}(\mathcal{O}(t))$ accounting for the mean force contributions.

II. APPLICATIONS

A. Single particle in bath

The proposed methodology is tested, first, to model the global effect of an heat bath on a single particle. Data regarding momentum and forces of the target particle, with mass $m = 1$, immersed in a bath of identical particles with masses $m_b = 1.0$ are gathered from equilibrium MD simulations. The interaction between any two particles i and j is modelled by the Lennard-Jones (LJ) potential:

$$v_{\text{LJ}}(\mathbf{r}_{ij}) = \begin{cases} 4\epsilon \left[(\sigma/r_{ij})^{12} - (\sigma/r_{ij})^6 \right] & \text{if } r_{ij} \leq r_c, \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the particles, $\epsilon = 1.0$ is the depth of the potential well, $\sigma = 1.0$ is the finite atom-atom distance at which the potential is zero, and $r_c = 2.5\sigma$ is a cut-off radius. In this work, numerical results involving particles dynamics are reported in reduced units, using σ and ϵ to scale lengths, energies and times.

The simulation box dimensions are $10\sigma \times 10\sigma \times 10\sigma$, and periodic boundary conditions are imposed along x , y and z axes. A Nosé-Hoover thermostat is used to equilibrate the system at a reduced temperature $T = 1.0$. In this study, we consider two bath densities: the low density limit (LDL) with $\rho = 0.699$, and the high density limit (HDL) with $\rho = 0.799$. Figs 3(a-b) show the comparison in the Laplace space of the exact memory kernel computed with MD and with our procedure for both LDL and HDL cases. It is worth underlying that the Laplace

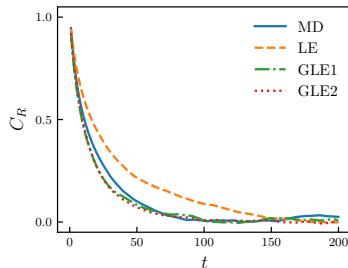


FIG. 4. Time correlation of the gyration radius of the particle chain in the bath at equilibrium computed from LE, GLE and MD simulations.

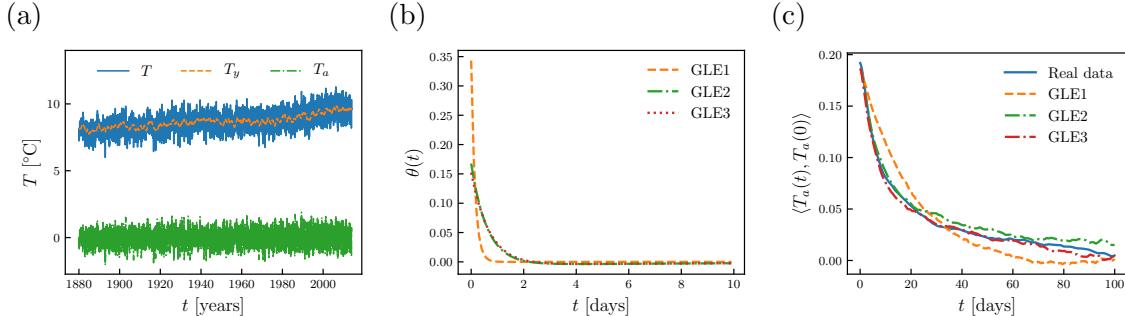


FIG. 5. (a) Global temperature $T(t)$, annual moving average temperature $T_y(t)$ and the daily anomaly $T_a(t) = T(t) - T_y(t)$ between 1880 and 2014. (b) Memory kernel approximations computed through MLP with 1,2 and 3 neurons in the hidden layer and (c) corresponding time correlations obtained from real data and GLE simulations.

space is used only for comparison purpose, since it allows to extract $\theta(t)$ numerically from MD data. From the comparison it emerges that the first order approximation obtained with the MLP outperforms the Markovian approximation at low λ , but is unable to catch the long term behavior in the Laplace space. In contrast, the second order approximations exactly overlay with the MD results for the entire spectrum of λ . The accuracy in the memory kernel approximation directly affects the velocity correlation functions obtained with the different methods, shown in Figs 3(c-d). In the log-log diagram clearly emerges the limitations of LE, which is able to replicate the correlation dynamics only on average. The first order approximation, on the contrary, is fairly accurate, but diverges for long times. Finally, the second degree approximation follows the exact autocorrelation within a tolerance lower than 1%. We also investigate the performances of GLE in reproducing MD dynamics out of equilibrium in the LDL case. By using the same θ evaluated in equilibrium conditions, we analyze the relaxation of momentum and position pdfs from a Dirac delta distribution to the equilibrium. In Figs 3(e-f), we report the standard error ϵ_p and ϵ_q between the reduced system pdfs and the full system pdf as function of the relaxation time. It emerges that GLE, if compared with LE, dramatically reduce both ϵ_p and ϵ_q up to about 50% during the non equilibrium relaxation.

B. Particles chain in bath

As an additional test, we analyze the dynamics of a chain of $N = 20$ particles in a bath. A LJ potential v_{LJ} is used to model pairwise non bonded interactions among chain and bath particles. The chain particles interactions are modelled by the following multi-body Dreiding potential [25], already adopted in Ref. [26] to study polymer chains deformations,

$$v(\mathbf{r}_{i,j,k,l}) = v_{LJ}(\mathbf{r}_{ij}) + v_H(\mathbf{r}_{ij}) + v_\theta(\mathbf{r}_{ijk}) + v_\phi(\mathbf{r}_{ijkl}), \quad (5)$$

where $v_H(\mathbf{r}_{ij}) = k_H(\mathbf{r}_{ij} - \mathbf{r}_0)^2$, $v_\theta(r_{ijk}) = k_\theta(\theta_{ijk} - \theta_0)^2$ and $v_\phi(r_{ijkl}) = k_\phi(1 + \cos(2\phi_{ijkl}))$ account for linear, angular and dihedral bonds, respectively (for more details see the Appendix). The bath has the same characteristics (density $\rho = 0.699$, temperature $T = 1$ and interaction potential v_{LJ}) of the LDL case for the single particle. This choice, together with the assumption that the potential of mean force acting among the chain particles is approximately equal to $v(\mathbf{r}_{i,j,k,l})$, allows us to use the same memory kernel obtained for the single particle (see Fig. 3). Particles chains are usually used to model polymers, whose characteristic dimensions are described by the gyration radius, defined as $R_G^2 = \frac{1}{N} \sum_{k=1}^N (\mathbf{r}_k - \mathbf{r}_{CM})^2$, where \mathbf{r}_k and \mathbf{r}_{CM} are the position of the particle k and of the center of mass of the chain respectively. Fig. 4 shows the radius of gyration autocorrelation [27, 28], computed as $C_R = [\langle R_G(t)^2 R_G(0)^2 \rangle - \langle R_G(0)^2 \rangle^2] [\langle R_G(0)^4 \rangle - \langle R_G(0)^2 \rangle^2]^{-1}$, for the particle chain dynamics at equilibrium simulated with LE, GLE and MD. It is interesting to observe that GLE, already with a single neuron, is able to accurately reproduce the bath effects on the chain and outperforms the commonly used Markovian approximation.

C. Modelling global temperature

Several stochastic models have been developed to reproduce and forecast global and local temperature dynamics, for example in Refs [29–31]. In the present work, we show that GLE, parametrized through our method, is able to accurately model the global daily temperature fluctuations with respect to a properly chosen moving average. It is worth underlying that the methodology can be also employed to model local temperature dynamics. We consider the daily land-average global temperature $T(t)$ measured during the period 1880-2014, published by the Berkeley Earth [32, 33]. Despite the local temperature showing cyclical trends in the short period due, for instance, to season changes, global temperature does not exhibit

a significant seasonal behavior, being a result of the energy balance between solar and earth radiations [34]. Nevertheless, $T(t)$ reveals non-stationarity features due to a long period increasing trend related to global warming, as visible in Fig. 5(a). Hence, we first compute the long term dynamics $T_y(t)$ as an yearly moving average. Then, we define the observable of interest as $T_a(t) = T(t) - T_y(t)$, so that the corresponding time series is stationary (see Appendix). Consequently, we model $T_a(t)$ with the GLE $\partial_t T_a(t) = - \int_0^t \theta(t-\tau) T_a(\tau) d\tau + R(t)$. In fact, this is a generalization of the Markovian model for weather derivatives proposed by Ref. [29]. In Fig. 5(b), we plot various degrees of approximations of the memory kernel extracted with our MLP-based method, while Fig. 5(c) shows the corresponding correlations functions. In first place, it emerges an excellent agreement between the correlations obtained with GLE dynamics and the real world data, especially when three neurons are adopted in the hidden layer. Then, matching the relaxation times of memory kernel (\sim days) with the characteristic time of the variable T_y (\sim years), we can obtain the evolution of $T(t)$ as sum of a Markovian yearly (long term) contribution and a non-Markovian daily (short term) term, namely:

$$\partial_t T(t) = (\partial_t + \theta_c) T_y(t) - \int_0^t \theta(t-\tau) T(\tau) d\tau + R(t), \quad (6)$$

where we introduced the constant $\theta_c = \int_0^t \theta(\tau)$. (6), originating directly from data, reflects the main features of global temperature multi-scale dynamics and, thus, gives clear insights into current questions regarding, for instance, global warming.

D. A stock market model: the Nikkei index

In more than one study, stochastic models have been employed to model financial instruments, such as bonds and stock prices [35–37]. In fact, operations such as financial risk management and portfolio optimization require accurate predictions of markets dynamics to maximize profits. However, most of the models used in finance relies on Markovian assumptions, which can potentially introduce inaccuracies. In this work, we overcome such limitations by modelling with a GLE the daily price of the Japanese financial index Nikkei $NI(t)$ between May 1949 and May 2018 [38]. As many other financial instruments, $NI(t)$ exhibits a non-stationary behavior in both mean and variance. Thus, we build an observable defined as $NI_a(t) = [NI(t) - NI_y(t)] / \sigma_y(t)$, with $NI_y(t)$ and $\sigma_y(t)$ be-

ing respectively a moving average and a moving standard deviation computed over a period $[t-y, t-1]$. The parameter y is then properly chosen in order to obtain a stationary $NI_a(t)$; In this work we find $y = 10$ days to be the optimal value (for more details see Appendix). Hence, we model the normalized stock price $NI_a(t)$ with the following non-Markovian model $\partial_t NI_a(t) = - \int_0^t \theta(t-\tau) NI_a(\tau) d\tau + R(t)$. Figs 6(b-c) show various degrees of approximations obtained with our method and the corresponding correlations functions. In contrast with the global temperature trend, $NI_a(t)$ do not exhibit a clear time-scale separation between memory kernel and autocorrelation decay. The comparisons between the correlations obtained with GLE dynamics and the real data exhibit a growing accuracy with an increasing number of neurons in the hidden layer. In fact, with the third order approximation we are able to reproduce the correlation decay with a maximum relative errors of order 10^{-2} . The proposed GLE equation, parametrized with a MLP equipped with 3 neurons, is employed in a comparison between the predicted probability distribution and actual market data for four time windows, each ten market days long, between 15 Jun 2018 and 10 Aug 2018 (Figs 6(d)). It emerges that our model is able not only to predict most of the actual market trend, but, more important, gives quite accurate information on the local variance of the trend, thus giving the chance of optimizing risk management in short term (\sim weakly) investments.

