Data-driven reconstruction of a multivariate Langevin equation to model complex systems

Antonio Malpica-Morales [1,*] Miguel A. Durán-Olivencia [1,2,†] and Serafim Kalliadasis [1,‡]
[1]*Department of Chemical Engineering, Imperial College, London SW7 2AZ, United Kingdom*
[2]*Research, Vortico Tech, Málaga 29100, Spain*

(Received 13 September 2024; revised 6 May 2025; accepted 11 July 2025; published 11 August 2025)

Obtaining an accurate description of complex systems is challenging, particularly when their elements exhibit intricate interactions. We propose a data-driven multivariate Langevin equation (LE) to approximate real-world complex systems' observables. By reconstructing the drift and diffusion terms of the LE through a nonparametric technique following the definition of the Kramers-Moyal coefficients, our approach unravels the main features of a complex system without requiring *a priori* knowledge about the underlying governing mechanisms. We illustrate our framework's adaptability, reliability, and capability to extract pertinent information through three case studies. First, we benchmark the framework with a simple prototypical example from mechanics, a particle confined by a bistable potential energy well. We then turn to two more involved examples from financial markets, the electricity day-ahead prices and currency-exchange rates, where the nonparametric multivariate LE has not previously been applied. In all cases, our framework accurately identifies the equilibrium values, metastabilty regions, and distinct diffusion behaviors, in a functional agnostic manner, as opposed to price-equation models that require specific domain knowledge.

DOI: 10.1103/zncf-n4y3

I. INTRODUCTION

Complex systems are ubiquitous in nature, engineering, and technological processes. Their study often employs theoretical frameworks, mathematical models, and computational tools aiming to provide structured approaches for analyzing and interpreting intricate behaviors at both micro- and macroscopic scales. At the microscopic level, the complete and precise description of a complex system's constituents and their nontrivial interactions typically requires the resolution of a large number of degrees of freedom (DoF) and, inevitably, poses significant challenges in terms of computational and modeling tractability. On the other hand, at the macroscopic level, phenomenological continuum models for the time evolution of the system's observables are often employed. Such models provide effective descriptions aiming to retain the main effects at the macroscopic level. Macroscopic continuum modeling enables a functional and practical formulation of complex systems capturing their prevailing features [1]. The corresponding evolution equations include two key ingredients: nonlinearity and fluctuations.

Statistical mechanics bridges the micro- and macrodomains by developing mathematical routes to connect the system's DoF with the macroscopic observables. A long-standing result in this direction is the so-called projection-operator (PO) formalism [2,3]. Under certain assumptions and simplifications, the formalism establishes the relationship of the microscopic DoF with the time-evolution equations for the macroscopic observables [4]—and this without the need to know *a priori* the functional form that relates the micro- and macroscopic variables. Specifically, the PO formalism gives rise to formally exact evolution equations for the observable dynamics, which can be cast as a set of Langevin equations (LEs) [5].

An LE is typically composed of deterministic and stochastic terms whose interplay leads to a drift-diffusion stochastic process [6]. Specifically, an LE includes a term accounting for the dynamic variation of an observable, a mean force/drift term that informs about the expected change of the average value of the observable, and a diffusion process term that consists of a stochastic driving and a diffusion coefficient that controls the influence of the driving. Constructing an LE for a complex system not only provides a mathematical representation of its behavior but also facilitates understanding of its governing mechanisms; in particular, it enables quantification of the strength and interactions of the underlying forces between the constituent agents [5,7–9]. The simplicity of the LE in retaining only the most dominant effects, makes it a generic versatile model for numerical and mathematical scrutiny of complex systems [5,10].

From stochastic process theory, the drift and diffusion terms can be estimated from empirical records of the system's observables using the Kramers-Moyal (KM) coefficients, establishing in particular a one-to-one correspondence between drift and diffusion with the first and second KM coefficients, respectively [11]. Data-driven reconstruction of the

*Contact author: a.malpica-morales21@imperial.ac.uk
†Contact author: miguel@vortico.tech
‡Contact author: s.kalliadasis@imperial.ac.uk

KM coefficients has been extensively applied to canonical, prototypical, and theoretical models of complex systems. We now summarize previous efforts in this direction:

Prototypical complex systems: Univariate parametric/nonparametric methods. Data-driven reconstruction of the KM coefficients has been applied to prototypical models of complex systems [12–18], like a noisy genetic model or diffusion over bistable potentials. These applications utilize parametric methods, such as regression fitting [12,13,15,18], and maximum likelihood [14], or nonparametric methods, like kernel density estimation (KDE) [16,17]. The works in this category share the common characteristic of reducing the study of the complex system to a single observable, resulting in a univariate LE formulation that is straightforward to interpret and has relatively low computational requirements.

Prototypical complex systems: Multivariate parametric methods. In contrast, Ref. [19] lays the foundation for the multivariate approach but still within the parametric framework. By using an expansion of the KM coefficients together with polynomial and nonpolynomial fitting, the authors estimate the governing parameters of the drift and diffusion terms. Still, application of the methodology is confined to model systems with well-defined equations, such as Kuramoto-Sakaguchi oscillators, or the Lorenz system, which are normally derived from first principles following a bottom-up approach. The first step of the methodology is to solve the equations in order to generate the dataset used in the subsequent KM approximation. This process facilitates the direct comparison between the obtained KM coefficients and the known analytical expressions.

Real-world complex systems: Univariate parametric/nonparametric methods. Unfortunately, when dealing with real-world instances of complex systems, we often lack either precise theoretical models or the well-defined mathematical expressions for the constituent terms of such systems, which poses a great validation challenge. Examples of KM approximation for real-world complex systems span various fields [8,11], including neuroscience [9], wind power [20,21], financial markets [22], and cardiology [23] to name but a few. A common starting point in these studies is to assume an LE with unknown drift and diffusion terms as a model for the dynamics of the real-world system. To reconstruct the unknown drift and diffusion terms, empirical data obtained from the real-world complex system at hand is utilized in the KM approximation, following a top-down approach. The crucial point is that the studies in this category follow a univariate approach for modeling real-world complex systems encompassing either parametric or nonparametric methods in the KM estimation. Yet, actual complex systems inherently involve multiple interrelated state variables.

Real-world complex systems: A multivariate nonparametric method. To advance to the next level, the way forward is the development of a multivariate framework for describing real-world complex systems. This is precisely the framework we put forward. It entails a high-dimensional LE implying a transition from a one-dimensional stochastic process to a multidimensional stochastic one, where the number of interrelated observables and dependencies can easily be customized. A similar data-driven methodology was already formulated in the recent effort in Ref. [24] but was specific to the particular application, the power output of a wind turbine, while here the developed framework is tested across different application domains. Indeed, multivariate formulations constitute a challenge and several factors often preclude the generalization and expansion of existing univariate formulations to multivariate ones. The "curse of dimensionality" [25], i.e., the exponential growth in data requirements when adding extra dimensions, leads to sparse data regions in high-dimensional setups. Computational requirements also increase dramatically with dimensionality, as the number of terms to estimate rises. Moreover, thorough methodology adjustments and refinements are necessary as highlighted in Ref. [19], and pairwise interactions between state variables introduce additional complexity. For example, in a univariate formulation, the diffusion term is a single element, while in a multivariate approach, the diffusion term becomes a matrix that captures the covariance between the fluctuations of different state variables. Therefore, it requires careful estimation to ensure it is positive (semi-)definite, along with a detailed understanding of the strength and direction of each interaction. In this direction, a nonparametric multivariate approach, not attempted before, could be a powerful method for actual complex systems, overcoming the challenges mentioned earlier. Unlike parametric methods which rely on imposing certain functional forms to represent the behavior of a complex system which might not necessarily be true, nonparametric methods, which do not require any functional form to be imposed *a priori*, offer both the flexibility and adaptability required for a top-down approach to complex systems. In this sense, the observational data directly inform the KM reconstruction in contexts free of any idealizations and where the equations driving the complex system behavior are unknown and thus revealing the prevailing relationships and dynamics of the system conditioned on the historical available records of the system's state variables.

Specifically, we put forward a data-driven reconstruction of the KM coefficients based on KDE to approximate the drift and diffusion terms of the LE. Our overarching objective is to disentangle the stochastic behavior of the complex system's observables, without having to specify the exact mathematical expression for the components of the multidimensional LE (the KDE technique is nonparametric). We adopt a model paradigmatic system from physics, a classical particle in a bistable potential, to illustrate the new framework. This enables us to establish a first benchmark, providing a proof of concept for the framework and demonstrating its efficiency and robustness. We then apply the framework to two complex systems from the realm of financial markets: the electricity day-ahead market (rather topical in view of the energy crisis in different parts of the world as a result of certain geopolitical events) and the currency-exchange market. For these financial-markets examples, we reconstruct a multivariate LE in which each drift depends only on one component and each diffusion term on a pair of components, thereby omitting potential higher-order interactions. Nonetheless, as we shall see, this form will be sufficient to elucidate the fundamental principles driving the time evolution of the observed prices, provided that small local steps dominate the price changes.

The present study makes the following main contributions:

(1) Puts forward a data-driven framework whose cornerstone is a multivariate LE, providing a flexible and powerful tool for unravelling the intricate mechanisms governing real-world complex systems.

(2) The nonparametric multivariate LE is applied for the first time in financial markets, thus taking to the next level existing approaches based on univariate LE.

(3) Unlike traditional price-equation models in financial markets, our methodology does not require *a priori* specific domain knowledge.

Section II presents the methodology to approximate the drift-diffusion terms of the multivariate LE by means of the nonparametric KM estimation. The three case studies mentioned above are examined in Sec. III. Specifically, for the electricity day-ahead and the currency-exchange market, we start with a comprehensive review of different approaches, moving on to the results obtained using our framework and their detailed analysis, unravelling the prevailing dynamics for each system. Finally, a conclusion and discussion of open problems and possible future research avenues are offered in Sec. IV.

II. THEORY

The dynamical behavior of complex systems typically involves a large number of DoF but their governing equations often elude us. However, more often than not our focus is on a subset of DoF, or certain functions of them, commonly referred to as observables, which encapsulate the essential features of the system's behavior. And this is the main point of coarse graining: obtain an effective description that judiciously removes certain information, by, e.g., averaging out microscopic properties, and retains the main effects at the level of description we are interested in, typically the macroscopic level. For instance, in colloidal fluids, the positions and momenta of the colloidal particles are the primary DoF of interest, and these are only a small subset of the total DoF which also includes the positions and momenta of the bath particles [26–28]. Taking climate as an example, the average global temperature is influenced by multiple factors, such as atmosphere composition, solar radiation, and wind-driven ocean circulation. Yet, at least from the perspective of policymakers, the average global temperature itself represents a crucial observable of interest that encompasses climate conditions [29].

Projection-operator techniques offer a powerful means to bridge the gap between the microscopic dynamics of the underlying DoF and the macroscopic behavior captured by the observables. The principal requirement is that of separation of scales which enables us to split the dynamics into slow (resolved) and fast (unresolved) modes. [An extension to PO to include the renormalization group technique and stochastic mode reduction to obtain the statistical properties of the fast modes, is particularly suitable for dissipative systems, such as the generalized Kuramoto-Sivashinsky equation [30], a paradigmatic weakly nonlinear prototype for complex systems with instability-energy production, stability-energy dissipation, dispersion and nonlinearity.] By systematically coarse-graining the dynamics of the system, PO allows us to derive (finite-dimensional) effective equations of motion for the observables of interest and incorporate both deterministic forces and stochastic fluctuations. Under the Markovian approximation, which neglects memory effects, the effective equations that emerge from the derivation are LEs. They are compelling paradigmatic models describing the dynamic evolution of observables. Thus, with reference to colloidal fluids and climate mentioned earlier, both the positions of colloidal particles [26–28], and the value of the average global temperature [31], can be conceptualized as observables evolving according to a specific LE. This rationale provides a robust justification and strong basis for employing LEs as a generic model for the dynamics of complex systems, focusing on a subset of underlying DoF, the resolved or dominant modes.

Unfortunately, more often than not, the nature of the complex system being examined is such that is not always straightforward to identify these modes. They may remain elusive or poorly understood, rendering the direct derivation of LEs from first principles unfeasible. Even when the underlying governing equations of the complex system under consideration are known, the resulting LEs may defy analytical tractability, except in some limiting cases, posing significant obstacles to their practical application and implementation. A way forward is to approach the problem of low-dimensional representation of complex systems from a different angle. The focus can instead shift to the formulation of methodologies for disentangling the basic ingredients of LEs from empirical data, in particular, the drift and diffusion components. The development of such data-driven techniques is of utmost significance in the study of complex systems, as it would enable the modeling of a system's behavior without requiring to fully analyze and characterize its underlying interactions and dynamics. Indeed, data-driven techniques not only facilitate the calibration and validation of LE-based models against observational data but also pave the way for a deeper understanding of the emergent behavior exhibited by complex systems.

In what follows we introduce our proposed methodology for the reconstruction of the components of a multivariate LE from empirical records of a given real-world complex system. We first briefly outline the connection between the KM coefficients and the drift and diffusion terms of the LE. We then describe the KDE procedure used to obtain the KM coefficients. Figure 1 illustrates a diagrammatic representation of our framework, summarizing its main steps.

A. KM coefficients

Consider a set of l observables of the complex system being analyzed. Consider also the following multivariate LE:

$$dX_t^i = \mu^i(\mathbf{X}_t)dt + \sigma_{ij}(\mathbf{X}_t)dW_t^j, \tag{1}$$

where X^i is the ith observable, $i = \{1, \ldots, l\}$. Hence, the system follows a multidimensional stochastic process $\mathbf{X}_t = \{X_t^i\}$, with drift μ^i and diffusion σ_{ij}, $j = \{1, \ldots, l\}$. W_t^j is an uncorrelated Wiener process vector with Gaussian increments, i.e., $W_{t+dt}^j - W_t^j \sim \mathcal{N}(0, dt)$ and $\mathrm{Cov}(W_t^i, W_t^j) = 0$. Equation (1) inherently satisfies the Markov property, as it does not incorporate a memory kernel to account for past influences [5]. We adopt this simplified LE as a first step to uncover the

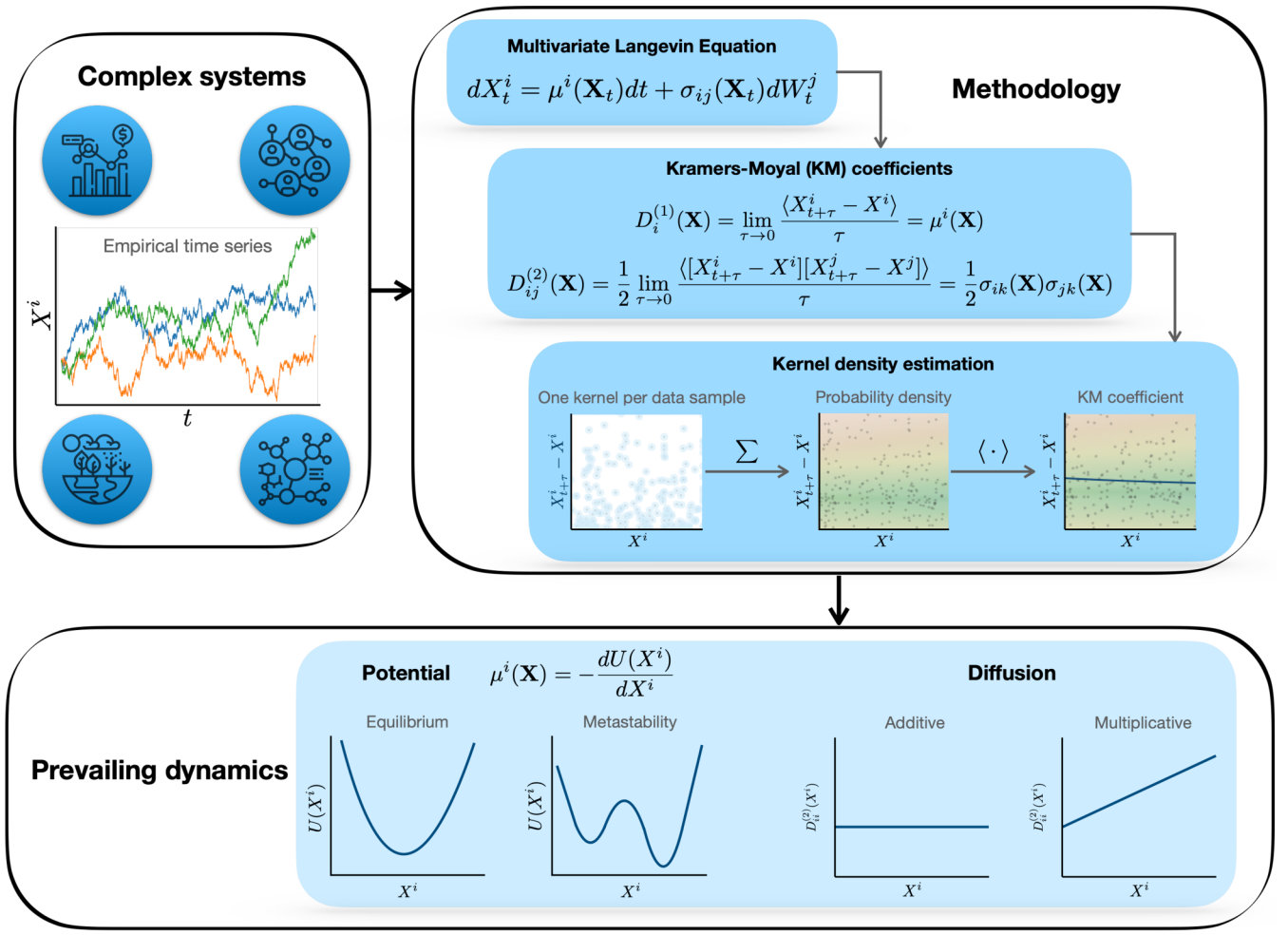

FIG. 1. Overview of our framework to model complex systems. A data-driven multivariate LE is at its core. Empirical times series of complex systems' observables are used to reconstruct the drift and diffusion terms of the multivariate LE. This reconstruction makes use of the KM coefficients, which are approximated using a KDE technique. Detailed analysis of the obtained drift and diffusion terms enables a comprehensive understanding of the prevailing mechanisms driving complex system dynamics.

dominant features of the observables. While this assumption may not fully capture the system's complexity, it provides a foundational approach for analyzing its stochastic nature.

The stochastic process $\mathbf{X}_t$ has an underlying probability density function (PDF), $\rho(\mathbf{X}, t)$ to observe a set of values $\mathbf{X} = \{X^i\}$ at time t. This PDF obeys the KM forward expansion [32]:

$$\frac{\partial \rho(\mathbf{X}, t)}{\partial t} = \sum_{n=1}^{\infty} \frac{(-\partial)^n}{\partial X^{k_1} \dots \partial X^{k_n}} D^{(n)}_{k_1,\dots,k_n}(\mathbf{X}, t)\rho(\mathbf{X}, t), \quad (2)$$

where $k_i \equiv i$, $D^{(n)}_{k_1,\dots,k_n}(\mathbf{X}, t)$ is the nth KM coefficient defined as follows:

$$D^{(n)}_{k_1,\dots,k_n}(\mathbf{X}, t) = \frac{1}{n!}\lim_{\tau\to 0}\frac{1}{\tau}\left\langle \left[X^{k_1}_{t+\tau} - X^{k_1}\right] \cdots \left[X^{k_n}_{t+\tau} - X^{k_n}\right] \right\rangle, \quad (3)$$

where $\langle\cdot\rangle$ is shorthand notation for the conditional expectation:

$$\langle X_{t+\tau} - x \rangle = \mathbb{E}[X_{t+\tau} - X_t \,|\, X_t = x]. \quad (4)$$

The first and second KM coefficients maintain the following relationship with the drift and diffusion terms of Eq. (1) (further details are given in Appendix A):

$$\begin{aligned} D_i^{(1)}(\mathbf{X}) &= \lim_{\tau\to 0} \frac{\langle X_{t+\tau}^i - X^i \rangle}{\tau} = \mu^i(\mathbf{X}) \\ D_{ij}^{(2)}(\mathbf{X}) &= \frac{1}{2}\lim_{\tau\to 0} \frac{\langle [X_{t+\tau}^i - X^i][X_{t+\tau}^j - X^j] \rangle}{\tau} \\ &= \frac{1}{2}\sigma_{ik}(\mathbf{X})\sigma_{jk}(\mathbf{X}). \end{aligned} \quad (5)$$

Since we impose a Gaussian model, Eq. (1), we do not consider higher-order KM coefficients for $n \geqslant 3$. When Gaussian behavior holds, Pawula's theorem states that higher-order KM coefficients vanish [32]. This is precisely why we truncate the infinite series in Eq. (2) at $n = 2$. As also highlighted in Refs. [32,33], truncating the infinite KM expansion of Eq. (2) at a finite order $n \geqslant 3$ may result in negative values for the PDF. A scenario that is numerically feasible but physically inconsistent. Thus, in the absence of further specifications about the underlying distribution governing the complex system, the Gaussian assumption is the best second-order and *a*

priori approximation of the stochastic process $\mathbf{X}_t$. This approach, retaining only the first two KM coefficients, ensures a well-posed Fokker-Planck equation, aligning with established practices [8]. Nevertheless, we will examine higher-order KM coefficients in our real-world examples to assess the accuracy and limitations of the Gaussian approximation.

B. KDE

From the results of the previous section, estimating the dynamics of the observables of the complex system under examination reduces to computing the KM coefficients of Eq. (5). Easy as it may seem at a conceptual level, it is a nontrivial task in practice, especially in the nonparametric and multivariate formulation of our framework, as already highlighted in Sec. I. The definitions in Eq. (5) involve the conditional expectation as in Eq. (4), which tacitly requires the PDF of the observables. Yet, we do not have access to this PDF. This prompts us to estimate the expected values in Eq. (5) using an empirical approach. While the so-called frequentist approach, a histogram-based method, may be conceptually simple and computationally straightforward, it is rather sensitive to the input data and does not provide an accurate solution for multidimensional problems. Moreover, this technique is limited by the binning criteria, leading to a piecewise constant approximation, which often lacks continuity between adjacent bins. Conversely, the KDE technique stands as a robust approach to better estimate the PDFs, allowing the calculation of the conditional expectation present in the KM coefficients [16]. It offers a smooth and continuous reconstruction of the PDFs, offering local adaptability while also handling sparse data more efficiently, as opposed to the frequentist approach [34], and succeeding in overcoming the "curse of dimensionality" highlighted in Sec. I.

We now introduce the procedure for applying the KDE technique to estimate the first and second KM coefficients of Eq. (5). First, we calculate the historical PDFs of $D_i^{(1)}$ and $D_{ij}^{(2)}$ to compute the conditional expectation of Eq. (5). These PDFs are the joint probabilities of the random variables:

$$\begin{aligned} P^i &= \left(P_1^i, P_2^i\right) = \left(X_t^i, X_{t+\tau}^i - X_t^i\right) \\ P^{ij} &= \left(P_1^{ij}, P_2^{ij}, P_3^{ij}\right) \\ &= \left(X_t^i, X_t^j, \tfrac{1}{2}\left(X_{t+\tau}^i - X_t^i\right)\left(X_{t+\tau}^j - X_t^j\right)\right). \end{aligned} \tag{6}$$

To approximate them we apply KDE, as described in Ref. [35], over the available data samples of the observables from the complex system under consideration:

$$\begin{aligned} \mathcal{P}^i &= \left\{\left(x_s^i,\ x_{s+\tau}^i - x_s^i\right)\right\} \\ \mathcal{P}^{ij} &= \left\{\left(x_s^i,\ x_s^j,\ \tfrac{1}{2}\left(x_{s+\tau}^i - x_s^i\right)\left(x_{s+\tau}^i - x_s^i\right)\right)\right\} \\ s &= 0, \ldots, N, \end{aligned} \tag{7}$$

where N is the total number of empirical records, with τ being lower bounded by the minimum sampling rate. KDE fits a series of predefined kernels, $K_H(\bullet)$, one per data sample of the datasets $\mathcal{P}^i$ and $\mathcal{P}^{ij}$, where H denotes the bandwidth of the kernel. We employ a Gaussian kernel and follow Scott's rule [36] for the bandwidth parameter:

$$H = \left(N^{-\frac{1}{d+4}}\right)\mathbf{I}_d, \tag{8}$$

where d indicates the input size of the kernel ($d = 2$ for P^i and $d = 3$ for P^{ij}), with $\mathbf{I}_d$ being a d-dimensional identity matrix.