III. CONCLUSIONS

In this work, we have propose a novel methodology to parametrize a GLE dynamics of an observable by means of deep neural networks. By using machine learning to eliminate the unknowns from GLEs, our methodology outperforms previous approaches in terms of efficiency and robustness. In fact, despite previous approaches using Laplace transform, we have shown that the presented methodology does not suffer random data fluctuations typically present in real system data-sets. The general applicability of our approach, guaranteed by the universal approximation theorem, makes its use appealing in a variety of applications. In fact, our methodology is tested against several prototypical examples, from a colloidal systems to particle chains in a bath, to climatology and financial models. In all cases, we show an excellent agreement between the actual and the approximated dynamics of the observables under consideration. Thus, coupling deep learning with the most general equation

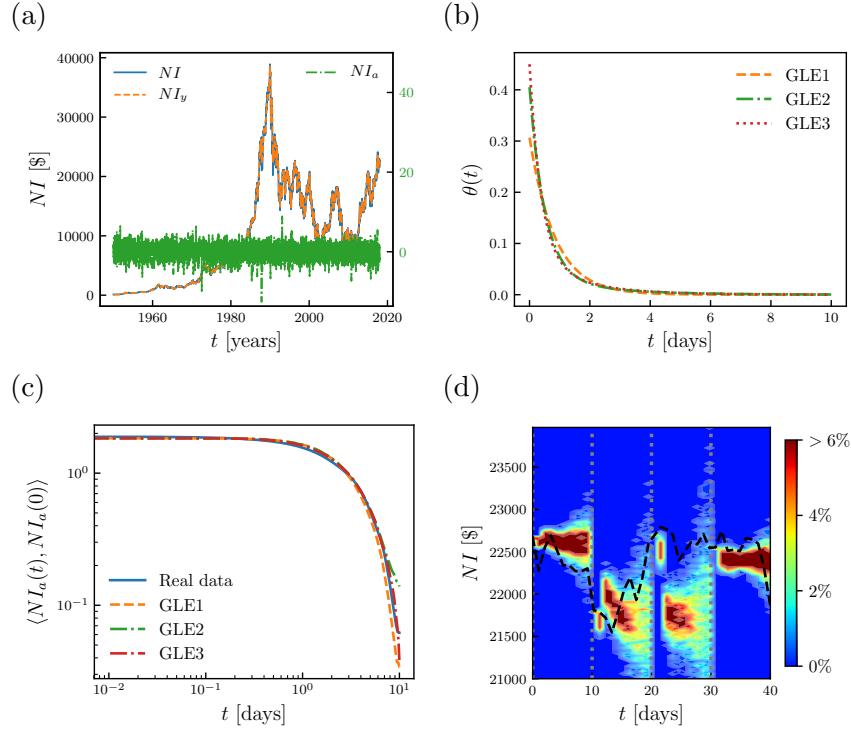


FIG. 6. (a) Daily close price of Nikkei index $NI(t)$, moving average index $NI_y(t)$ computed over a window of 10 days preceding the time t and the normalized index $NI_a(t)$ between 1949 and 2018. (b) Memory kernel approximations computed through MLP with 1,2 and 3 neurons in the hidden layer and (c) the corresponding time correlations obtained from real data and GLE simulation. (d) Comparison between predicted probability distribution (color-map) and actual market data (dashed black line). Dotted lines in gray delineate the 10 days long investment windows.

of statistical physics, namely GLE, opens the doors for a new way of modelling and understanding complex systems. Future developments of our method will involve MLPs equipped with complex exponential functions, since this may lead to enhanced approximations of oscillatory memory kernels.

IV. ACKNOWLEDGMENTS

We acknowledge financial support from the Engineering and Physical Sciences Research Council of the UK via grants No. EP/L020564 and EP/L025159 and from the European Research Council via Advanced Grant No. 247031.

Appendix A: Mori-Zwanzig's formalism

Let us consider the following (linear or non-linear) deterministic dynamical system:

$$\begin{cases} \frac{d\mathbf{z}}{dt} = \mathbf{f}(\mathbf{z}) \\ \mathbf{z}(0) = \mathbf{z}_0 \end{cases} \quad (\text{A1})$$

where $\mathbf{z} \in \mathbb{R}^n$ is a vector of independent variables. For the system in (A1), it can be defined a set of observables $\mathcal{O}(\mathbf{z}, t) = \phi(\mathbf{z}(t))$, where ϕ represent the transformation map between \mathbf{z} and \mathcal{O} . By using the chain rule, it is easy to show that the evolution equation of $\mathcal{O}(\mathbf{z}, t)$ can be written as:

$$\begin{cases} \frac{\partial \mathcal{O}}{\partial t}(\mathbf{z}, t) = \mathcal{L}\mathcal{O} \\ \mathcal{O}(\mathbf{z}, 0) = \mathcal{O}_0 \end{cases} \quad (\text{A2})$$

where it was introduced the operator $\mathcal{L} = \mathbf{f}(\mathbf{z}) \cdot \nabla_{\mathbf{z}}$. It follows that the solution of (A2) can be written as:

$$\mathcal{O}(\mathbf{z}, t) = e^{\mathcal{L}t} \mathcal{O}_0 \quad (\text{A3})$$

where the exponential has to be intended as the power series that defines the exponential map between matrices.

If we are only interested in the dynamics of some observables \mathcal{O} , rather then the whole solution $\mathbf{z}(t)$, we can define a projection operator \mathcal{P} , which maps functions of \mathbf{z} into function of \mathcal{O} . It is worth underlining that, in general, the set of observables \mathcal{O} may be defined by a linear or nonlinear transformation of \mathbf{z} , but in any case the evolution of \mathcal{O} is supposed to be unitary, i.e. $|\mathcal{O}(t)|^2 = |\mathcal{O}(0)|^2$. A simple, but still important, scenario is given by \mathcal{O} being a subset of \mathbf{z} . As we will see later, this case plays a fundamental role in dimensional reductions of multi-component systems, i.e. colloidal particles in a thermal bath. Given a projection operator \mathcal{P} , namely a transformation from a vector space to itself such that $\mathcal{P}^2 = \mathcal{P}$, one can follow Mori-Zwanzig's formalism[3–5] to obtain a form of (A2) suitable for system dimensionality reduction. Note that at this point no constrain is put on the form of the projection operator. After defining the operator $\mathcal{Q} = \mathbf{1} - \mathcal{P}$, orthogonal to \mathcal{P} , (A2) can be rewritten as:

$$\frac{\partial \mathcal{O}}{\partial t}(\mathbf{z}, t) = \mathcal{L}e^{\mathcal{L}t}\mathcal{O}_0 = e^{\mathcal{L}t}\mathcal{P}\mathcal{L}\mathcal{O}_0 + e^{\mathcal{L}t}\mathcal{Q}\mathcal{L}\mathcal{O}_0 \quad (\text{A4})$$

Duhamel-Dyson's formula allows to rewrite the exponential term $e^{\mathcal{L}t}$ as:

$$e^{\mathcal{L}t} = e^{\mathcal{Q}t} + \int_0^t e^{\mathcal{L}(t-\tau)} \mathcal{P} e^{\mathcal{Q}\tau} d\tau \quad (\text{A5})$$

and, consequently, (A4) becomes the so called Mori-Zwanzig's equation:

$$\frac{\partial \mathcal{O}}{\partial t}(\mathbf{z}, t) = e^{\mathcal{L}t} \mathcal{P} \mathcal{L} \mathcal{O}_0 + \int_0^t e^{\mathcal{L}(t-\tau)} \mathcal{P} \mathcal{L} e^{\mathcal{Q}\tau} \mathcal{Q} \mathcal{L} \mathcal{O}_0 d\tau + e^{\mathcal{Q}\mathcal{L}t} \mathcal{Q} \mathcal{L} \mathcal{O}_0 \quad (\text{A6})$$

The first term is the Markovian contribution, the second constitutes the memory term and the last one is often interpreted as the noise. It is worth noticing that, at this stage, (A6) is exactly equivalent to (A1) and is valid independently from the specific choice of the projection operator \mathcal{P} . Mori and Zwanzig [4, 5, 13] proposed two different projection operators leading to different forms of GLE, that we will briefly discuss in next sections.

If we name the noise term $\mathbf{R}(t) = e^{\mathcal{Q}\mathcal{L}t} \mathcal{Q} \mathcal{L} \mathcal{O}_0$, then the following dynamical system remains determined:

$$\begin{cases} \frac{\partial \mathbf{R}}{\partial t}(\mathcal{O}_0, t) = \mathcal{Q} \mathcal{L} \mathbf{R}(\mathcal{O}_0, t), \\ \mathbf{R}(\mathcal{O}_0, t) = \mathcal{Q} \mathcal{L} \mathcal{O}_0. \end{cases} \quad (\text{A7})$$

Projecting (A7) according to \mathcal{P} , it follows:

$$\begin{cases} \mathcal{P} \frac{\partial \mathbf{R}}{\partial t}(\mathcal{O}_0, t) = \mathcal{P} \mathcal{Q} \mathcal{L} \mathbf{R}(\mathcal{O}_0, t) = \mathbf{0}, \\ \mathcal{P} \mathbf{R}(\mathcal{O}_0, t) = \mathcal{P} \mathcal{Q} \mathcal{L} \mathcal{O}_0 = \mathbf{0}, \end{cases} \quad (\text{A8})$$

where we have used the property of the projection operator $\mathcal{P} \mathcal{Q} = \mathbf{0}$. This shows that $\mathbf{R}(t)$ is orthogonal to the range of \mathcal{P} for any time t , and consequently is orthogonal to \mathcal{O} . However, in order to express $\mathbf{R}(t)$ as a stochastic process, it is necessary to have either time scale separation or weak coupling between resolved and unresolved variables[39]. When at least one of such conditions occurs, at least asymptotically, the influence of the unresolved variables may be interpreted as sum of many uncorrelated events, and consequently can be treated with Central Limit Theorem[40]. Thus, it is the Central Limit Theorem that determines the Gaussian shape for the distribution of $\mathbf{R}(t)$, while its time correlation follows from the fluctuation dissipation theorem, as shown in what follows.

1. Mori's projection operator

The projection operator introduced by Mori[4], when applied to a general variable $\mathbf{A}(\mathbf{z})$, is given by:

$$\mathcal{P}\mathbf{A}(\mathbf{z}) = \langle \mathbf{A}, \mathcal{O}_0 \rangle \langle \mathcal{O}_0, \mathcal{O}_0 \rangle^{-1} \mathcal{O}_0 \quad (\text{A9})$$

where the inner product $\langle \mathbf{A}, \mathbf{B} \rangle$ is defined as

$$\langle \mathbf{A}, \mathbf{B} \rangle = \int \rho(\mathbf{z}) \mathbf{A}(\mathbf{z}) \mathbf{B}^*(\mathbf{z}) d\mathbf{z} \quad (\text{A10})$$

with $\rho(\mathbf{z})$ being a normalized probability density function defined in the phase space of the original system and \mathbf{B}^* the conjugate transpose of \mathbf{B} . In case of systems with Hamiltonian \mathcal{H} in a canonical ensemble, the probability density function is $\rho(\mathbf{z}) = Z^{-1} e^{-\beta \mathcal{H}(\mathbf{z})}$, where Z is the partition function and $\beta = k_B T$. Employing Mori's operator in (A6), we obtain the Markovian term:

$$e^{\mathcal{L}t} \mathcal{P} \mathcal{L} \mathcal{O}_0 = \langle \mathcal{L} \mathcal{O}_0, \mathcal{O}_0 \rangle \langle \mathcal{O}_0, \mathcal{O}_0 \rangle^{-1} \mathcal{O}(t). \quad (\text{A11})$$

Moreover, from the definition of $\mathbf{R}(t)$, we obtain the memory term:

$$\int_0^t e^{\mathcal{L}(t-\tau)} \mathcal{P} \mathcal{L} e^{\mathcal{Q}\tau} \mathcal{Q} \mathcal{L} \mathcal{O}_0 d\tau = - \int_0^t \boldsymbol{\theta}(\tau) \mathcal{O}(t-\tau) d\tau \quad (\text{A12})$$

where the memory kernel is defined as $\boldsymbol{\theta}(t) = -\langle \mathcal{L} \mathbf{R}(t), \mathcal{O}_0 \rangle \langle \mathcal{O}_0, \mathcal{O}_0 \rangle^{-1}$. Since $\mathcal{Q} \mathbf{R}(t) = \mathbf{R}(t)$, and \mathcal{L} is an anti-Hermitian operator [13], it follows that $\langle \mathcal{L} \mathbf{R}(t), \mathcal{O}_0 \rangle = -\langle \mathbf{R}(t), \mathcal{L} \mathcal{O}_0 \rangle = -\langle \mathbf{R}(t), \mathcal{Q} \mathcal{L} \mathcal{O}_0 \rangle = -\langle \mathbf{R}(t), \mathbf{R}(0) \rangle$. Hence, we obtain the following relation:

$$\boldsymbol{\theta}(t) = \langle \mathbf{R}(t), \mathbf{R}(0) \rangle \langle \mathcal{O}_0, \mathcal{O}_0 \rangle^{-1}, \quad (\text{A13})$$

which constitutes the fluctuation dissipation theorem.

2. Zwanzig projection operator

As Zwanzig pointed out, Mori's projection operator leads to a linearized GLE[13]. Zwanzig [3, 13] defined the projection operator applied to the variable $\mathbf{A}(\mathbf{z})$ through the following conditional expectation:

$$\mathcal{P}\mathbf{A}(\mathbf{z}) = \frac{\int \rho(\mathbf{z}) \mathbf{A}(\mathbf{z}) \delta(\mathcal{O} - \phi(\mathbf{z})) d\mathbf{z}}{\int \rho(\mathbf{z}) \delta(\mathcal{O} - \phi(\mathbf{z})) d\mathbf{z}}, \quad (\text{A14})$$

where $\delta(\mathcal{O} - \phi(\mathbf{z})) = \prod_j \delta(\mathcal{O}_j - \phi_j)$.

In molecular dynamics, the set of observables is often defined as a subset of the original coordinates, namely $\mathcal{O} \subseteq \mathbf{z}$. In this case, Zwanzig's projection operator allows to express the Markovian term in (A6) as function of the potential of mean force. To show this, let us consider an isothermal Hamiltonian system of N particles with coordinates $\mathbf{z} = \{\mathbf{r}, \mathbf{p}\}$, where $\mathbf{r} = \{\mathbf{r}_1 \dots \mathbf{r}_N\}$ and $\mathbf{p} = \mathbf{p}_1 \dots \mathbf{p}_N$ are position and momenta, respectively. With $\mathbf{f}(\mathbf{z}) = -\nabla_{\mathbf{z}} V(\mathbf{z})$, (A1) gives the Newton's equations of motion for a system of interacting particles. Suppose one is interested in the dynamical evolution of only n of the original N particles, whose coordinates (called relevant variables) and are indicated as $\tilde{\mathbf{z}} = \{\mathbf{r}_1 \dots \mathbf{r}_n, \mathbf{p}_1 \dots \mathbf{p}_n\}$. The remaining variables, called unresolved variables, are denoted by $\hat{\mathbf{z}} = \{\mathbf{r}_{n+1} \dots \mathbf{r}_N, \mathbf{p}_{n+1} \dots \mathbf{p}_N\}$. Hence, inserting Zwanzig's operator in (A6), we obtain the Markovian term in the form:

$$\mathcal{P}\mathcal{L}\tilde{\mathbf{z}} = \frac{\int -\nabla_{\mathbf{z}} V(\mathbf{z}) e^{-\beta\mathcal{H}(\mathbf{z})} \delta(\mathbf{z} - \tilde{\mathbf{z}}) d\mathbf{z}}{\int e^{-\beta\mathcal{H}(\mathbf{z})} \delta(\mathbf{z} - \tilde{\mathbf{z}}) d\mathbf{z}} = -\nabla_{\tilde{\mathbf{z}}} V^{\text{PMF}}(\tilde{\mathbf{z}}) \quad (\text{A15})$$

where V^{PMF} is known as potential of mean force. Moreover, the memory term can be written in terms of the noise term as:

$$\int_0^t e^{\mathcal{L}(t-\tau)} \mathcal{P}\mathcal{L}\mathbf{R}(\tau) d\tau. \quad (\text{A16})$$

Ref.[10] has shown that the term in (A16) is null for the position coordinates \mathbf{r} , while can expressed for the momentum coordinates \mathbf{p} as:

$$- \int_0^t \boldsymbol{\theta}(\tau) \mathbf{p}(t - \tau) d\tau \quad (\text{A17})$$

Another interesting result was derived by Ref. [7] in the framework of particles coarse-graining. In fact, the authors proved that, given the position of the coarse-grained particles $\alpha = 1 \dots N$, defined as

$$\mathbf{r}_\alpha = \frac{\sum_i m_{\alpha,i} \mathbf{r}_{\alpha,i}}{M_\alpha}, \quad (\text{A18})$$

where $M_\alpha = \sum_i m_{\alpha,i}$, and the corresponding momentum

$$\mathbf{p}_\alpha = \sum_i \mathbf{p}_{\alpha,i}, \quad (\text{A19})$$

then the momentum equation for the coarse-grained particle σ takes the following form:

$$\partial_t \mathbf{p}_\sigma(t) = \beta^{-1} \frac{\partial}{\partial \mathbf{r}_\sigma} \ln \omega(\mathbf{r}) - \sum_{\alpha=1}^N \beta \int_0^t \boldsymbol{\theta}_\alpha(\tau) \frac{\mathbf{p}_\alpha}{M_\alpha} (t - \tau) d\tau + \mathbf{R}(t). \quad (\text{A20})$$

The term $\omega(\mathbf{r})$ is given by

$$\omega(\mathbf{r}) = \frac{\int d\mathbf{r} \delta(\hat{\mathbf{r}} - \mathbf{r}) e^{\beta U}}{\int d\mathbf{r} e^{\beta U}}. \quad (\text{A21})$$

with U being the potential energy. Moreover, the memory $\boldsymbol{\theta}_\alpha(\tau)$ satisfies the fluctuation dissipation relation:

$$\boldsymbol{\theta}_\alpha(\tau) = \langle \mathbf{R}_\sigma(\tau) \mathbf{R}_\alpha^T(0) \rangle. \quad (\text{A22})$$

Appendix B: Multi-layer perceptron structure

Artificial neural networks are an interesting substitute of conventional methods in the parameterization of the GLE because of their enhanced capabilities in function approximations. Developed in analogy with the biological processes in the brain, artificial neural networks can be series of linear and non-linear transformations of some inputs in some output. Among the different possible variants, MLPs have gained popularity because of their versatility and capability in non-linear function approximations [21]. MLPs consist of at least three layers, known as input, hidden and output layers, each one including several nodes (Fig. 1). Each node i in the layer $l - 1$ is connected with any other node j in the successive layer l and every connection is characterized by a parameter called weight $w_{j,i}^l$. In addition, to every neuron in the network corresponds a parameter called bias b_j^l . The transformation of the dataset at each node is determined by an activation function $\sigma(z_j^l)$. It follows that the output a_j^l of the neuron j of the layer l is computed as $a_j^l = \sigma(z_j^l)$, with $z_j^l = \sum_i w_{j,i}^l a_i^{l-1} + b_j^l$. The network learning process consists in an optimization algorithm aiming to find the weights $w_{j,i}^l$ and the bias b_j^l that minimize a cost (or error) function C computed at the output of the MPL. In this work, we employ a quadratic cost function

$$C = \sum_{t_j}^{N_t} \frac{1}{2N_t} (y_j(t_j) - a_j^L(t_j))^2, \quad (\text{B1})$$

where N_t is the number of data samples. Hence, an algorithm is used to cyclically back-propagate the information about the error evaluated at the output to update weights and bias. Every repetition of this algorithm is called epoch e and the whole procedure is commonly known as learning process.

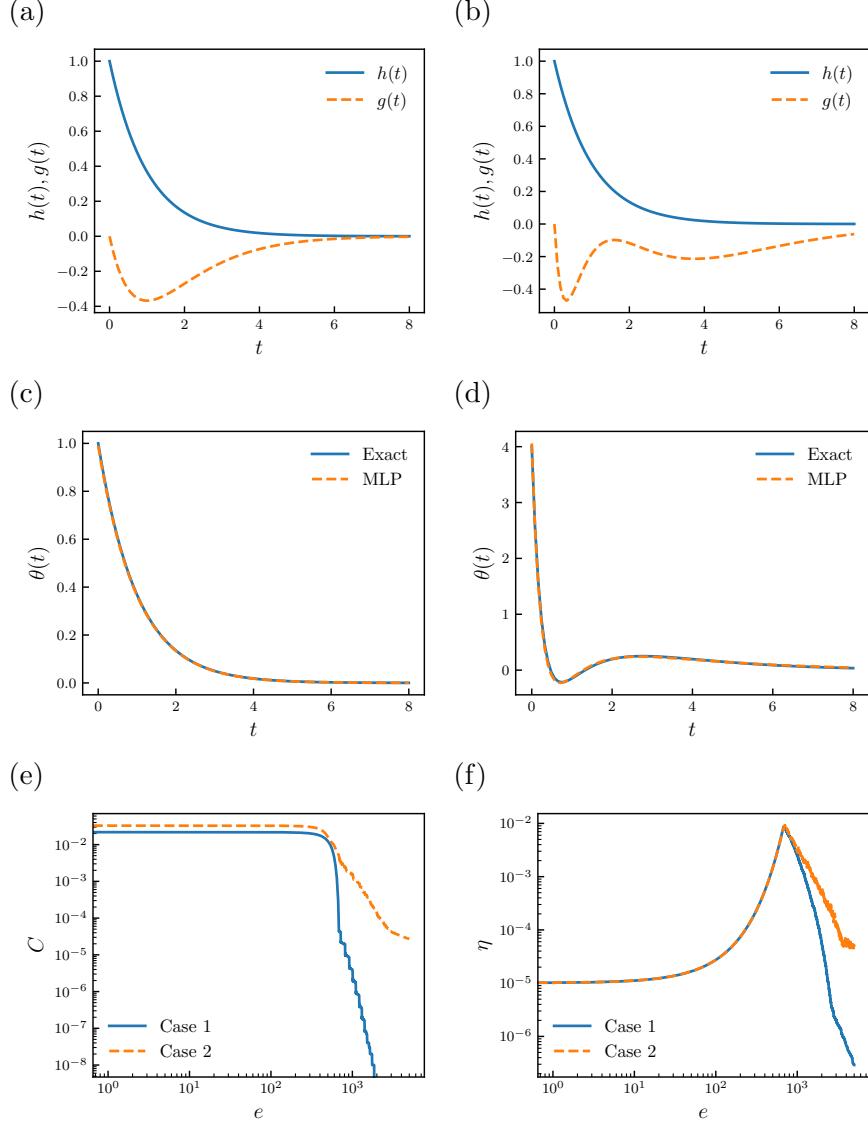


FIG. 7. In order to test our methodology two representative cases (discussed in the main text) are reported: Case 1 in (a-c) and case 2 in (b-d). The functions $h(t)$ and $g(t)$ (a-b), discretized at 800 points, are provided to the MLP. The comparison between the memory kernel θ computed numerically with our MLP and the exact one is given in (c-d). In (e-f) we show the cost function and learning rate for the two analyzed scenarios. In both cases, the numerical approximation is obtained with a MLP trained for 5000 epochs.