Summing and normalizing all kernels for each $\mathcal{P}^i$ and $\mathcal{P}^{ij}$ dataset leads to an approximation of the PDFs, $\hat{P}^i \approx P^i$, $\hat{P}^{ij} \approx P^{ij}$. These approximated PDFs can be sampled through a set of collocation points defining a spatial mesh. The mesh resolution depends on the particular application being studied. Finally, we use the obtained samples from the spatial mesh to calculate $D_i^{(1)}$ and $D_{ij}^{(2)}$:

$$\begin{aligned} D_i^{(1)}(\mathbf{X}) &= \mathbb{E}\left[\hat{P}_2^i \middle| \hat{P}_1^i = X^i\right] \\ D_{ij}^{(2)}(\mathbf{X}) &= \mathbb{E}\left[\hat{P}_3^{ij} \middle| \hat{P}_1^{ij} = X^i, \hat{P}_2^{ij} = X^j\right]. \end{aligned} \tag{9}$$

It is important to note that the dimensionality of the complex system, l, and hence the LE used to describe the system, is completely independent of the dimensionality of the KDE. From Eq. (5), it is evident that $D_i^{(1)}$ depends only on X^i. The full l-dimensional state vector of the system, $\mathbf{X}$, is not involved in the calculation of the expected value that leads to $D_i^{(1)}$. Therefore, each element of the drift vector is reconstructed from a two-dimensional PDF, P^i in Eq. (6), which is approximated using a KDE with two input variables. The same procedure is replicated for $D_{i,j}^{(2)}$, resulting in a three-dimensional PDF, P^{ij} in Eq. (6), and three input variables for the KDE. At most, our methodology requires KDEs with two or three input variables, where the curse of dimensionality remains manageable.

III. APPLICATIONS

A. Single particle in a bistable potential

The first case study to verify our methodology is a canonical prototype from classical physics: the dynamics of a single one-dimensional Brownian particle confined in a bistable potential and in contact with a thermal bath. The time evolution of the particle position is governed by the following stochastic differential equation (in dimensionless form):

$$dX_t = -\frac{dU(X_t)}{dX_t}dt + \sqrt{2D}dW_t, \tag{10}$$

where X_t is the particle position at time t, U is a double-well potential energy function defined as $U(X) = (1/4)X^4 - (1/2)X^2$, D is the noise strength, and W_t is a Wiener process. We note that the force F acting on the particle is given by $F = -dU/dX$, i.e., it is conservative, or, equivalently, it depends at most on position and does not depend directly on other variables of motion, such as velocity, acceleration, or time. We also note that the equilibrium points of $U(X)$, obtained from $U' = 0$, are $X = -1$, $X = 0$, and $X = 1$, where $U'' > 0$, $U'' < 0$, and $U'' > 0$, respectively, corresponding to a local minimum, local maximum, and local minimum, respectively; hence $X = -1$ and $X = 1$ are stable equilibria and $X = 0$ is an unstable equilibrium.

Equation (10) is one of the simplest LEs. It can be viewed as a specific univariate version of the general LE in Eq. (1) with $i \equiv j = 1$ omitted for simplicity. And since the present

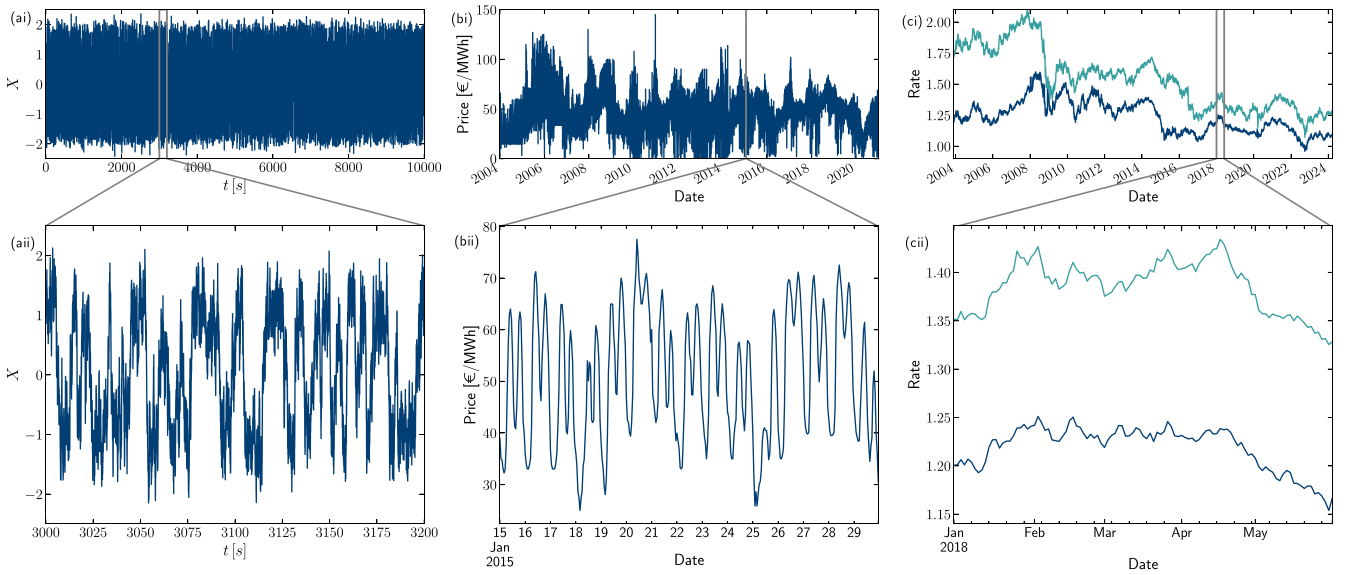

FIG. 2. Empirical time series of the three applications studied. Panels (ai) and (aii) correspond to the (one-dimensional) position of a single Brownian particle in a bistable potential. Panels (bi) and (bii) depict the time evolution of the Spanish electricity day-ahead price for the period spanning 2004–2020. Panels (ci) and (cii) illustrate the daily closing exchange rate for the EURUSD (blue line) and GBPUSD (green line) currency pairs between December 2003 and March 2024.

theoretical application is formally represented exactly by an LE, our data-driven framework should perform optimally when reconstructing the behavior of the system. In particular, the drift and diffusion terms of Eq. (10) have closed-form functional expressions, which should allow us to confirm and quantify the reliability and robustness of our framework.

We simulate one single trajectory for $D = 0.5$ using the Euler-Maruyama [37] scheme over $T = 10^4$ s with a time step $dt = 0.005$ s. Figure 2(ai) represents the simulated trajectory, consisting of a total of 2×10^6 data points. Figure 2(aii) displays a selected portion of the trajectory exhibiting random transitions between the two stable equilibrium points located at $X = -1$ and $X = 1$.

Figure 3 reports the results of applying our methodology to infer the LE terms for the problem of a particle in a bistable potential. We approximate the first and second KM coefficients using a mesh of $(10^2, 10^2)$ and $(10^2, 5 \times 10^2)$ collocation points, respectively. Overall, there is an excellent agreement between the actual drift, potential, and diffusion functions and the approximated ones. The first and second KM coefficients calculated from our framework almost perfectly replicate the drift and diffusion terms governing Eq. (10). The red dashed lines in Figs. 3(a) and 3(c), corresponding to the drift and diffusion estimations, respectively, are calculated as the expected values of the approximated PDFs obtained from the KDE technique in Eq. (9). The

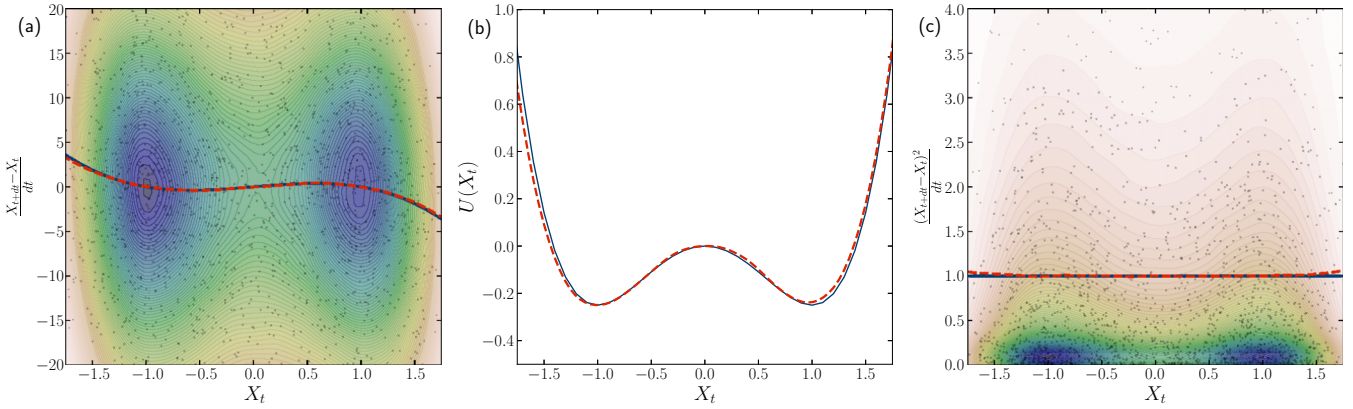

FIG. 3. LE terms' reconstruction of the particle in a bistable potential from the simulated trajectory. (a) Actual (blue solid line) and approximated drift (red dashed line) as functions of the particle position, X_t. (b) Actual (blue solid line) and approximated potential (red dashed line) as a function of X_t. (c) Actual (blue solid line) and approximated diffusion (red dashed line) as a function of X_t. Colored shadows in (a) and (c) represent the estimated PDF from KDE, ranging from shades of red (low probability) to shades of blue (higher probability). Black dots in (a) and (c) represent 2×10^3 random samples from the simulated trajectory computed as the particle position (horizontal axis) and instantaneous displacement (vertical axis) in (a) and squared value of instantaneous displacement (vertical axis) in (c).

estimated drift reveals a cubic function with roots at the three equilibrium points at $X = -1$, $X = 0$, and $X = 1$, as expected from the double-well potential $U(X)$ applied to the simulation. On the other hand, the estimated diffusion is a scalar with value of 1 which matches the diffusion coefficient used, $\sqrt{2D} = 1$, across 96% of the particle position domain, i.e., $X_t \in [-1.6, 1.6]$. It should be noted though that the accuracy of the estimated diffusion slightly deteriorates at the extremes of the domain, $X_t < -1.6$ and $X_t > 1.6$. This effect is primarily due to the limited available samples, comprising only 4% of the total data points, in these regions. Finally, the true potential applied, $U(X)$, and the one computed from the approximated drift [red dashed line in Fig. 3(b)] match extremely well. The minor discrepancies arise because of the mesh resolution employed, which affects the numerical integration of the potential and can be alleviated by refining the mesh.

B. Electricity day-ahead market

Having demonstrated the efficiency and robustness of our framework with a theoretical application, we take a step forward towards real-world complex systems. Among such systems, electricity markets are of particular interest due to the associated technological, social, and economic challenges [38] and especially their impact on the global economy. Understanding their governing mechanisms and the processes affecting electricity price formation and its time evolution is crucial for both market players, who need to adopt to ever changing customer preferences, and policymakers, who must create regulations to respond to uncertainty but also promote green economies [39]. This is especially true in the aftermath of the COVID-19 pandemic and the global energy crisis that led to surging and volatile energy prices.

Among electricity markets, the day-ahead market accounts for the bulk of the electricity system operation and electricity traded in competitive economies. In the day-ahead market, the market operator calls a daily group of auctions, with each auction corresponding to a certain time-block, gathering electricity generators and demand agents to determine the electricity prices. The agents then submit their respective offers for electricity energy across each auctioned time-block. In American and European markets, each time-block usually corresponds to 1 h of the following day. Thus, the outcome of the day-ahead market consists of a 24-dimensional array of prices reflecting the electricity value for each hour of the following day. Through these 24 separate time-block auctions, the day-ahead market entails an efficient mechanism to yield an affordable electricity price while ensuring the security of supply.

Understanding the behavior of the day-ahead prices is pivotal in the decision-making process of market players. However, the time evolution of the electricity prices is influenced by a myriad of both endogenous and exogenous variables, including weather conditions, commodities prices (coal, gas, uranium, etc.), and market players' bidding expectations. Consequently, the electricity-price time series exhibit various widely known features, like regular seasonal patterns, mean-reversion effects, and price spikes [40].