In the present work, we adopt a three layer MPL with a single input and a single output function. The hidden layer has an arbitrary number of neurons N_n , determining the degree of accuracy of the memory kernel approximation. The universal approximation theorem guarantees that such a structure of the network is able to approximate any continuous function

defined on a compact subset of \mathbb{R}^d [19, 20]. Initialization of the MPLs is obtained providing Gaussian distributed random numbers to the weights, and zeros to the bias. Moreover, no bias is added at the output layer. As regards the activation functions, in the hidden layer we adopt $\sigma(z) = \int_0^t h(t-\tau)e^{z(\tau)}d\tau$, with $h(t)$ being known a priori, while we employ $\sigma(z) = z$ at the output layer. The learning algorithm adopted is the resilient back-propagation algorithm (Rprop) [22], which can be synthesized as follows:

$$\eta(e) = \begin{cases} \eta^+ \cdot \eta(e-1) & \text{if } \frac{\partial C}{\partial \alpha}(e) \cdot \frac{\partial C}{\partial \alpha}(e-1) > 0, \\ \eta^- \cdot \eta(e-1) & \text{if } \frac{\partial C}{\partial \alpha}(e) \cdot \frac{\partial C}{\partial \alpha}(e-1) < 0, \\ \eta(e-1) & \text{otherwise,} \end{cases} \quad (\text{B2})$$

where $\alpha = [w_{j,i}^l; b_j^l]$, e indicates the epoque, η is the adaptive learning rate and $0 < \eta^- < 1 < \eta^+$ are fixed parameters. In our experience and according to the literature [22], Rprop algorithm gives an optimal compromise between fastness of the response and solution convergence. Provided the MLP with the two matrices $g(t)$ and $h(t)$, the memory kernel is then extracted in the form of an exponential series, namely as:

$$\theta(t) \sim \sum_{k=1}^{N_n} w_k^3 e^{b_k^2 t} = \sum_{k=1}^{N_n} A_k e^{B_k(t)}, \quad (\text{B3})$$

where N_n is the number of nodes in the hidden layer, $A_k = w_k^3 e^{b_k^2}$ are real numbers and $B_k = w_k^2$ are strictly real negative coefficients. The algorithm presented so far has been adopted to extract the memory kernel in case of diagonal $\theta(t)$. However, in the next subsection we present a generalization of our approach to non diagonal memory kernel matrices.

1. MLP for general memory kernel tensor

In this section, we discuss a generalization of the MLP structure that allows to compute also non-diagonal memory kernel matrices. To this aim, it is useful to rewrite the m th raw of the GLE explicitly as:

$$\left\{ \begin{array}{l} g_{m1}(t) = - \int_0^t \sum_k \theta_{mk}(t-\tau) h_{k1}(\tau) d\tau \\ g_{m2}(t) = - \int_0^t \sum_k \theta_{mk}(t-\tau) h_{k2}(\tau) d\tau \\ \dots \\ g_{mn}(t) = - \int_0^t \sum_k \theta_{mk}(t-\tau) h_{kn}(\tau) d\tau \end{array} \right. \quad (\text{B4})$$

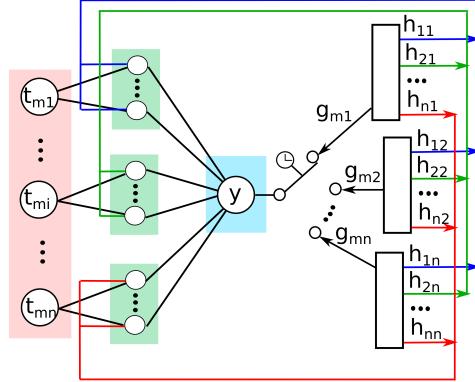


FIG. 8. MLP structure for systems with non diagonal memory kernel matrix

This system of equations can then be solved to compute $\theta_{mk}(t)$, with $k = \{1, n\}$. In line with the approach followed to find diagonal memory kernel matrices, we propose to employ the computational structure illustrated in Fig. 8. Such a network is made of n parallel MLPs, each responsible for the approximation of a function $\theta_{mk}(t)$. A switch allows to provide both the hidden layer and the output with different components of \mathbf{g} and \mathbf{h} , such that all the relations in (B4) are sequentially employed to train the MLPs.

The network training process is based on the updates of weights and bias. For the quadratic cost function, weights and bias updates are defined as:

- Error at the output layer: $\delta^{(3)} = \frac{\partial C}{\partial z^{(3)}} = \frac{\partial C}{\partial a^{(3)}} \frac{\partial a^{(3)}}{\partial z^{(3)}} = \frac{1}{N_t} \sum_{t_j}^{N_t} (y_j(t_j) - a_j^L(t_j))$;
- Error of the neuron 'j' at the layer 'l': $\delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_k w_{j,i}^{l+1} \delta_k^{l+1} \frac{\partial \sigma(z_j^l)}{\partial z_j^l}$;
- update of weight: $\Delta w_{j,i}^l = -\eta a_k^{l-1} \frac{\partial C}{\partial z_j^l}$;
- update of bias: $\Delta b_j^l = -\eta \frac{\partial C}{\partial z_j^l}$.

As a preliminary test of our approach, we consider three simple functions $h(t)$, θ and $g(t)$ for which can be analytically shown that $g(t) = - \int_0^t \theta(t - \tau) h(\tau) d\tau$. Given $h(t)$ and $g(t)$, an approximation of θ is computed with our methodology and is compared with the exact analytical θ . Two tests with different sets of functions are reported here. The functions used for the first test are the following:

$$h(t) = e^{-t}, \quad \theta(t) = e^{-t}, \quad g(t) = -te^{-t} \quad (B5)$$

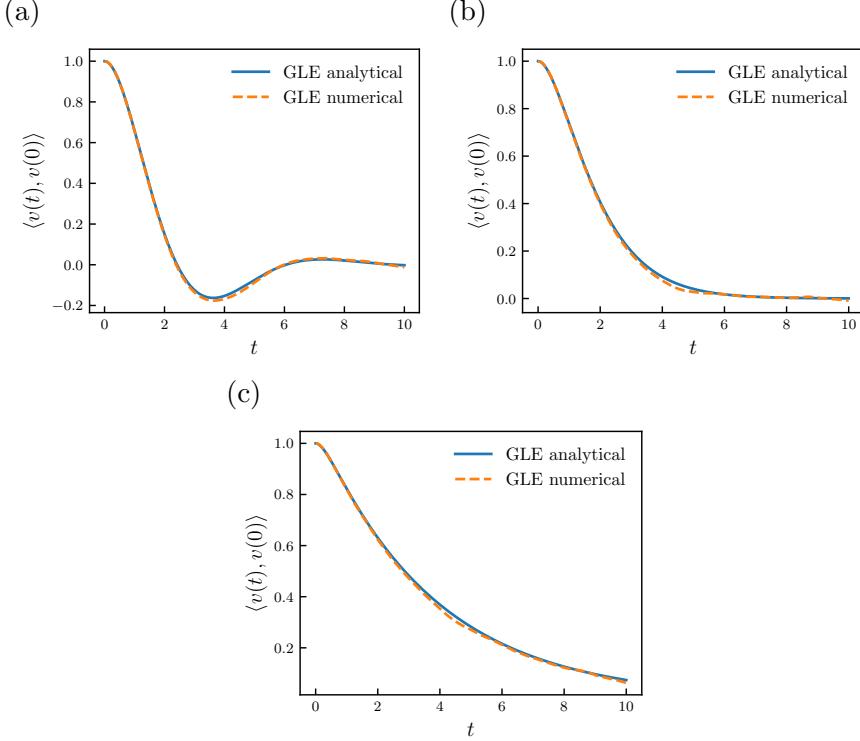


FIG. 9. Comparison between numerical and analytical time correlation computed over 10^4 independent trajectories for a GLE with memory kernel in the form of a single exponential function[23]. The correlation is computed in the under-damped limit with $A=1$ and $B=1$ (a), in the damped case with $A=1$ and $B=-2$ (b) and in the over-damped limit with $A=1$ and $B=-4$ (c). In all cases the temperature is set to $T = 1$.

In this test, because of the single exponential form of θ , a MLP with a single neuron in the hidden layer is adopted, namely $N_n = 1$.

The functions adopted for the second test are the following:

$$\begin{aligned} h(t) &= e^{-t}, \quad \theta(t) = 6e^{-4t} - 4e^{-t} + 2e^{-t/2} \\ g(t) &= - (2e^{-t} - 2e^{-4t} - 4te^{-t} + 4e^{3t/2} - 4e^{-t}) \end{aligned} \tag{B6}$$

For this latter example, we impose $N_n = 3$ neurons in the hidden layer.

Fig. 7(a-b) reports $h(t)$ and $g(t)$ provided as input to the MLP for both tests. The comparisons between numerical approximations and analytical θ reported in Fig. 7(c-d) clearly shows the accuracy of our methodology. The behaviors of cost function and learning rate during the learning process for both tests is also shown in Fig. 7(e-f).

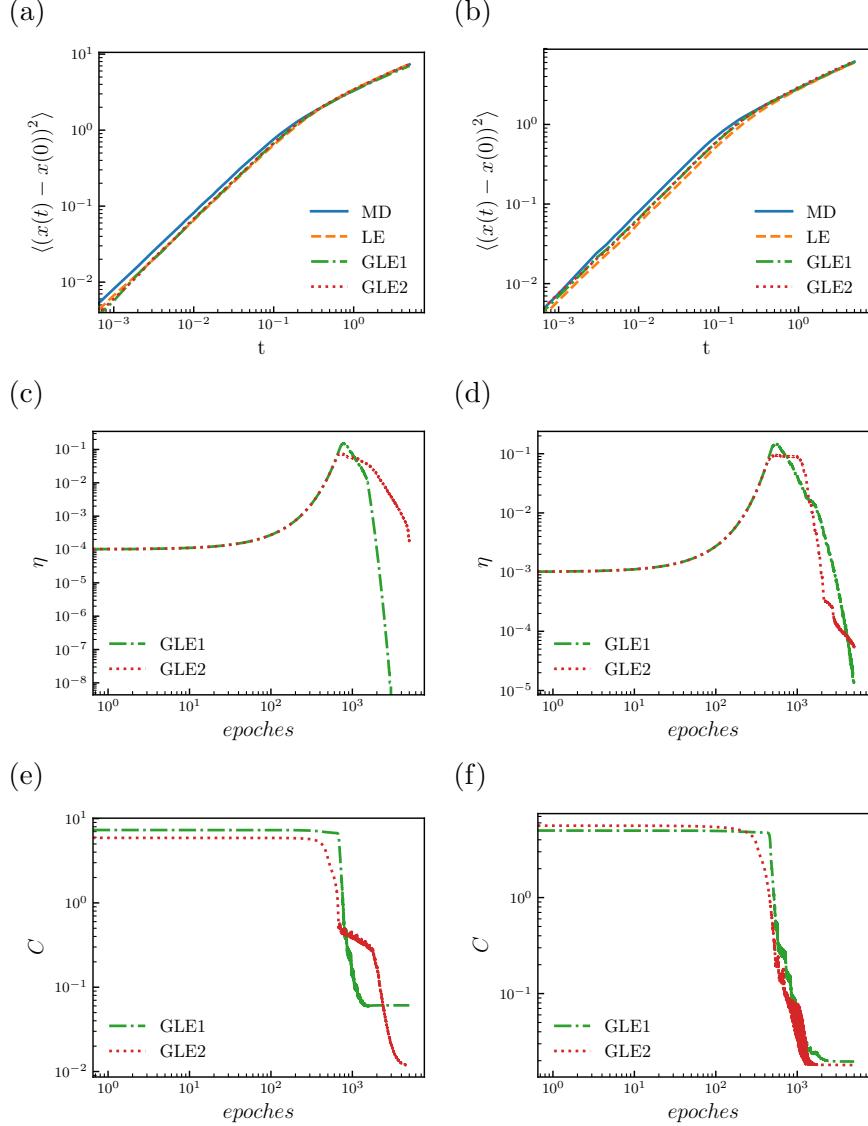


FIG. 10. Comparison of the mean square displacement (MSD) computed with the MLP and with MD, for both LDL (a) and HDL (b) cases. GLE1 and GLE2 refer to the memory kernel approximations obtained respectively with 1 and 2 neurons in the hidden layer.

Appendix C: Numerical methods: GLE integration

The integration of the GLE dynamics is a not-trivial task for two reasons: first, the convolution integral depends on the full history of the observable, and second, the noise term correlations have to be consistent with the fluctuation dissipation theorem. Several approaches based on the introduction of a set of auxiliary variable have been proposed to face these issues, i.e. Refs [18, 23, 24]. In the present work, we take advantage of the exponential

structure of $\boldsymbol{\theta}(t)$ to implement an integration algorithm, which, for a scalar memory kernel, reduces to the one proposed in Ref. [24].

1. Convolution decomposition

The history-dependent convolution term is written as a sum of the additional variable vectors $\mathbf{Z}_k(t)$, whose components are $Z_{k,i}(t) = \int_0^t A_{k,ij} e^{B_{k,ij}(t-\tau)} \mathcal{O}_j(\tau) d\tau$. Applying Leibniz's integral rule, and taking advantage of the symmetry of the matrices \mathbf{B}_k it follows that the time evolution of $\mathbf{Z}_k(t)$ can be expressed as:

$$\dot{\mathbf{Z}}_k(t) = \mathbf{B}_k \mathbf{Z}_k(t) - \mathbf{A}_k \mathcal{O}(t) \quad (\text{C1})$$

Hence, the original GLE is rewritten in form of the equivalent system:

$$\begin{cases} \partial_t \mathcal{O}(t) = \mathcal{P}\mathcal{L}\mathcal{O} - \sum_k \mathbf{Z}_k(t) + \mathbf{R}(t) \\ \partial_t \mathbf{Z}_k(t) = \mathbf{B}_k \mathbf{Z}_k(t) + \mathbf{A}_k \mathcal{O}(t) \end{cases} \quad (\text{C2})$$

2. Random force decomposition

In this section we provide the theoretical derivation of the random force decomposition for a general real tensor function $\boldsymbol{\theta}(t)$. It is worth noticing that such formulation is valid for any form of the memory kernel, not just exponential ones. First, let us notice that, because of the symmetry between t and t' in the fluctuation dissipation theorem, $\boldsymbol{\theta}(t)$ is an even function of time, i.e. $\boldsymbol{\theta}(t) = \boldsymbol{\theta}(-t)$. We define the Fourier transform of $\boldsymbol{\theta}(t)$ as $\tilde{\boldsymbol{\theta}}(\omega) = \int_{-\infty}^{-\infty} \boldsymbol{\theta}(t) e^{-i\omega t} dt$. Since $\boldsymbol{\theta}(t)$ is real and even in time, also $\tilde{\boldsymbol{\theta}}(\omega)$ is real and even for real ω . It follows that both zeros and singular points of $\tilde{\boldsymbol{\theta}}(\omega)$ are symmetric with respect to both real and imaginary axis in the ω -plane. Then we introduce the function $\tilde{\chi}(\omega)$ given by

$$\tilde{\chi}(\omega) = \sum_k -i (\omega \mathbf{1} + i \mathbf{B}'_k)^{-1} \mathbf{b}_k \quad (\text{C3})$$

where the real matrices \mathbf{b}_k and \mathbf{B}'_k are such that:

$$\tilde{\boldsymbol{\theta}}(\omega) \langle \mathcal{O}, \mathcal{O} \rangle = 2 \tilde{\chi}(\omega) \tilde{\chi}^T(-\omega), \quad (\text{C4})$$

and the singular points of $\tilde{\chi}^{-1}(\omega)$ lie in the lower-half complex ω -plane. Moreover, we define the two matrices:

$$\tilde{\zeta}(\omega) = \tilde{\chi}^{-1}(\omega) = \sum_k (\omega \mathbf{1} + i\mathbf{B}'_k) (-i\mathbf{b}_k)^{-1}, \quad (\text{C5})$$

and

$$\tilde{\mathbf{k}}_k(\omega) = -i (\omega \mathbf{1} + i\mathbf{B}'_k)^{-1} \mathbf{b}_k \tilde{\zeta}(\omega), \quad (\text{C6})$$

and we denote their Fourier inverse transform with $\mathbf{h}(\mathbf{t})$ and $\mathbf{k}_k(t)$. Combining (C3), (C5) and (C6), it follows that:

$$\sum_k \tilde{\mathbf{k}}_k(\omega) = \mathbf{1} \quad (\text{C7})$$

or, equivalently,

$$\sum_k \mathbf{k}_k(t) = \mathbf{1}\delta(t). \quad (\text{C8})$$

Moreover (C6) can be rewritten as $(i\omega \mathbf{1} - \mathbf{B}'_k) \tilde{\mathbf{k}}_k(\omega) = \mathbf{b}_k \tilde{\zeta}(\omega)$, that in the time domain gives:

$$\frac{d}{dt} \mathbf{k}_k(t) - \mathbf{B}'_k \mathbf{k}_k(t) = \mathbf{b}_k \zeta(t) \quad (\text{C9})$$

Finally, the following vector variables are introduced:

$$\boldsymbol{\xi}(t) = \int_0^{+\infty} \zeta(t-t') \mathbf{R}(t') dt' \quad (\text{C10})$$

and

$$\mathbf{R}_k(t) = \int_0^{+\infty} \mathbf{k}_k(t-t') \mathbf{R}(t') dt'. \quad (\text{C11})$$

From (C11) and (C8) it follows that:

$$\sum_k \mathbf{R}_k(t) = \mathbf{R}(t), \quad (\text{C12})$$

while, combining (C11) and (C9)

$$\frac{d}{dt} \mathbf{R}_k(t) = \mathbf{B}'_k \mathbf{R}_k(t) + \mathbf{b}_k \boldsymbol{\xi}(t). \quad (\text{C13})$$

(C13) and (C12) are the main result of the section since they allow to express the correlated noise of the original GLE as a function of white noises $\xi(t)$.

In what follows, we discuss the properties of the stochastic process $\xi(t)$. First, since all the singularities of $\tilde{\zeta}(\omega) = \tilde{\chi}^{-1}(\omega)$ lie in the lower-half complex ω -plane, then for $\tau > 0$:

$$\begin{aligned}\zeta(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \tilde{\zeta}(\omega) e^{i\omega\tau} = \lim_{a \rightarrow \infty} \frac{1}{2\pi} \int_{-a}^a d\omega \tilde{\zeta}(\omega) e^{i\omega\tau} = \\ &= \frac{1}{2\pi} \oint_{C^+} d\omega \tilde{\zeta}(\omega) e^{i\omega\tau} - \lim_{a \rightarrow \infty} \frac{1}{2\pi} \int_{\text{arc}(a \rightarrow -a)} d\omega \tilde{\zeta}(\omega) e^{i\omega\tau} = \mathbf{0}\end{aligned}\quad (\text{C14})$$

where $\oint_{C^+} d\omega$ indicates the integral over a closed contour C^+ that goes along the real line from $-a$ to a and then along a semicircle centered at 0 from a to $-a$, while $\int_{\text{arc}(a \rightarrow -a)} d\omega$ is the integral along an arc centered at 0 from a to $-a$. Hence, for $t > 0$ we can write

$$\xi(t) = \int_0^{+\infty} \zeta(t-t') \mathbf{R}(t') dt' = \int_{-\infty}^{+\infty} \zeta(t-t') \mathbf{R}(t') dt' \quad (\text{C15})$$

Thus, the correlation function of $\xi(t)$ at t_1 and t_2 is given by:

$$\begin{aligned}\langle \xi(t_1) \xi^T(t_2) \rangle &= \\ &= \left\langle \int_{-\infty}^{+\infty} dt'_1 \int_{-\infty}^{+\infty} dt'_2 \zeta(t_1 - t'_1) \mathbf{R}(t'_1) [\zeta(t_2 - t'_2) \mathbf{R}(t'_2)]^T \right\rangle = \\ &= \int_{-\infty}^{+\infty} dt'_1 \int_{-\infty}^{+\infty} dt'_2 \zeta(t_1 - t'_1) \langle \mathbf{R}(t'_1) \mathbf{R}^T(t'_2) \rangle \zeta^T(t_2 - t'_2) = \\ &= \int_{-\infty}^{+\infty} dt'_1 \int_{-\infty}^{+\infty} dt'_2 \zeta(t_1 - t'_1) \boldsymbol{\theta}(t'_1 - t'_2) \langle \mathcal{O}, \mathcal{O} \rangle \zeta^T(t_2 - t'_2)\end{aligned}\quad (\text{C16})$$

where we used the fluctuation dissipation theorem. Using the definition of Fourier transform of $\boldsymbol{\theta}$, it follows

$$\begin{aligned}\langle \xi(t_1) \xi^T(t_2) \rangle &= \\ &= \int_{-\infty}^{+\infty} dt'_1 \int_{-\infty}^{+\infty} dt'_2 \zeta(t_1 - t'_1) \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \tilde{\boldsymbol{\theta}}(\omega) e^{i\omega(t'_1 - t'_2)} \\ \langle \mathcal{O}, \mathcal{O} \rangle \zeta^T(t_2 - t'_2) &= \\ &= \int_{-\infty}^{+\infty} dt'_1 \int_{-\infty}^{+\infty} dt'_2 \zeta(t_1 - t'_1) \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \tilde{\boldsymbol{\theta}}(\omega) e^{i\omega(t'_1 - t'_2)} \\ \langle \mathcal{O}, \mathcal{O} \rangle \zeta^T(t_2 - t'_2) e^{-i\omega(t_1 - t_2)} e^{i\omega(t_1 - t_2)} &= \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \int_{-\infty}^{+\infty} dt'_1 \zeta(t_1 - t'_1) e^{-i\omega(t_1 - t'_1)} \tilde{\boldsymbol{\theta}}(\omega) \langle \mathcal{O}, \mathcal{O} \rangle \\ &\quad \int_{-\infty}^{+\infty} dt'_2 \zeta^T(t_2 - t'_2) e^{i\omega(t'_2 - t'_1)} e^{i\omega(t_1 - t_2)}\end{aligned}\quad (\text{C17})$$

Applying the definition of Fourier transform of $\zeta(t)$, and taking advantage of (C4) and (C5), we finally obtain:

$$\langle \boldsymbol{\xi}(t_1)\boldsymbol{\xi}(t_2)^T \rangle = \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \tilde{\boldsymbol{\zeta}}(\omega) \tilde{\boldsymbol{\theta}}(\omega) \langle \mathcal{O}, \mathcal{O} \rangle \tilde{\boldsymbol{\zeta}}^T(-\omega) e^{i\omega(t_1-t_2)} = \quad (\text{C18})$$

$$= \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \tilde{\boldsymbol{\zeta}}(\omega) 2\tilde{\boldsymbol{\chi}}(\omega) \tilde{\boldsymbol{\chi}}(-\omega)^T \tilde{\boldsymbol{\zeta}}^T(-\omega) e^{i\omega(t_1-t_2)} = \quad (\text{C19})$$

$$= \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega 2\mathbf{1} e^{i\omega(t_1-t_2)} = 2\delta(t_1 - t_2). \quad (\text{C20})$$

It follows that $\boldsymbol{\xi}(t)$ is a delta correlated stochastic process.