There is a wealth of literature dedicated to mathematical models and data-driven techniques aiming to obtain a meaningful representation of the time evolution of electricity prices. Within the domain of electricity day-ahead prices, the so-called spot prices, three primary techniques are often employed: stochastic (drift-diffusion) processes, autoregressive models, and machine learning. A comparison of electricity day-ahead price modeling from different studies can be found in Table I. As already highlighted in Sec. I, our framework offers the first multivariate and nonparametric formulation.

Within machine learning techniques, deep-learning models may provide accurate price forecasting [41,42]. However, the interpretability of the results obtained by deep-learning models is seriously hampered by their black-box nature, characterized by the absence of an equation-based formulation. This lack of interpretability is a major drawback of such models that can seriously hinder the understanding of the relationship between price fluctuations and the underlying market dynamics. Moreover, deep-learning models do not generate uncertainty quantification by default, which further complicates the analysis of spot prices.

Examples of stochastic processes range from basic ones, including Gaussian and Ornstein-Uhlenbeck (OU) [43], which belong to the family of drift-diffusion processes, Poisson [44] and regime-switching frameworks [45,46], that extend drift-diffusion processes to account for jumps, to more sophisticated ones, such as normal inverse Gaussian [47] and stochastic-volatility models [48,49], where the variance of a stochastic process is itself randomly distributed. Each one of these stochastic models can be seen as a particular case of an LE, with its own distinctive features which provide both advantages but also limitations in the electricity-price representation. But a common shortcoming for all previous stochastic models is that they typically consider the electricity price as a univariate stochastic process, which inherently assumes a higher degree of uncertainty as time evolves. This implies that later hours on the same day may have higher dispersion than earlier ones. However, this assumption may not necessarily hold true, as the 24 hourly prices are cleared simultaneously, with each one having the same level of uncertainty upfront. Consequently, the existing stochastic models based on drift-diffusion behavior are continually refined to account for an increasingly detailed electricity-price representation based on empirical observations. A procedure which is ineffective but also impractical. Even more so, such refinements still rely on prior assumptions to identify certain electricity-price features effectively. This leads to increasingly elaborate and complex models that require a comprehensive knowledge and expertise in the electricity-market domain often based on a constellation of experiences interpreted judiciously.

Autoregressive frameworks on the other hand, embody a simpler approach, modeling the electricity price as a linear combination of previous price values [52], called regressors. Both univariate and multivariate autoregressive methods are standard [53], where the latter is more in accordance with the character and idiosyncrasies of the day-ahead market [54], as each dimension is in one-to-one correspondence with the hourly price. However, linear regression, although attractive and certainly useful for computational scrutiny, could mask

TABLE I. Comparison of previous research works and this study on modeling electricity day-ahead markets. AR: autoregressive, ML: machine learning, SP: stochastic process.

Reference	Technique	Univariate	Multivariate	Parametric	Nonparametric	Uncertainty	Comment
[41]	ML	–	✓	–	✓	–	Support vector, random forest, and deep-learning models
[42]	ML	✓	✓	–	✓	–	Benchmark deep-learning models
[50]	ML	✓	✓	–	✓	–	Review of different techniques
	SP	✓	–	✓	–	✓	
	AR	✓	✓	✓	–	–	
[51]	SP	✓	–	✓	–	✓	Survey of SPs
[43]	SP	✓	–	✓	–	✓	Ornstein-Uhlenbeck process
[44]	SP	✓	–	✓	–	✓	Poisson process for price jumps
[45,46]	SP	✓	–	✓	–	✓	Regime-switching framework
[47]	SP	✓	–	✓	–	✓	Normal inverse Gaussian
[48,49]	SP	✓	–	✓	–	✓	Stochastic volatility
[52]	AR	✓	–	✓	–	–	Automatic selection regressors
[53]	AR	✓	✓	✓	–	–	Univariate and multivariate frameworks
[54]	AR	–	✓	✓	–	–	Functional time series
This study	SP	–	✓	–	✓	–	Drift-diffusion as KM coefficients using KDE

important nonlinear price dynamics which must be accounted for. Introducing higher-order terms and extensions in the linear regression function, as, e.g., done in Ref. [53], would be an ad hoc step, demanding not only an in-depth knowledge of electricity markets but also extensive technical knowledge in time-series modeling, and would not guarantee that critical nonlinear features can be captured.

The drawbacks and increasing complexity of existing electricity-price-equation models highlight the need for a framework based on a multivariate stochastic process (previous approaches considered multivariate frameworks as discussed in Sec. I but not in the context of electricity prices). This is precisely the modeling approach we follow here. Rather than constructing a new electricity-price model based on heuristics and/or ansatz solutions relying on previous studies, we put forward an adaptive framework with the capability to extract the drift-diffusion terms in an agnostic manner. A strict requirement is that the prevailing price features must be captured in an unbiased manner. The core of our framework is the multivariate LE in Eq. (1), where $i, j = \{1, \ldots, 24\}$ and the time evolution accounts for the changes on a daily scale. This implies that each dimension of our LE models the variability of one hourly price as well as its dependency on the remaining hours through the diffusion matrix.

We apply our methodology to the Spanish electricity day-ahead prices for the period 2004–2020. Figures 2(bi) and 2(bii) depict the time series used to uncover the drift and diffusion terms of the LE that models the time evolution of the spot prices. Both figures exhibit certain characteristic price features, such as the mean-reversion in Fig. 2(bi), or a daily and weekly seasonality in Fig. 2(bii). We assume that the drift function and the first KM coefficients are solely a function of the hourly price, X_t^i, calculated through a mesh resolution of $(10^3, 10^3)$ data points for each hour i. Conversely, for the sake of simplicity, we assume that the diffusion matrix and the second KM coefficients are state-independent, i.e., $D_{ij}^{(2)}(\mathbf{X}) = \mathbb{E}[\hat{P}_3^{ij}]$ in Eq. (9) for this particular application. Their computation involves a mesh resolution of $(10^2, 10^2, 5 \times 10^2)$ data points for each $D_{ij}^{(2)}$, except for the $i = j$ coefficients, where dimensionality reduction due to $\hat{P}_1^{ij} = \hat{P}_2^{ij}$ allows us to use a mesh of $(10^3, 10^3)$ data points.

The results of the first KM coefficient for each hour are included in Appendix B. A clearer understanding of the first KM coefficient, which equates the drift μ^i, emerges by associating the latter with a conservative force, hence the gradient of an effective potential V:

$$D_i^{(1)}(\mathbf{X}) = \mu^i(\mathbf{X}) \doteq -\frac{dV(X^i)}{dX^i}, \tag{11}$$

which follows the same rationale captured in the drift term of Eq. (10). The situation is then similar to Refs. [55,56] where a potential function was used to model paleoclimatic data and identify equilibrium points in a financial market, respectively. But here the potential function is not stipulated but data driven, extracted directly from the empirical data.

Figure 4 displays the effective potential function, for each hour (dark-blue line), obtained after integrating the estimated first KM coefficient. Overall, the potential function maintains a consistent parabolic shape across all hours but with slight deviations unique to each hour. In essence, the spot price for each hour exhibits a specific equilibrium value (minimum of the parabola) towards which the current spot-price value is pushed by an either positive or negative force. The magnitude of this force depends on the steepness of the potential curve for each hour, revealing that the intensity of dynamics varies for each hour. Besides, the equilibrium values vary across hours: valley hours (3h–6h) range from 30 to 35 € /MWh, mid-day hours (12h–16h) exhibit moderate values ranging from 46 to 48 € /MWh, and, finally, peak hours (20h–22h) reach 51 to 55 € /MWh. The drift approaches zero when the spot price for each hour is close to its respective equilibrium value. This indicates that the change of the hourly spot price is purely random when it revolves around the corresponding equilibrium value, reminiscent of a particle fluctuating in a harmonic potential. This effect implies that, under

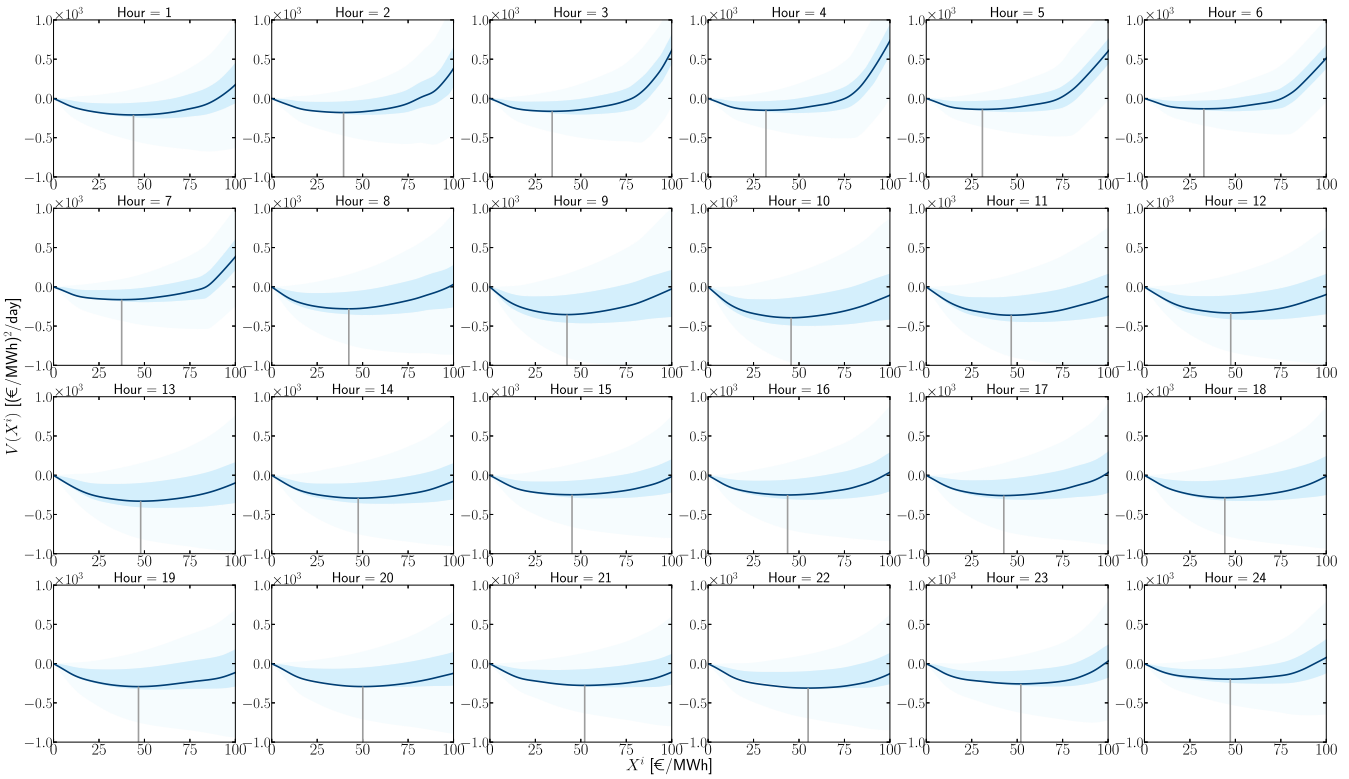

FIG. 4. Effective potential of each hour as a function of the spot price, X^i. Dark-blue lines represent the expected potential. Shaded areas enclose the [40, 60] (blue) and [20, 80] (light blue) percentile ranges.

stationary conditions assumed in the reconstruction of the KM coefficients (Appendix A), each hourly price in the Spanish day-ahead market represents a single stable market state.

The parabolic structure of the potential function aligns with the mean-reverting behavior observed in the spot-price dynamics. As discussed earlier, current stochastic models address this characteristic by incorporating an explicit biased term into their univariate mathematical formulation [43,44,47,48]. However, our approach reveals 24 different mean or equilibrium values, depending on the hour, rather than a single mean value that the univariate family models postulate. Furthermore, unlike OU-related processes [43,44,47,48], which typically impose a fixed mean-reverting intensity regardless of the spot-price value, our methodology reveals that the mean-reverting intensity is in fact price dependent.