Since we adopted a approximation of $\boldsymbol{\theta}(t)$ whose components are in the exponential form $\theta_{i,j}(t) = \sum_k A_{k,ij} e^{B_{k,ij}(t)}$, its Fourier transform is given by:

$$\tilde{\boldsymbol{\theta}}(\omega) = \sum_k [-i\mathbf{A}_k \oslash (\omega\mathbf{J} + i\mathbf{B}_k) + i\mathbf{A}_k \oslash (\omega\mathbf{J} - i\mathbf{B}_k)] \quad (\text{C21})$$

where \oslash indicates the Hadamard division and \mathbf{J} is the $n \times n$ matrix of ones. Now $\boldsymbol{\theta}(t)$ is a real and even function of t , then $\tilde{\boldsymbol{\theta}}(\omega)$ has to be real and even for real values of ω . As a consequence, the singular points of $\tilde{\boldsymbol{\theta}}(\omega)$ has to be symmetrical with respect to the real and imaginary axes, namely in the form of pairs $\pm i\mathbf{B}_k$. For the same reason, the roots of $\tilde{\boldsymbol{\theta}}(\omega)$ have to be symmetric with respect to the real and imaginary axes. Thus, putting (C21) into a common denominator, factorizing, and calling $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_n^*$ the conjugate matrices containing the zeros of the numerator, we obtain:

$$\tilde{\boldsymbol{\theta}}(\omega) = \mathbf{K} \odot \left(\prod_n (\omega\mathbf{J} - \boldsymbol{\beta}_n) \odot (\omega\mathbf{J} - \boldsymbol{\beta}_n^*) \right) \oslash \quad (\text{C22})$$

$$\left(\prod_k (\omega\mathbf{J} + i\mathbf{B}_k) \odot (\omega\mathbf{J} - i\mathbf{B}_k) \right) \quad (\text{C23})$$

where \odot is the Hadamard product, \mathbf{K} is a matrix of positive real numbers and it is assumed that $\text{Im}(\beta_n) > 0$ and $\text{Im}(\beta_n^*) < 0$. It is worth noticing that, since $\tilde{\boldsymbol{\theta}}(\omega)$ is non-negative, then \mathbf{K} contains only positive values[24]. Now we define the function $\tilde{\boldsymbol{\chi}}(\omega)$ as:

$$\begin{aligned} \tilde{\boldsymbol{\chi}}(\omega) &= \frac{\mathbf{K}^{1/2}}{\sqrt{2}} \odot \langle \mathcal{O}, \mathcal{O} \rangle^{\odot 1/2} \odot \prod_n i(\omega\mathbf{J} - \boldsymbol{\beta}_n^*) \oslash \prod_k i(\omega\mathbf{J} + i\mathbf{B}_k) = \\ &= \sum_k -i(\omega\mathbf{1} + i\mathbf{B}_k)^{-1} \mathbf{b}_k \end{aligned} \quad (\text{C24})$$

(C24) has to be solved to find the matrices \mathbf{b}_k .

In case of diagonal memory kernel matrix, \mathbf{b}_k can be easily found by solving the easier relation:

$$\begin{aligned} \frac{\mathbf{K}^{1/2}}{\sqrt{2}} \langle \mathcal{O}, \mathcal{O} \rangle^{1/2} \prod_n i(\omega \mathbf{1} - \boldsymbol{\beta}_n^*) \prod_k i(\omega \mathbf{1} + i\mathbf{B}_k)^{-1} = \\ = \sum_k -i(\omega \mathbf{1} + i\mathbf{B}_k)^{-1} \mathbf{b}_k \end{aligned} \quad (\text{C25})$$

Finally, for a one dimensional GLE the presented formulation reduces to the one derived by Kawai[24], therefore the coefficients b_k can be evaluated from the following relation:

$$\langle \mathcal{O}, \mathcal{O} \rangle A_k = -2b_k \sum_n \frac{b_n}{B_k + B_n} \quad (\text{C26})$$

obtained by (C3), (C4) and (C21).

3. Extended dynamics and integration algorithm

For a general $\boldsymbol{\theta}(t)$, the extended dynamics is then expressed as:

$$\begin{cases} \partial_t \mathcal{O}(t) = \mathcal{P}\mathcal{L}\mathcal{O} - \int_0^t \boldsymbol{\theta}(\tau) \mathcal{O}(t-\tau) d\tau + \sum_k \mathbf{R}_k(t) \\ \partial_t \mathbf{R}_k(t) = \mathbf{B}'_k \mathbf{R}_k(t) + \mathbf{b}_k \boldsymbol{\xi}(t) \end{cases} \quad (\text{C27})$$

where the convolution can be decomposed in different ways depending on the structure of $\boldsymbol{\theta}(t)$. In our case $\boldsymbol{\theta}(t)$ has an exponential form, thus $\mathbf{B}'_k = \mathbf{B}_k$ and the variables $\mathbf{S}_k(t) = -\mathbf{Z}_k(t) + \mathbf{R}_k(t)$ can be defined, so that the GLE can be rewritten in following form:

$$\begin{cases} \partial_t \mathcal{O}(t) \mathcal{O} = \mathbf{F}(\mathcal{O}(t)) + \sum_{k=1}^{N_n} \mathbf{S}_k(t) \\ \partial_t \mathbf{S}_k(t) = \mathbf{B}_k \mathbf{S}_k(t) - \mathbf{A}_k \mathcal{O}(t) + \mathbf{b}_k \boldsymbol{\xi}(t), \end{cases} \quad (\text{C28})$$

with $\mathbf{F}(\mathcal{O}(t)) = \mathcal{P}\mathcal{L}\mathcal{O}$ accounting for the conservative mean force contributions.

The numerical algorithm adopted to solve the system is the following splitting method, with Euler-Maruyama scheme for $S_k(t)$:

$$\mathcal{O}^{(n+1/2)} = \mathcal{O}^{(n)} + \frac{\Delta t}{2} F^c(\mathcal{O}^{(n)}) + \frac{\Delta t}{2} \sum_{k=1}^{N_n} S_k^{(n)}, \quad (\text{C29})$$

$$S_k^{(n+1)} = (1 + B_k \Delta t) S_k^{(n)} - A_k \mathcal{O}^{(n+1/2)} \Delta t + b_k \boldsymbol{\xi}_k^{(n)}, \quad (\text{C30})$$

$$\mathcal{O}^{(n+1)} = \mathcal{O}^{(n+1/2)} + \frac{\Delta t}{2} F^c(\mathcal{O}^{(n+1)}) + \frac{\Delta t}{2} \sum_{k=1}^{N_n} S_k^{(n+1)} \quad (\text{C31})$$

where $\xi_k^{(n)} \sim \mathcal{N}(0, 2\Delta t)$ are independent Gaussian distributed random values.

In order to test the numerical stochastic integrator, similarly to Ref. [23], we consider a one dimensional GLE with single exponential memory kernel and no conservative forces. In this specific case, the time correlation is analytically solvable. Thus, we compare the auto-correlation function computed numerically with the analytical one, which can be expressed as:

$$\frac{\langle \mathcal{O}(t)\mathcal{O}(0) \rangle}{\langle \mathcal{O}(0)\mathcal{O}(0) \rangle} = \begin{cases} e^{\frac{tB}{2}} (\cos(\Omega t) - \frac{B}{2\Omega} \sin(\Omega t)) & \Omega \neq 0, \\ e^{\frac{tB}{2}} (1 - \frac{Bt}{2}) & \Omega = 0, \end{cases} \quad (\text{C32})$$

where it is introduced the complex parameter $\Omega = \sqrt{A - B^2/4}$. Fig. 9 shows that the numerical integrator is able to accurately reproduce the analytical correlation in the under damped limit ($A = 1$ and $B = 1$), in the damped case ($A = 1$ and $B = -2$) and in the over damped limit ($A = 1$ and $B = -4$).

Appendix D: Single particle in a bath: simulations details and additional results

A target particle, with mass $m = 1$, immersed in a bath of identical particles with masses $m_b = 1$ is simulated. Two systems are studies: We simulated a low density limit (LDL) with 700 particles in total, while the high density limit (HDL) with 800 particles. The interaction between any two particles i and j is modelled by the Lennard-Jones potential:

$$v_{\text{LJ}}(\mathbf{r}_{ij}) = \begin{cases} 4\epsilon_{ij} [(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^6] & \text{if } r_{ij} \leq r_c, \\ 0 & \text{otherwise,} \end{cases} \quad (\text{D1})$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the particles, $\epsilon_{ij} = 1.0$ is the depth of the potential well, $\sigma_{ij} = 1.0$ is the finite atom-atom distance at which the potential is zero, and $r_c = 2.5$ is a cut-off radius. The simulation box dimensions are $10\sigma \times 10\sigma \times 10\sigma$, and periodic boundary conditions are imposed along x , y and z axes. A Nosé-Hoover thermostat is used to equilibrate the system at a reduced temperature $T = 1.1$. The time step adopted is $\Delta t = 10^{-3}$. The following procedure is followed to run the MD simulations. First, the bath particles are randomly generated inside the simulation box. Then, a minimization algorithm is employed to avoid overlaps between particles. Hence, a run of 10^5 time steps is used to equilibrate the system. Finally, data on forces and momenta are gathered over 10^5 time

steps. This process is repeated for 10^2 trajectories in order to enhance the accuracy of the correlations, and consequently, of the memory kernels.

In Figs 10(a-b), we report the mean square displacement $\langle (x(t) - x(0))^2 \rangle$ computed with MD, LE and GLE in the LDL and HDL cases. From the comparison, it emerges that both the Markovian and the non-Markovian coarse-graining are able to accurately reproduce the mean square displacement. Moreover, in the HDL case, GLE shows better performances with respect to LE. Figs 10(c-d) show the values of the adaptive learning rate η during the MLP learning process. The log-log plot highlights the wide range of η values, that spans up to 8 orders of magnitude. This variability exemplifies the advantages of an adaptive learning rate over a fixed one. The error (or cost function) evolution during the learning process is reported in Figs 10(e-f). The monotonically decreasing trend of C at some point shows a plateau, which corresponds to the end of the learning process.

The effectiveness and the limitations of the coarse-graining out of equilibrium is also tested, by analyzing the probability density function ρ . A target particle with zero initial position x and momentum p is immersed in an equilibrated bath of 699 particles identical to the one adopted in the LDL case at equilibrium. Then, 10^5 trajectories of the system relaxation to equilibrium are simulated. This relaxation corresponds to the evolution of a Dirac delta to the equilibrium distribution in the phase space. Similarly, the relaxation of ρ obtained by coarse-graining the bath with GLE and LE is followed. The comparison reported in Figs 11(a-f) shows that GLE, even if parametrized with a memory kernel evaluated in equilibrium conditions, significantly outperforms LE. As expected, at equilibrium the distributions obtained with MD, GLE and LE converge. During the relaxation, ρ relaxes faster for LE and GLE with respect to MD. A quantitative estimation of the accuracy of GLE in reproducing the density relaxation is provided by the mean square errors in position ϵ_q and momentum ϵ_p , shown in Figs 11(g-h). As expected, both errors are null at the beginning and, asymptotically, when the system reaches equilibrium. During the first instants of the relaxation, the error reaches a peak, whose value for GLE is lower than LE of about 50% and 35% if considering ϵ_q and ϵ_p , respectively.

TABLE I. Values of the interaction potentials parameters, adopted to simulate the particle chain.

Parameters	K_H	K_γ	K_ϕ	r_0	γ_0	$\epsilon_{i,j}$	$\sigma_{i,j}$
Values	100	10	10	1.5	109.5	1	1

Appendix E: Particle chain in a bath: simulations details

A chain of $N = 20$ particles in a Lennard-Jones bath is also simulated (Fig. 12). The chain particles interactions are modelled by the following multi-body Dreiding potential [25, 26]

$$v(\mathbf{r}_{i,j,k,l}) = v_{LJ}(\mathbf{r}_{ij}) + v_H(\mathbf{r}_{ij}) + v_\gamma(\mathbf{r}_{ijk}) + v_\phi(\mathbf{r}_{ijkl}). \quad (\text{E1})$$

Linear covalent bonds are modelled with the harmonic potential $v_H(\mathbf{r}_{ij}) = k_H(\mathbf{r}_{ij} - \mathbf{r}_0)^2$, where \mathbf{r}_0 is the equilibrium position and k_H is a positive constant. Similarly, angular covalent bonds are modelled by $v_\gamma(r_{ijk}) = k_\gamma(\gamma_{ijk} - \gamma_0)^2$, where γ_{ijk} is the angle in i formed by the particles i , j and k , γ_0 is the equilibrium angle and k_γ is a positive constant. Finally, we consider torsional (dihedral) bonds through the potential $v_\phi(r_{ijkl}) = k_\phi(1 + \cos(2\phi_{ijkl}))$, with ϕ_{ijkl} being the angle between the two planes defined by $\{\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k\}$ and $\{\mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l\}$ respectively, and k_ϕ being a positive parameter. Table I reports the values of the constant parameters characterizing the interactions adopted in the present work. The bath contains 69900 particles interacting with Lennard-Jones potential v_{LJ} . The simulation box measures $50\sigma \times 50\sigma \times 40\sigma$, and periodic boundary conditions are imposed along x , y and z axes. A Nosé-Hoover thermostat is used to equilibrate the system at a reduced temperature $T = 1.1$ with a time step $\Delta t = 10^{-2}$. The following procedure is followed to run the MD simulations. First, the bath particles are randomly generated inside the simulation box. Then, the chain particle are placed along a straight line, and a minimization algorithm is employed to avoid particle overlaps. Hence, a run of 1.5×10^5 time steps is used to equilibrate the system. Finally, data are gathered over 2×10^7 time steps.

Appendix F: GLE for time series

To model a general time series of an observable \mathcal{O} by means of a GLE with a zero Markovian contribution, the following conditions have to be satisfied:

- $\mathcal{O}(t) \sim \mathcal{N}(\mu, \sigma^2)$

TABLE II. Results of the augmented DickeyFuller(ADF) test for modified global temperature.

ADF Statistic:	-21.377945
p-value:	< 10^{-16}
lags:	54
Critical Values:	
1%:	-2.566
5%:	-1.941
10%:	-1.617

- $\langle \mathcal{O}(t) \rangle = 0 \quad \forall t$
- $\langle \mathcal{O}^2(t) \rangle = \sigma^2 \quad \forall t$
- $\langle \mathcal{O}(t)\mathcal{O}(t') \rangle = \langle \mathcal{O}(t-t')\mathcal{O}(0) \rangle \quad \forall t \geq t'$

If the original data of the \mathcal{O} presents non-stationary features, some techniques can be adopted to obtain stationarity.

1. Case 1: Global temperature dynamics

$T(t)$ reveals non-stationary features due to a long period increasing trend related to global warming. Hence, we first compute the long term dynamics T_y as an yearly moving average:

$$T_y(t) = \frac{1}{y} \sum_{i=t-y-1}^{t-1} T(i) \quad (\text{F1})$$

The observable of interest is then defined as $T_a(t) = T(t) - T_y(t)$, so that the corresponding time series is approximately stationary and $T_a(t)$ can modelled by the GLE $\partial_t T_a(t) = -\int_0^t \theta(t-\tau) T_a(\tau) d\tau + R(t)$.

In order to test the statistical properties of T_a , we adopted some qualitative and quantitative tests. Fig. 13(a) shows the QQ (quantile-quantile) plot, which compares the data distribution against the normal one for each quantile. It emerges that the time series data are well approximated by a normal distribution, especially in the theoretical quantile range $-3 < Q < 3$. Some tail effects are visible, but the overall agreement is quantitatively verified by the R-squared test, which gives a value $R^2 = 0.9955$.

TABLE III. Results of the augmented DickeyFuller(ADF) test for modified Nikkei index.

ADF Statistic:	-29.805726
p-value:	$< 10^{-16}$
lags:	10
Critical Values:	
1%:	-2.566
5%:	-1.941
10%:	-1.617

In order to test the stationarity of mean variance, and time correlations, we split the data in 5 windows. Fig. 13(b) shows that, assuming the stationarity of the series, we commit maximum errors for mean and standard deviation of 0.0183 and 0.0430, respectively. Moreover, as reported in Fig. 13(c), the maximum standard error between the windows time correlation and their mean is 0.0246.

In order to test the stationarity of the modified time series, also the augmented Dickey-Fuller (ADF) test is adopted[41]. ADF test is useful to establish if a unit root is present in the stochastic data series. Specifically, the null hypothesis of a unit root is rejected in favor of the stationary alternative if the test statistic is more negative than some critical values. The results of the ADF test reported in Table II allows us to reject the unit root hypothesis with a probability higher than 99%.

2. Case 2: Nikkei index

Similarly to many financial instruments, Nikkei index exhibits a non-stationary behavior in both mean and variance. To overcome this issue, we define the observable as $NI_a(t) = [NI(t) - NI_y(t)] / \sigma_y(t)$, with $NI_y(t)$ and $\sigma_y(t)$ being respectively the moving average and the moving standard deviation computed over a period $[t - y, t - 1]$ as:

$$NI_y(t) = \frac{1}{y} \sum_{i=t-y-1}^{t-1} NI(i) \quad (F2)$$

$$\sigma_y(t) = \sqrt{\frac{1}{y} \sum_{i=t-y-1}^{t-1} (NI(i) - NI_y(t))^2} \quad (F3)$$

The parameter y is properly chosen in order to obtain a stationary $NI_a(t)$; preliminary tests has shown that $y = 10$ days is an optimal value. Hence, we model the normalized stock price $NI_a(t)$ with the following non-Markovian model $\partial_t NI_a(t) = -\int_0^t \theta(t-\tau) NI_a(\tau)d\tau + R(t)$

Fig. 14(a) shows the QQ (quantile-quantile) plot. The time series distribution is well approximated with a normal distribution in the theoretical quantile range $-2.5 < Q < 2.5$. Heavy tails effects are present. This means that the Gaussian approximation, and consequently the GLE for $NI(t)$, remains valid as long as extreme market events, such as market crashes or crisis, are not considered. The overall agreement is quantitatively verified by the R-squared test, which gives $R^2 = 0.9894$.

In order to test the stationarity of mean variance, and time correlations, we split the data in 5 equally sized sets and, for each one, we analyze the statistical properties. Fig. 14(b) shows that, assuming the stationarity of the series, we commit maximum errors for mean and standard deviation of 0.2787 and 0.0234, respectively. Moreover, as reported in Fig. 14(c), the maximum standard error between the windows time correlation and their mean is 0.1082. The results of the ADF test reported in Table III allows us to reject the unit root hypothesis with a probability higher than 99%.

- [1] P. A. Clarkson and E. L. Mansfield, *Nonlinearity* **7**, 975 (1994).
- [2] A. J. Archer, *J. Phys. Condens. Matter* **18**, 5617 (2006).
- [3] R. Zwanzig, *J. Chem. Phys.* **33**, 1338 (1960).
- [4] H. Mori, *Prog. Theor. Phys.* **33**, 423 (1965).
- [5] R. Zwanzig, *J. Stat. Phys.* **9**, 215 (1973).
- [6] A. Chorin and P. Stinis, *Commun. Appl. Math. Comput. Sci.* **1**, 1 (2006).
- [7] T. Kinjo and S.-a. Hyodo, *Phys. Rev. E* **75**, 051109 (2007).
- [8] C. Hijón, P. Espa  ol, E. Vanden-Eijnden, and R. Delgado-Buscalioni, *Faraday Discuss.* **144**, 301 (2010).
- [9] S.-a. Hyodo, *Japan J. Appl. Ind. Math.* **28**, 69 (2011).
- [10] M. Chen, X. Li, and C. Liu, *J. Chem. Phys.* **141**, 064112 (2014).
- [11] J. Comer, J. C. Gumbart, J. H  nin, T. Leli  vre, A. Pohorille, and C. Chipot, *J. Phys. Chem. B* **119**, 1129 (2015).

- [12] T. Baştuğ, P.-C. Chen, S. M. Patra, and S. Kuyucak, *J. Chem. Phys.* **128**, 155104 (2008).
- [13] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, 2001).
- [14] A. J. Chorin, O. H. Hald, and R. Kupferman, *PNAS* **97**, 2968 (2000).
- [15] E. Darve, J. Solomon, and A. Kia, *PNAS* **106**, 10884 (2009).
- [16] A. Torres-Carabal, S. Herrera-Velarde, and R. Castañeda Priego, *Phys. Chem. Chem. Phys.* **17**, 19557 (2015).
- [17] O. F. Lange and H. Grubmüller, *J. Chem. Phys.* **124**, 214903 (2006).
- [18] H. Lei, N. A. Baker, and X. Li, *PNAS* **113**, 14183 (2016).
- [19] K. Hornik, M. Stinchcombe, and H. White, *Neural Netw.* **2**, 359 (1989).
- [20] K. Hornik, *Neural Netw.* **4**, 251 (1991).
- [21] M. A. Nielsen, *Neural Networks and Deep Learning* (Determination Press, 2015).
- [22] M. Riedmiller, *Comp. Stand. Inter.* **16**, 265 (1994).
- [23] A. D. Baczewski and S. D. Bond, *J. Chem. Phys.* **139**, 044107 (2013).
- [24] S. Kawai, *J. Chem. Phys.* **143**, 094101 (2015).
- [25] S. L. Mayo, B. D. Olafson, and W. A. Goddard, *J. Phys. Chem.* **94**, 8897 (1990).
- [26] D. Hossain, M. Tschopp, D. Ward, J. Bouvard, P. Wang, and M. Horstemeyer, *Polymer* **51**, 6071 (2010).
- [27] M. Bishop, M. H. Kalos, and H. L. Frisch, *J. Chem. Phys.* **70**, 1299 (1979).
- [28] D. I. Dimitrov, A. Milchev, K. Binder, L. I. Klushin, and A. M. Skvortsov, *J. Chem. Phys.* **128**, 234902 (2008).
- [29] P. Alaton, B. Djehiche, and D. Stillberger, *Appl. Math. Fin.* **9**, 1 (2002).
- [30] F. E. Benth and J. S. Benth, *Int. J. Stoch. Anal.* **2011**, 1 (2011).
- [31] E. Moreles and D. Martínez-López, *Atmósfira* **29**, 279 (2016).
- [32] Berkeley-Earth, “Experimental land-average temperature for period 1880–2014,” (2018), http://berkeleyearth.lbl.gov/auto/Global/Complete_TAVG_daily.txt.
- [33] R. Rohde, R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham, and S. Mosher, *Geoinfor. Geostat.: An Overview* **1** (2013), 10.4172/gigs.1000103.
- [34] E. Friis-Christensen and K. Lasses, *Science* **254**, 698 (1991).
- [35] M. Takahashi, *Fin. Eng. Japanese Markets* **3**, 87 (1996).
- [36] M. G. Lee, A. Oba, and H. Takayasu, “Parameter estimation of a generalized langevin equation of market price,” in *Empirical Science of Financial Fluctuations: The Advent of Econo-*