Figure 5 depicts the second KM coefficient matrix. Given that this matrix follows the same pattern as the diffusion matrix, as evidenced by Eq. (5), it is vital to understand the effect of the diffusion matrix on the LE for the spot price. The block of hours between 8h–13h and 17h–19h in the main diagonal exhibits the largest second KM coefficients, with values between 56 and 74 $(€/\mathrm{MWh})^2/\mathrm{day}$. The cross-coefficients, i.e., entries with $i \neq j$, for this block of hours are positive, indicating that price variations occur in the same direction. Furthermore, substantial diffusion values are observed for the $i = \{9, 10, 11\}$ hours and $j = \{17, 18, 19\}$ hours groups. Conversely, clusters $i = \{1, 2, 3, 4, 5\}$ and $j = \{19, 20, 21, 22, 23, 24\}$—the upper right portion of the matrix—are slightly negative, $D^{(2)}_{2\,20} = -3.19$, $D^{(2)}_{4\,22} = -2.82$. This suggests that the interactions of the hourly price fluctuations within these clusters are dominated by low spot price fluctuations occurring in opposite directions. Higher-order KM coefficients are analyzed in Appendix C 1, revealing that truncating the expansion at second order, i.e., invoking the Gaussian drift-diffusion assumption, captures the essential picture, though finer details lie beyond the framework's scope

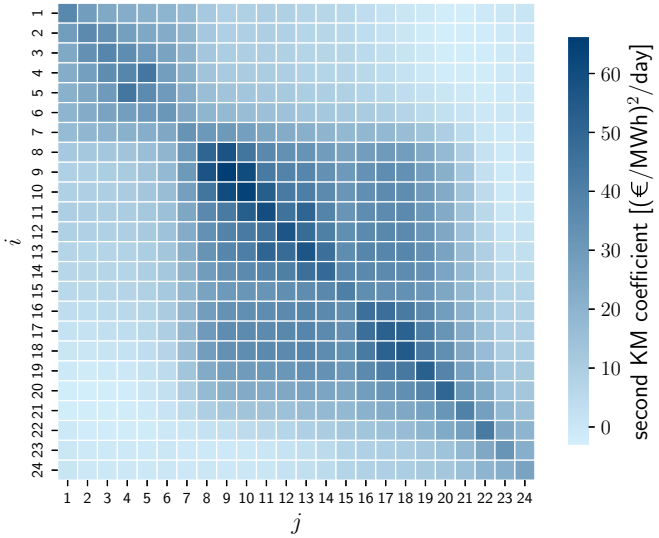

FIG. 5. Second KM coefficients matrix. Each cell $(i; j)$ contains the result of $D^{(2)}_{ij}$.

since the Spanish electricity day-ahead market exhibits large but infrequent price jumps.

C. Currency exchange market

In 1973, following the elimination of the US dollar's convertibility to gold under the gold standard, global currencies started to be traded in financial markets following a free-floating system [57]. Supply and demand of the currencies became the main drivers of their valuation. This led to the establishment of the foreign exchange (FX) market, a decentralized market operating 24 hours a day, providing currency valuation and liquidity worldwide. Since then, currency-exchange, or FX, rates have been recorded to measure the time evolution of global currencies relative to each other. Macroeconomic overviews, countries' monetary policies, as well as public and private growth strategies implemented by countries to revitalize their economies, heavily rely on the quantitative analysis of past, current, and future exchange rates. Their fluctuations affect a plethora of global trade and business-related decisions spanning both high-frequency trading and long-term investments.

A particular currency's value may change due to macroeconomic fundamentals, such as inflation (or deflation), economic growth (or recession) rate, and interest rates [58]. Simultaneously, both the process and outcome of exchanging assets under explicit trading mechanisms, the so-called market microstructure [59], also impact FX rates. At this microscopic level, information flow, market participants' sentiments, trading volumes, and price quotes all contribute to determining the short-run behavior of currency-exchange rates, of the order of days at most (the long-term behavior is affected by other factors such as banks, policies, or company financial statements) [60,61]. In fact, numerous variables at different spatiotemporal scales interact to distill the instantaneous currency-exchange valuations. Therefore, the FX market constitutes a paradigmatic real-world example of a complex system. By analyzing currency-exchange time series, we can extract insights into the dynamics governing the currency-exchange time evolution.

The peculiarities and unique qualities of financial markets, in general, and the FX market, in particular, have attracted considerable attention from the scientific community and sparked a wave of research leading to the proposal of a wide variety of modeling techniques in the field. Here we focus on analyzing the time evolution of the financial assets under consideration by adopting a top-down macroscopic approach as described in Secs. I and II.

Simple Brownian motion and its extension, geometric Brownian motion (GBM), remain the paradigmatic methods for modeling the time evolution of financial assets prices (or their returns) [62,63]. They constitute the foundation for the valuation of derivative instruments, such as options, swaps, or future contracts, for risk-modeling purposes. GBM can be viewed as a particular instance of a univariate LE in Eq. (1) ($i = 1$), with $\mu^1(X_t) = \mu X_t$, and $\sigma_1(X_t) = \sigma X_t$, where μ is the risk-free rate and σ is the historical volatility of the asset. The GBM and related methods are continually extended to provide a more consistent price representation according to empirical market data [64,65]. For example, GBM assumes a constant diffusion, σ, or a linear dependence with the asset-related variable being modelled. This entails a zeroth-order approximation of the true underlying dynamics generating the observed price (or returns) [65]. To capture a more detailed representation of the observed volatility, stochastic-volatility models propose a more elaborate state-dependent and even stochastic diffusion term [66]. By means of their random volatility, these models refine the variability of the second-order moment of the observed process, as opposed to the fixed volatility of the GBM. While GBM-related and stochastic-volatility models belong to the family of continuous-time models, there are also notable efforts in the discrete-time domain, with autoregressive conditional heteroskedasticity (ARCH) being the primary method [67,68].

Although the majority of the financial literature addresses stock and bond markets, applications to currency exchange are somewhat limited compared to the rest of financial assets, despite the fact that the FX market is the largest market in terms of trading volume. We attribute this imbalance to factors like the availability of historical data, investors' preferences, and company-performance metrics. Furthermore, the popular Black-Scholes-Merton model for option pricing [63], which relies on the GBM, is inadequate for FX options. References [69,70] outlined the first formulations for FX options, introducing adjustments to the classical GBM to model the time evolution of the currency rates. The main difference between the two works is that Ref. [69] considers a univariate GBM involving domestic and foreign exchange rates, whereas Ref. [70] proposes a two-dimensional GBM to model the domestic and foreign exchange rates in a stochastic manner. Extensions of stochastic-volatility models to address FX options can also be found in Refs. [71,72]. Additionally, Refs. [73,74] present autoregressive models specifically focused on currency-exchange rates, with the objective to uncover a common stochastic trend driving the valuation of multiple currencies simultaneously.

The construction of GBM-related and stochastic-volatility models requires notable knowledge of financial markets, including experience with financial instrument analysis and economic trends. For example, Ref. [75] proposes a phenomenological univariate LE to model stock prices from the S&P500. It is derived from *a priori* assumptions about the market microstructure, which impose multiple parameters into the LE that are calibrated to historical prices. Through these parameters, several desired effects arise in the price representation but artificially. On the other hand, Ref. [65] extends the GBM by incorporating two additional parameters into the drift-diffusion coefficients that encapsulate the market microstructure and the inflow/outflow of money, thus considering the financial market as an open system. This results in a univariate equation that has been extensively analyzed from a statistical mechanics point of view, aiming to uncover market states and associated equilibrium points.

The investigation of asset prices and financial instruments through the lens of statistical mechanics has yielded significant research outcomes. The corresponding studies share the common feature of not requiring in-depth knowledge of finance or market microstructure. Instead, they treat the observed time series for the prices as a single trajectory generated primarily by a drift-diffusion process. Generally, they

TABLE II. Comparison of previous research works and this study on modeling the time evolution of financial assets.

Reference	Asset type	Method	Univariate	Multivariate	Parametric	Nonparametric	Comment
[63]	Stocks	GBM	✓	–	✓	–	Application to options pricing
[64]	Stocks	GBM	✓	–	✓	–	Extension with memory kernel
[65]	Stocks	GBM	✓	–	✓	–	Market microstructure to uncover market states and equilibrium points
[66]	Bonds & FX	Stochastic volatility	✓	–	✓	–	Stochastic diffusion term for options pricing
[67,68]	Stocks, bonds & currencies	ARCH	✓	✓	✓	✓	Review of ARCH models in finance
[69]	FX	GBM	✓	–	✓	–	Options with constant drift-diffusion for domestic and foreign rates
[70]	FX	GBM	–	✓	✓	–	Options for domestic and foreign rates (2 dimensions)
[71,72]	FX	Stochastic volatility	–	✓	✓	–	Extensions of Ref. [66] for multiple currencies (>2 dimensions)
[73,74]	FX	ARCH	✓	✓	✓	–	Common stochastic trend for several currencies
[75]	Stocks	LE	✓	–	✓	–	Phenomenological LE from market microstructure
[76]	Stocks	LE	✓	–	✓	–	Price modeling with polynomial fitting for drift term
[22]	Stocks	LE	✓	–	✓	–	Price increments modeling with polynomial fitting for drift-diffusion terms
[77]	FX	LE	✓	–	✓	–	Rates increments modeling with polynomial fitting for drift-diffusion terms
[78]	FX	LE	✓	–	✓	–	Log-returns modeling with polynomial fitting for drift-diffusion terms
[79]	Stocks	LE	✓	–	–	✓	Cross-correlation modeling with KDE for drift-diffusion terms
[56]	Stocks	LE	✓	–	–	✓	Cross-correlation modeling with binning for drift-diffusion terms
This study	FX	LE	–	✓	–	✓	Rates modeling with KDE for drift-diffusion terms

make use of the KM coefficients to reconstruct the drift and diffusion terms of a univariate LE, which in turn is used to reproduce stylized facts in the observed financial descriptors under consideration: price [76], price increment [22,77], log-return [78], and cross-correlation [56,79]. At the same time, the obtained LE enables the examination of multiple interesting concepts, including postulating a polynomial form for the potential energy function, hence drift, to uncover the equilibrium points [76] (while here the potential is obtained directly from the data as part of our data-driven framework), the time evolution of the potential under different market regimes [56], the notion of market states and transitions between them [79], the realization that the data indicates the presence of multiplicative noise [22], as well as evidence of Markov property in the price increments over different timescales [77].

Here we model the variability of currency-exchange rates without relying on any financial-related assumptions and by adopting a multivariate framework. Our aim is to extract information about the stochastic process governing FX markets by employing a data-driven framework. While the KDE technique has already been applied to drift-diffusion equations in stock markets [79], and parametric techniques, such as polynomial fitting [76–78] have been used to obtain returns and prices in FX markets, utilizing multidimensional stochastic processes to analyze the time evolution of currency-exchange rates (instead of the returns) together with nonparametric reconstruction of their drift-diffusion terms, have not yet been explored, as detailed in Table II, and highlighted in Sec. I. This exploration is precisely the scope of the present case study.

We apply our methodology to two of the most traded currency pairs worldwide, the Euro (EUR) and British pound (GBP), against the US dollar (USD), denoted as EURUSD and GBPUSD, respectively. We utilize the daily closing exchange rates of the EURUSD and GBPUSD from December 2003 to March 2024, depicted in Fig. 2(ci) and 2(cii). Our objective is to use these historical rates to approximate a two-dimensional LE, $i, j = \{1, 2\}$ in Eq. (1). This LE provides an effective equation that reveals the prevailing dynamics influencing these currencies. For this purpose we employ the same drift-diffusion reconstruction procedure developed for the previous applications. We assume that the first and second KM coefficients are given in Eq. (9). For the mesh resolution, we apply a grid of $(10^3, 10^3)$ data points for $D_i^{(1)}$ and $D_{ij}^{(2)}$, $i = j$. Conversely, for $D_{ij}^{(2)}$, $i \neq j$, the grid contains $(5 \times 10^2, 5 \times 10^2, 5 \times 10^2)$ data points.

Figure 6 presents the first KM coefficients and the reconstructed potential for EURUSD and GBPUSD. We observe multiple equilibrium values for both currencies

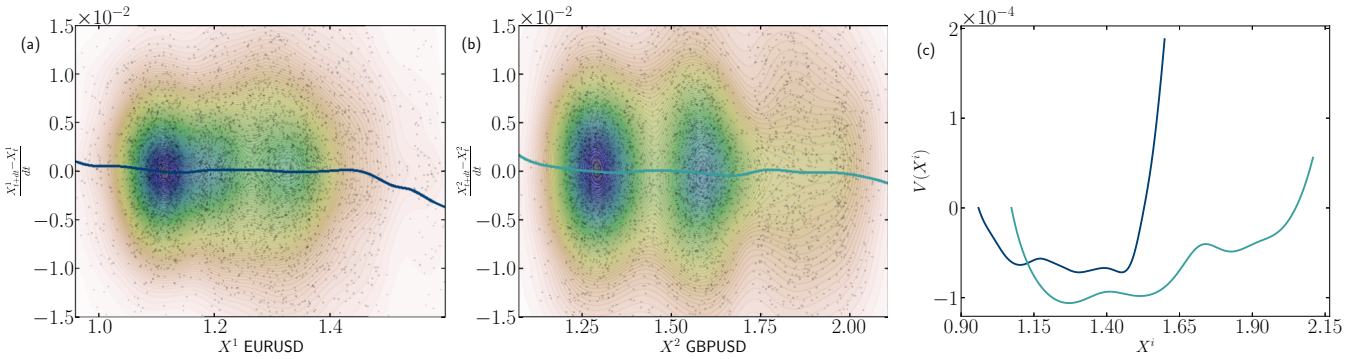

FIG. 6. Drift and potential approximation of the LE that models currency exchange rates. Colored shadows in (a) and (b) represent the estimated PDF from KDE, ranging from red (low probability) to blue (higher probability). Black dots in (a) and (b) represent empirical samples for the currency rate (horizontal axis) and its variation with respect to next day (vertical axis). Blue lines in (a) and (c) are the approximated drift and potential, respectively, for the EURUSD. Green lines in (b) and (c) are the approximated drift and potential, respectively, for the GBPUSD.

in (a) and (b), coinciding with the different minima of the approximated potentials in (c). These values indicate exchange rate regions where the currency tends to stay but fluctuates around these regions. However, the amount of time the exchange rate spends in these regions varies significantly, as indicated by the trajectories in Fig. 2(ci). The profile of the potentials evidences this phenomenon, suggesting metastability in currency-exchange rates, as opposed to the first case study with a particle in a bistable potential with coexistence of two equally stable equilibrium states. But metastability here is weak and the direction of the system is determined by the strength of fluctuations as discussed later. For example, the GBPUSD potential exhibits a minimum around $X^2 \approx 1.55$; however, it also exhibits a more stable point around $X^2 \approx 1.25$, a region of "least energy." This implies that, based on historical values, if the GBPUSD reaches the 1.55 region, then it would likely return to the 1.25 region after some time. This return journey is facilitated by market fluctuations, which push the currency towards the global minimum of the potential function in a stochastic manner.

Surprisingly, the global minimum of the potential function for the EURUSD is located at $X^1 \approx 1.45$, with a potential value almost similar to the next minimum at $X^1 \approx 1.32$. Although the former rate region is rarely explored by EURUSD in the empirical time series, we observe in Fig. 2(ci) that the EURUSD (blue line) spends a period of several years (2007–2012) fluctuating around this region. In fact, this was the period with the largest EURUSD volatility. And precisely because of this large volatility, the EURUSD abandoned this stable area in favour of other metastable regions ($X^1 < 1.4$).

As of March 2024, both currencies exhibit exchange rates around the leftmost minimum of their respective potential functions, i.e., $X^1 \approx 1.1$ and $X^2 \approx 1.25$. We would then expect both currencies to oscillate around these values until a large fluctuation drives them away to the next potential minimum. For that to occur, the currencies must exceed the energetic barrier that separates the current minima that revolve around from the nearest saddle points, $X^1 \approx 1.18$ and $X^2 \approx 1.4$.

We now turn our attention to the second KM coefficients which constitute the diffusion matrix and drive the fluctuations. The reconstruction of the coefficients is shown in Fig. 7. Overall, all elements of the second KM coefficient matrix are almost constant and one order of magnitude below the first KM coefficient (drift function). There are minor oscillations for $D^{(2)}_{22}$ along the X^2 domain [Fig. 7(c)], and for $D^{(2)}_{12}$ through the $\{X^1, X^2\}$ domain [Fig. 7(b)]. However, these variations are of $O(10^{-5})$. On the other hand, $D^{(2)}_{11}$ presents a notable variation around $X^1 \approx 1.53$, coinciding with the period of the largest EURUSD volatility. This suggests the presence of a subtle multiplicative noise for the EURUSD: In the neighbourhood of the global minimum of the potential energy function, $X^1 \approx 1.45$, the fluctuations are strong and the system drifts to the left of the global minimum and lower currency values. Despite this large deviation of $D^{(2)}_{11}$ occurring in a narrow region of X^1, $D^{(2)}_{ij}(\mathbf{X})$ could be approximated as $D^{(2)}_{ij}(\mathbf{X}) \approx \overline{D}_{ij}$, $\forall\, i, j = \{1, 2\}$. Therefore, one could assume from the outset a constant diffusion matrix as we did for the electricity day-ahead prices in Sec. III B; this, however, would be at the expense of missing out on subtle details of the system, such as multiplicative noise. Higher-order KM coefficients are analyzed in Appendix C 2, confirming that the Gaussian drift-diffusion assumption accurately captures the currency-exchange rates' dynamics.

IV. CLOSING REMARKS AND DISCUSSION

We have formulated a data-driven framework to analyze the stochastic behavior of real-world complex systems. The foundation of our framework consists of a multivariate LE with flexible drift and diffusion terms, eliminating the need for detailed prior understanding of the system being analyzed. Such flexibility relies on the relationship between the drift-diffusion terms and the KM coefficients. The estimation then of the KM coefficients through the KDE, a nonparametric technique, guarantees a high level of adaptability and intricate functional representation.

We have exemplified the efficiency and robustness of our framework through a number of case studies. We first looked at a simple prototypical example from classical mechanics, namely a particle in a bistable potential energy function,

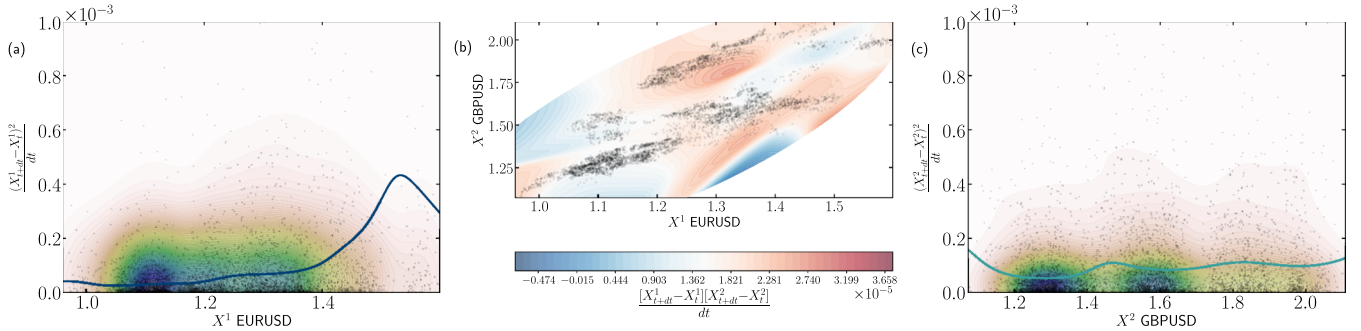

FIG. 7. Reconstruction of the elements of the matrix obtained by multiplying by 2 the second KM coefficient matrix in the LE that models currency exchange rates. The blue line in (a) and the green line in (c) indicate the mean value of the squared daily rate differences for the EURUSD, X^1, and GBPUSD, X^2, respectively. These are proportional to $D^{(2)}_{11}(\mathbf{X})$ and $D^{(2)}_{22}(\mathbf{X})$, as can be seen in Eq. (5). Colored values in (b) represent the multiplication of daily rate differences for X^1 and X^2 simultaneously, which in turn are proportional to $D^{(2)}_{12}(\mathbf{X})$. Colored shadows in (a) and (c) represent the estimated PDF from the KDE technique, ranging from red (low probability) to blue (higher probability). Black dots in (a) and (c) represent empirical samples for the currency rates (horizontal axis) and squared value of the variation with respect to next day (vertical axis). Black dots in (b) indicate empirical samples for both currencies rates (horizontal and vertical axes) simultaneously.

successfully reproducing the dynamics, and, in particular, reconstructing the metastability effect associated with this potential. We then moved on to two real-world examples from financial markets: the Spanish electricity-day ahead prices and the EURUSD and GBPUSD currency-exchange rates. While univariate formulations, such as OU- and GBM-related models, dominate standard practice in these application domains, our nonparametric multivariate LE represents an unprecedented approach, allowing us to uncover new patterns and insights. Its multivariate nature can extract subtle interactions between variables of the system that would otherwise remain elusive in univariate settings. Moreover, the nonparametric technique adopted ensures domain-agnostic applicability and low bias, removing the need for prior restrictive assumptions like the utilization of the risk-free rate in the drift term for financial assets modeling, while simultaneously revealing nontrivial characteristics in the drift-diffusion terms.

By solely relying on historical prices from the examined markets, our methodology has successfully reconstructed the deterministic and stochastic terms of the multivariate governing LE. This LE not only provides a versatile mathematical description of the time evolution of the considered complex system but also sheds light on the underlying mechanisms driving the observables' fluctuations. In particular, the electricity day-ahead LE helped us to identify the existence of different hourly equilibrium prices and significant correlations in the price movements between distinct groups of hours. Conversely, the currency pairs exhibited multiple equilibrium points and mild metastability as well as slight nonlinear diffusion with weak interactions between currencies.

Through these specific applications, we have also established our framework's capability to model various complexity landscapes, ranging from a univariate LE (classical particle) to a multivariate LE with state-dependent drift and constant diffusion (electricity market) to a multivariate LE with both state-dependent drift and diffusion terms (currency-exchange rates). The complementary analysis of the higher-order KM coefficients underpins the suitability of our framework. Currency-exchange rates can be modelled effectively as a drift-diffusion process with Gaussian noise, whereas the Spanish electricity day-ahead prices display more involved dynamics in which sparse price spikes and/or non-Gaussian behavior sometimes violate this modeling assumption. For such excursions in the system's phase space, large-deviation theory may be a useful complement to our modeling framework [80]. However, when no additional system details are available, the second-order KM truncation remains the most practical proxy in both scenarios, ensuring a well-posed PDF for the system's observables. In all cases, the thrust of our framework is to achieve maximum interpretability with minimal assumptions and to extract maximum information from the available time series of the complex system under consideration.

Future developments may exploit the multivariate interactions and dependencies that we omitted in our study. Considering drift, diffusion, and higher-order KM coefficients as functions of the full state vector, rather than of one or two components of the system, could enrich the insights gained from our framework. Additionally, the framework can be extended using tools and concepts from the statistical-mechanics field. In this direction, there is a number of interesting questions related to the analysis presented here. For instance, quantifying the escape rates associated with the metastability observed in the considered examples or verifying the variability in the dynamics over different timescales, e.g., from daily to yearly fluctuations. At the same time, it would be worthwhile to relate the prevailing dynamics revealed with empirical facts known about the system being studied. In the context of the electricity day-ahead market, we could contrast the time evolution of the extracted potential energy function with the particular characteristics of the underlying technology mix on the generation side. Conversely, we could link critical events, such as financial crises or changes in central banks' monetary policies, to the equilibrium points identified in the currency-exchange case study. Finally, we could refine our framework by considering other kernel types and bandwidth parameters for the KDE, a non-Markovian approach and/or by separating the empirical time series into stationary and nonstationary components. These extensions could enable us to identify key features such as kernel sensitivities