- physics*, edited by H. Takayasu (Springer Japan, Tokyo, 2002) pp. 260–270.
- [37] R. C. Merton, Rev. Econ. Stat. **51**, 247 (1969).
- [38] www.macrotrends.net, “Nikkei 225 index - 67 year historical chart,” (2018), <https://www.macrotrends.net/2593/nikkei-225-index-historical-chart-data>.
- [39] D. Givon, R. Kupferman, and A. Stuart, Nonlinearity **17**, R55 (2004).
- [40] G. A. Gottwald, D. T. Crommelin, and C. L. E. Franzke, “Stochastic climate theory,” in *Nonlinear and Stochastic Climate Dynamics*, edited by C. L. E. Franzke and T. J. O’Kane (Cambridge University Press, 2017) pp. 209–240.
- [41] D. A. Dickey and W. A. Fuller, J. Am. Stat. Assoc. **74**, 427 (1979).

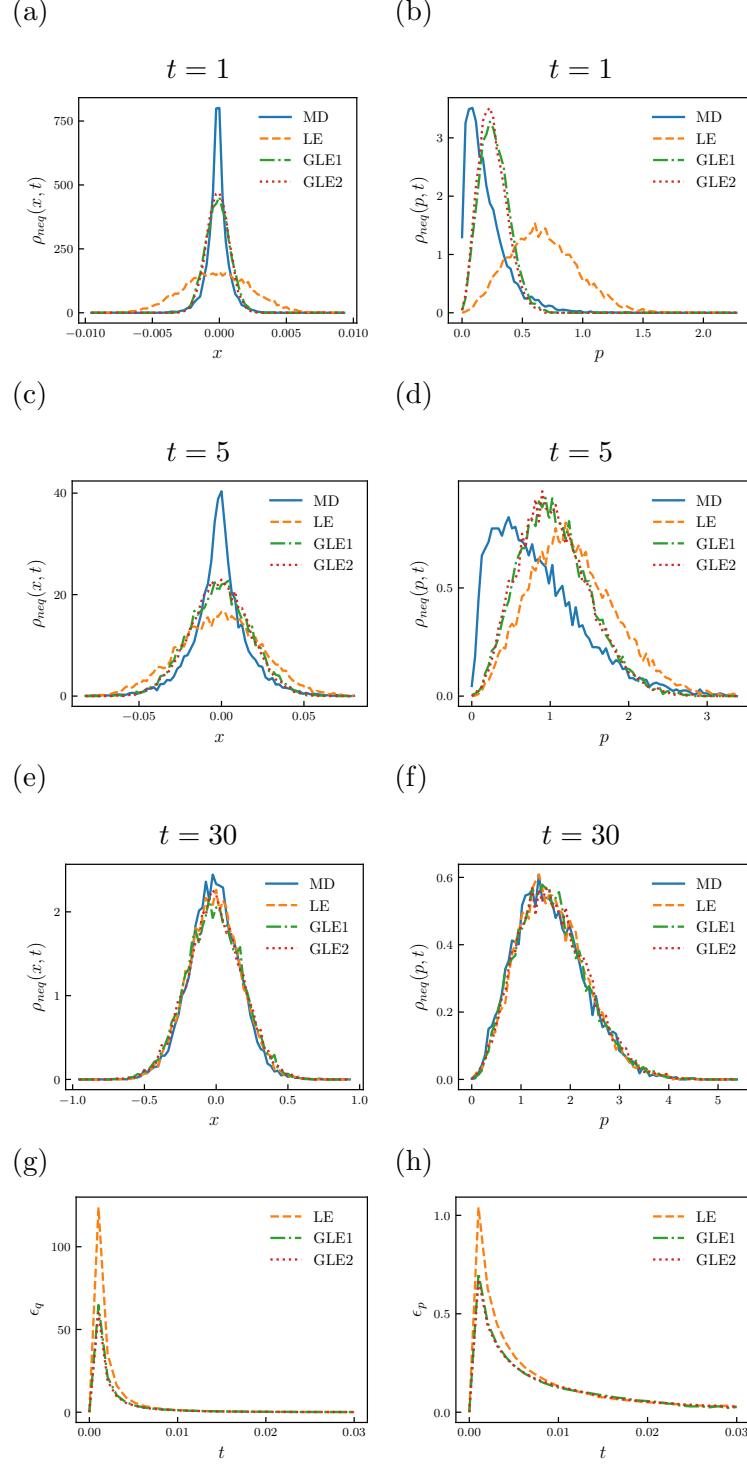


FIG. 11. Relaxation dynamics of position (a-c-e) and momentum (b-d-f) probability density function from Dirac delta to equilibrium condition computed with MD, LE and GLE over 10^4 trajectories. Corresponding mean square error of position (g) and momentum (h) probability density function in time.

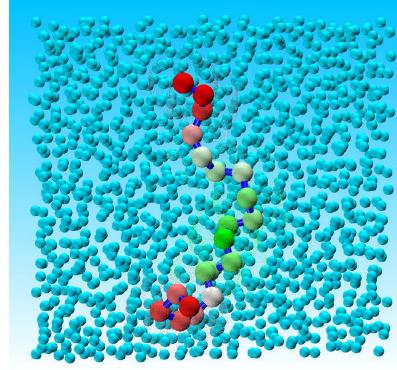


FIG. 12. Representation of the particle chain in the bath at equilibrium.

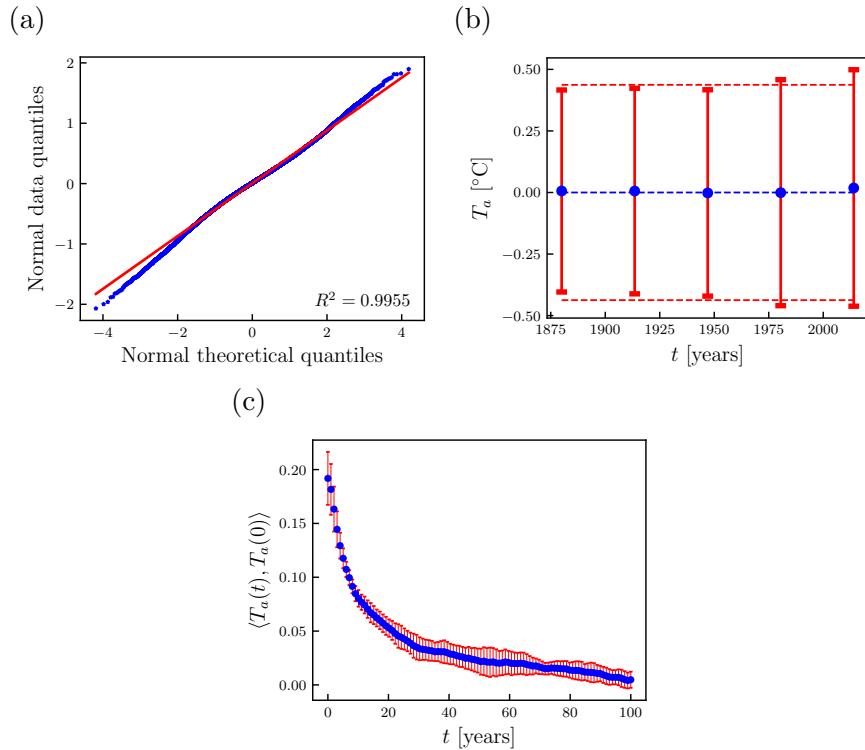


FIG. 13. (a) QQ-plot for $T_a(t) = T(t) - T_y(t)$. (b) Mean evaluated for 5 different data windows in time (blue dots) and corresponding standard deviations represented as red error bars. (c) Average time correlation function (blue dots) and standard error evaluated at each time from the time correlations of 5 different data windows.

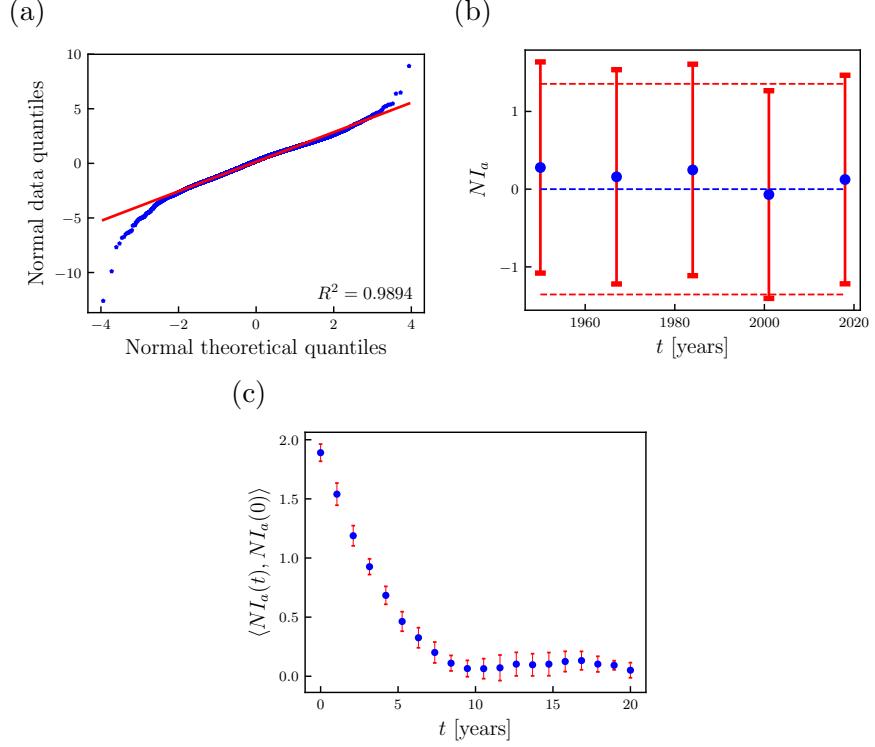


FIG. 14. (a) QQ-plot for $NI_a(t) = NI(t) - NI_y(t)$. (b) Mean evaluated for 5 different data windows in time (blue dots) and corresponding standard deviations represented as red error bars. (c) Average time correlation function (blue dots) and standard error evaluated at each time from the time correlations of 5 different data windows.