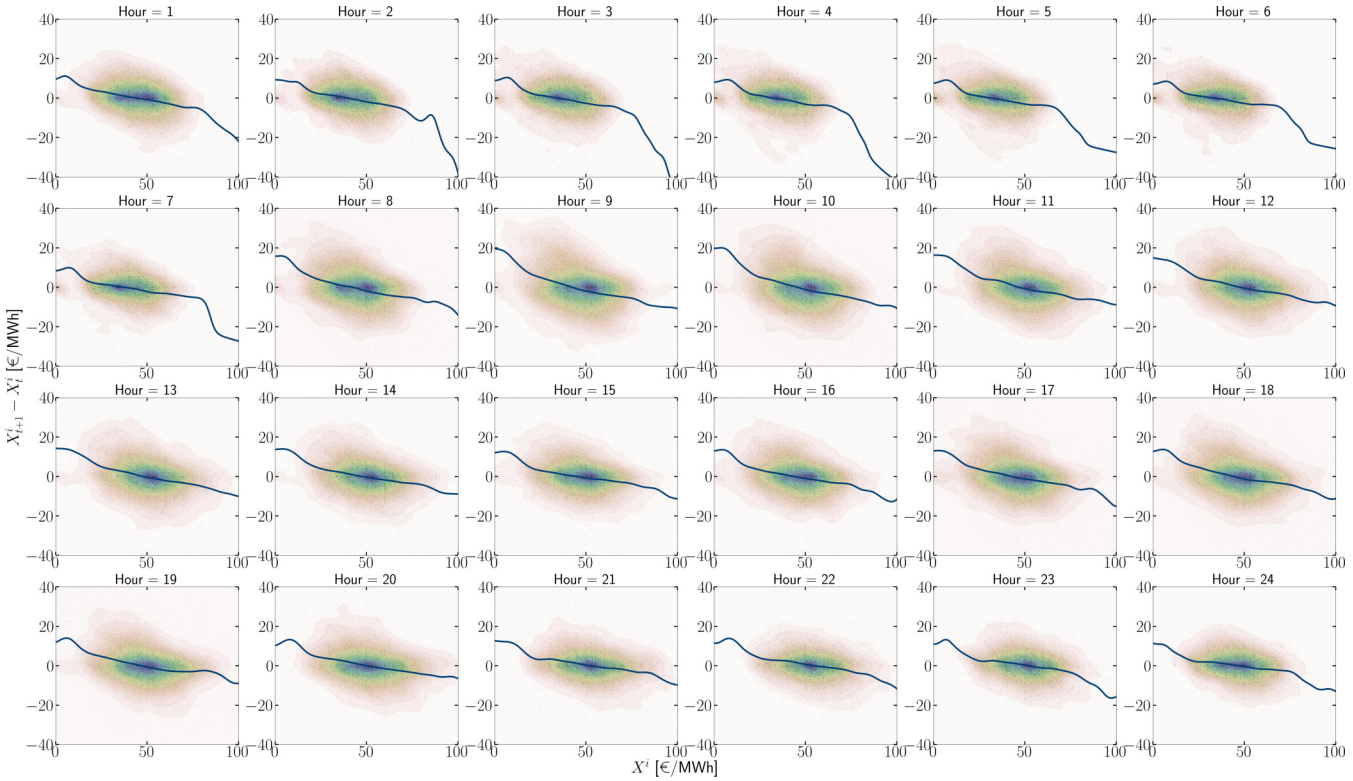

FIG. 8. Drift approximation of the multivariate LE that models the electricity day-ahead prices. Colored shadows represent the estimated PDF from KDE, ranging from red colors (low probability) to blue colors (higher probability). Blue lines for each hour represent the approximated drift as a function of the spot price, X^i.

and memory effects, and further facilitate understanding of the governing dynamics of complex systems under different dynamical regimes. We shall examine these and related questions in future studies.

ACKNOWLEDGMENTS

We are grateful to the anonymous referees for valuable comments and critical suggestions. We acknowledge financial support by the Imperial College London President's PhD Scholarship scheme, and by the ERC-EPSRC Frontier Research Guarantee through Grant No. EP/X038645, ERC through Advanced Grant No. 247031 and EPSRC through Grants No. EP/L025159 and No. EP/L020564.

APPENDIX A: RELATIONSHIP BETWEEN KM AND DRIFT-DIFFUSION COEFFICIENTS

Here we outline the derivation of the relationship between the KM coefficients and the drift and diffusion terms in Eq. (5). We first consider the Stieltjes integral over a time interval of length $\tau > 0$ with the initial condition $X^i_t = X^i$:

$$X^i_{t+\tau} - X^i = \int_t^{t+\tau} \mu^i(\mathbf{X}_{t'})dt' + \int_t^{t+\tau} \sigma_{ij}(\mathbf{X}_{t'})dW^j_{t'}. \quad \text{(A1)}$$

Consider next the Taylor expansions of $\mu^i(\mathbf{X}_{t'})$ and $\sigma_{ij}(\mathbf{X}_{t'})$:

$$\begin{aligned}\mu^i(\mathbf{X}_{t'}) &= \mu^i(\mathbf{X}) + \left(\frac{\partial}{\partial X^k}\mu^i(\mathbf{X})\right)\left(X^k_{t'} - X^k\right) + \dots \\ \sigma_{ij}(\mathbf{X}_{t'}) &= \sigma_{ij}(\mathbf{X}) + \left(\frac{\partial}{\partial X^k}\sigma_{ij}(\mathbf{X})\right)\left(X^k_{t'} - X^k\right) + \dots, \end{aligned} \quad \text{(A2)}$$

which are inserted into Eq. (A1) to obtain:

$$\begin{aligned} X^i_{t+\tau} - X^i = &\int_0^\tau \mu^i(\mathbf{X})d\tau' + \int_0^\tau \sigma_{ij}(\mathbf{X})dW^j_{\tau'} \\ &+ \int_0^\tau \left(\frac{\partial}{\partial X^k}\mu^i(\mathbf{X})\right)\left(X^k_{t+\tau'} - X^k\right)d\tau' \\ &+ \int_0^\tau \left(\frac{\partial}{\partial X^k}\sigma_{ij}(\mathbf{X})\right)\left(X^k_{t+\tau'} - X^k\right)dW^j_{\tau'} + \dots . \end{aligned} \quad \text{(A3)}$$

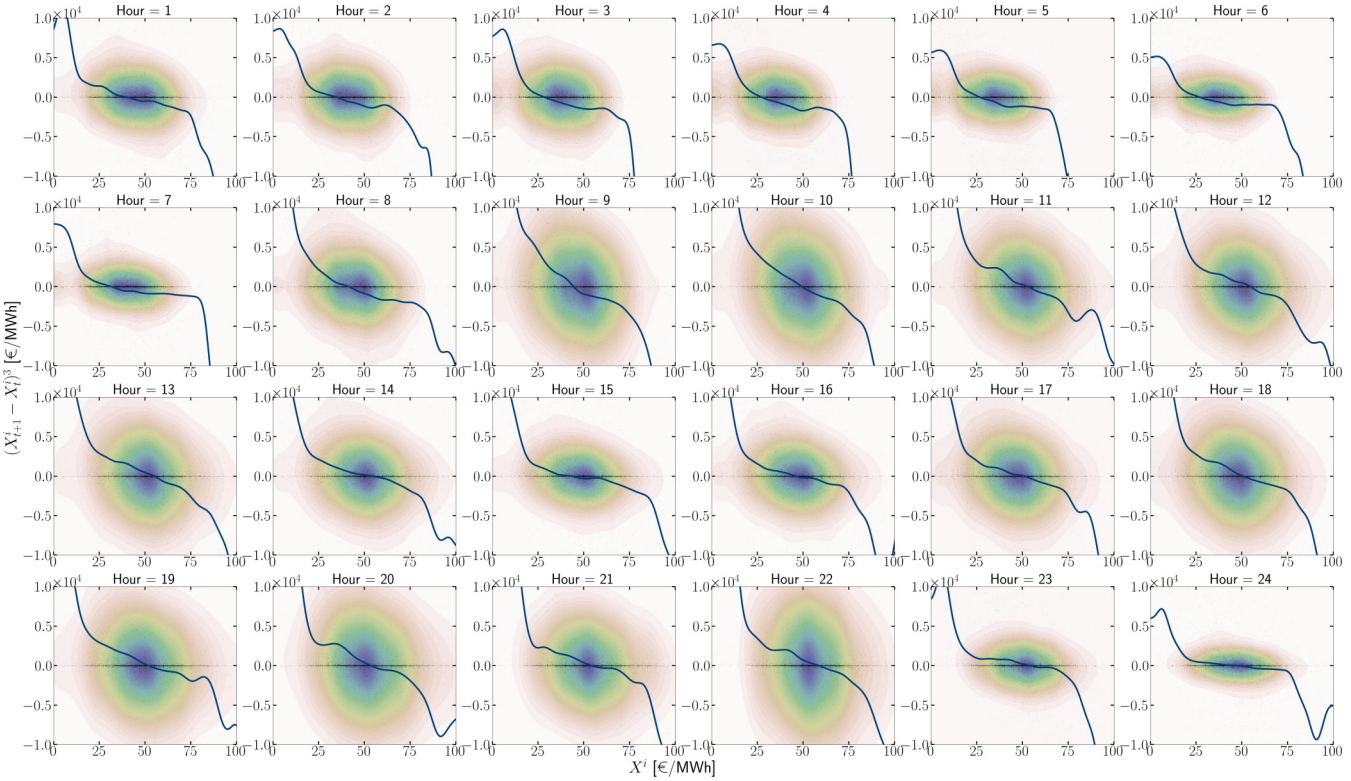

FIG. 9. Reconstruction of the elements in the main diagonal of the third KM coefficients tensor for the electricity day-ahead price time series. Colored shadows represent the estimated PDF from KDE, ranging from red colors (low probability) to blue colors (higher probability). Blue lines for each hour represent the mean value of the price differences to the power of three as a function of X^i. These are related to $D^{(3)}_{iii}$ by a factor of 1/3!.

We can rewrite the left-hand-side term of Eq. (A3) as:

$$X^i_{t+\tau} - X^i = \mu^i(\mathbf{X})\tau + \sigma_{ij}(\mathbf{X})W^j_\tau + \dots. \tag{A4}$$

Iterating Eq. (A3) by replacing $(X^k_{t+\tau'} - X^k)$ with Eq. (A4) yields:

$$\begin{aligned} X^i_{t+\tau} - X^i = &\int_0^\tau \mu^i(\mathbf{X})d\tau' + \int_0^\tau \sigma_{ij}(\mathbf{X})dW^{h_j}_{\tau'} \\ &+ \int_0^\tau \left(\frac{\partial}{\partial X^k}\mu^i(\mathbf{X})\right)\big(\mu^k(\mathbf{X})\tau' \\ &+ \sigma_{kj}(\mathbf{X})W^j_{\tau'} + \dots\big)d\tau' \\ &+ \int_0^\tau \left(\frac{\partial}{\partial X^k}\sigma_{ij}(\mathbf{X})\right)\big(\mu^k(\mathbf{X})\tau' \\ &+ \sigma_{kj}(\mathbf{X})W^j_{\tau'} + \dots\big)dW^j_{\tau'} + \dots. \end{aligned} \tag{A5}$$

Finally, applying $\langle\cdot\rangle$ on both sides of Eq. (A5), and keeping only terms that scale as τ:

$$\begin{aligned} \left\langle X^i_{t+\tau} - X^i\right\rangle = &\mu^i(\mathbf{X})\tau + \sigma_{ij}\left\langle W^j_\tau\right\rangle \\ &+ \left(\frac{\partial}{\partial X^k}\mu^i(\mathbf{X})\right)\sigma_{kj}(\mathbf{X})\left\langle\int_0^\tau W^j_{\tau'}d\tau'\right\rangle \\ &+ \left(\frac{\partial}{\partial X^k}\sigma_{ij}(\mathbf{X})\right)\sigma_{kj}(\mathbf{X})\left\langle\int_0^\tau W^j_{\tau'}dW^j_{\tau'}\right\rangle. \end{aligned} \tag{A6}$$

Recall that W^j_τ is a Wiener process, therefore:

$$\begin{aligned} \left\langle W^j_\tau\right\rangle &= 0 \\ \left\langle\int_0^\tau W^j_{\tau'}d\tau'\right\rangle &= 0 \\ \left\langle\int_0^\tau W^j_{\tau'}dW^j_{\tau'}\right\rangle &= 0. \end{aligned} \tag{A7}$$

Thus, $\langle X^i_{t+\tau} - X^i\rangle = \mu^i(\mathbf{X})\tau$, leading to the following relationship between the first KM coefficient and the drift:

$$D^{(1)}_i(\mathbf{X}) = \lim_{\tau\to 0}\frac{\left\langle X^i_{t+\tau} - X^i\right\rangle}{\tau} = \mu^i(\mathbf{X}). \tag{A8}$$

Following the same rationale for $\langle(X^i_{t+\tau} - X^i)(X^j_{t+\tau} - X^j)\rangle$, the second KM coefficient has the following relationship with the diffusion term:

$$\begin{aligned} D^{(2)}_{ij}(\mathbf{X}) &= \frac{1}{2}\lim_{\tau\to 0}\frac{\left\langle\left[X^i_{t+\tau} - X^i\right]\left[X^j_{t+\tau} - X^j\right]\right\rangle}{\tau} \\ &= \frac{1}{2}\sigma_{ik}(\mathbf{X})\sigma_{jk}(\mathbf{X}). \end{aligned} \tag{A9}$$

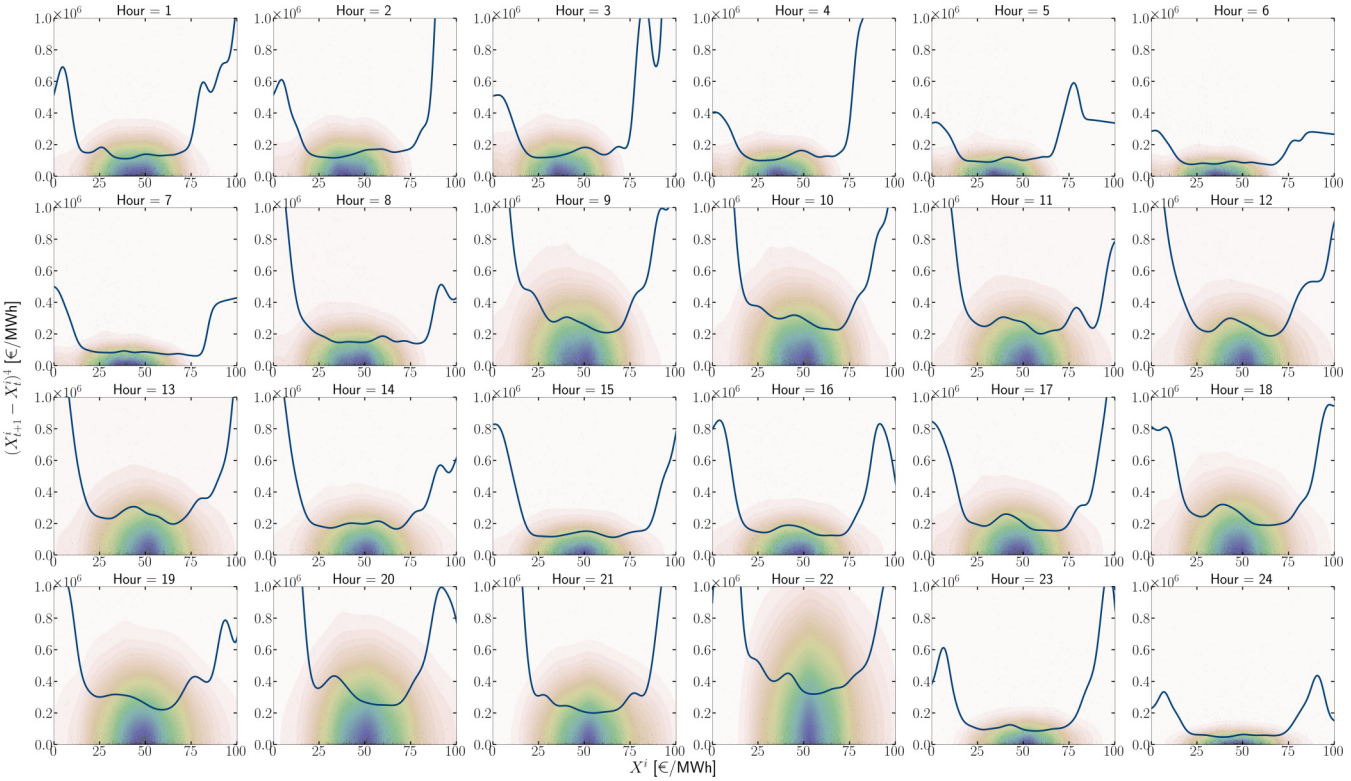

FIG. 10. Reconstruction of the elements in the main diagonal of the fourth KM coefficients tensor for the electricity day-ahead price time series. Colored shadows represent the estimated PDF from KDE, ranging from red colors (low probability) to blue colors (higher probability). Blue lines for each hour represent the mean value of the price differences to the power of four as a function of X^i. These are related to $D^{(4)}_{iiii}$ by a factor of 1/4!.

Through this derivation, we have assumed stationary conditions, i.e., $\partial\rho(\mathbf{X}, t)/\partial t = 0$, thus, removing the temporal dependence, t, on the KM coefficients.

APPENDIX B: FIRST KM COEFFICIENT OF THE ELECTRICITY DAY-AHEAD MARKET

Figure 8 illustrates the results of the reconstruction of the first KM coefficient for each hourly price of the electricity day-ahead market. As commented in Sec. III B, each hour has only one equilibrium point which coincides with the minimum of the parabola represented in Fig. 4. Also, different nonlinear regions are observed for each hour around $X^i < 20$ and $X^i > 75$ € /MWh.

APPENDIX C: HIGHER-ORDER KM COEFFICIENTS

We analyze here the higher-order KM coefficients for the two real-world applications under consideration. Specifically, we focus on $D^{(3)}$ and $D^{(4)}$ to validate the Gaussian assumption and its implications.

1. Electricity day-ahead market

Given that the LE for the electricity day-ahead prices has 24 components, the third-order KM coefficient tensor $D^{(3)}_{ijp}$ consists of 13 824 elements, while the fourth-order tensor $D^{(4)}_{ijpq}$ contains 331 776 elements, with $i, j, p, q = \{1, \ldots, 24\}$. This large number of elements raises computational constraints, making it impractical to explore the full behavior of higher-order KM coefficients in the particular application. To address this, we restrict our analysis to the elements along the main diagonal, $i = j = p = q$, for both coefficients. These will be sufficient to discuss the implications of the Gaussian assumption. We apply the same KDE technique described in Sec. III B to approximate $D^{(3)}_{iii}$, and $D^{(4)}_{iiii}$, using a $(10^3, 10^3)$ grid for the mesh resolution.

Figures 9 and 10 represent the reconstruction of the third and fourth KM coefficient, respectively, for each hour of the electricity day-ahead market. The expected values of $D^{(3)}_{iii}$ and $D^{(4)}_{iiii}$ are summarized in Table III. From these results, it is evident that the diagonal elements of the $D^{(3)}$ and $D^{(4)}$ tensors are sufficient to demonstrate that the electricity day-ahead market PDF does not truly follow a Gaussian distribution.

TABLE III. Selected higher-order KM coefficients for the electricity day-ahead prices time series. The large values for $D^{(4)}$ arise from the average of values raised to the fourth power and the presence of price spikes.

Coefficient	Expected value
$D^{(3)}_{1\,1\,1}$	–23.33
$D^{(3)}_{2\,2\,2}$	–58.02
$D^{(3)}_{3\,3\,3}$	–87.57
$D^{(3)}_{4\,4\,4}$	–85.59
$D^{(3)}_{5\,5\,5}$	–75.30
$D^{(3)}_{6\,6\,6}$	–63.60
$D^{(3)}_{7\,7\,7}$	–51.07
$D^{(3)}_{8\,8\,8}$	–37.20
$D^{(3)}_{9\,9\,9}$	0.55
$D^{(3)}_{10\,10\,10}$	32.84
$D^{(3)}_{11\,11\,11}$	41.57
$D^{(3)}_{12\,12\,12}$	47.85
$D^{(3)}_{13\,13\,13}$	51.51
$D^{(3)}_{14\,14\,14}$	27.72
$D^{(3)}_{15\,15\,15}$	–14.52
$D^{(3)}_{16\,16\,16}$	12.98
$D^{(3)}_{17\,17\,17}$	37.45
$D^{(3)}_{18\,18\,18}$	67.87
$D^{(3)}_{19\,19\,19}$	64.11
$D^{(3)}_{20\,20\,20}$	67.20
$D^{(3)}_{21\,21\,21}$	41.77
$D^{(3)}_{22\,22\,22}$	60.61
$D^{(3)}_{23\,23\,23}$	29.86
$D^{(3)}_{24\,24\,24}$	–2.73
$D^{(4)}_{1\,1\,1\,1}$	5470.45
$D^{(4)}_{2\,2\,2\,2}$	5785.67
$D^{(4)}_{3\,3\,3\,3}$	5903.55
$D^{(4)}_{4\,4\,4\,4}$	4938.28
$D^{(4)}_{5\,5\,5\,5}$	4311.98
$D^{(4)}_{6\,6\,6\,6}$	3475.38
$D^{(4)}_{7\,7\,7\,7}$	3553.98
$D^{(4)}_{8\,8\,8\,8}$	6713.12
$D^{(4)}_{9\,9\,9\,9}$	11363.75
$D^{(4)}_{10\,10\,10\,10}$	12038.22
$D^{(4)}_{11\,11\,11\,11}$	10622.97
$D^{(4)}_{12\,12\,12\,12}$	10115.90
$D^{(4)}_{13\,13\,13\,13}$	10595.58
$D^{(4)}_{14\,14\,14\,14}$	8204.78
$D^{(4)}_{15\,15\,15\,15}$	5546.84
$D^{(4)}_{16\,16\,16\,16}$	6526.70
$D^{(4)}_{17\,17\,17\,17}$	8297.28
$D^{(4)}_{18\,18\,18\,18}$	10265.40
$D^{(4)}_{19\,19\,19\,19}$	11687.38
$D^{(4)}_{20\,20\,20\,20}$	12595.26
$D^{(4)}_{21\,21\,21\,21}$	9634.31
$D^{(4)}_{22\,22\,22\,22}$	15668.43
$D^{(4)}_{23\,23\,23\,23}$	4187.51
$D^{(4)}_{24\,24\,24\,24}$	2384.74

TABLE IV. Higher-orders KM coefficients for the currency-exchange time series.

Order	Coefficient(s)	Expected Value
3	$D^{(3)}_{111}$	5.69×10^{-8}
	$D^{(3)}_{222}$	-5.61×10^{-8}
	$D^{(3)}_{112}$, $D^{(3)}_{121}$, $D^{(3)}_{211}$	2.17×10^{-11}
	$D^{(3)}_{221}$, $D^{(3)}_{212}$, $D^{(3)}_{122}$	7.09×10^{-10}
4	$D^{(4)}_{1111}$	5.90×10^{-8}
	$D^{(4)}_{2222}$	1.70×10^{-8}
	$D^{(4)}_{1112}, D^{(4)}_{1121}, D^{(4)}_{1211}, D^{(4)}_{2111}$	3.92×10^{-11}
	$D^{(4)}_{2221}, D^{(4)}_{2212}, D^{(4)}_{2122}, D^{(4)}_{1222}$	5.85×10^{-11}
	$D^{(4)}_{1122}, D^{(4)}_{1221}, D^{(4)}_{2211}, D^{(4)}_{2121}, D^{(4)}_{2112}, D^{(4)}_{1212}$	5.19×10^{-11}

However, as we discussed in Sec. II A, the Gaussian approximation remains the best proxy for capturing the prevailing dynamics of the system while avoiding potential misrepresentations, like the negative values in the PDF because of an arbitrary finite-order truncation of the infinite KM series in Eq. (3).

Regarding the mean values in Figs. 9 and 10, the significant deviation from 0 and the irregular shape observed arise from long-tailed conditional distributions, caused by the presence of infrequent but large price jumps across the whole X^i domain. These jumps are particularly concentrated at the extremes of the price range ($X^i < 25$ and $X^i > 75$). We acknowledge the limited capability of our model in capturing this particular feature, as the Gaussian assumption is not capable of accounting for abrupt changes. More suitable alternatives include ad hoc model parameters or frameworks like Poisson processes [44], beta distributions [81], or regime-switching frameworks [45,46]. However, introducing such alternatives in our framework would compromise its domain-agnostic nature and adaptability.

2. Currency exchange market

We replicate the KDE technique mentioned in Sec. III C to reconstruct $D^{(3)}$ and $D^{(4)}$. For the mesh resolution, we apply a $(10^3, 10^3)$ data points grid for $D^{(3)}_{ijp}$ and $D^{(4)}_{ijpq}$, when considering the elements along the main diagonal, $i = j = p = q$. Conversely, for off-diagonal elements, the grid contains $(5 \times 10^2, 5 \times 10^2, 5 \times 10^2)$ data points. To ensure clarity, Fig. 11 illustrates the reconstruction of the elements in the main diagonal of the third and fourth KM coefficients. Additionally, Table IV summarizes the expected value of all third and fourth KM coefficients. Overall, both $D^{(3)}$ and $D^{(4)}$ are several orders of magnitude smaller than $D^{(1)}$ and $D^{(2)}$. Therefore, we can assume these coefficients to be negligible, demonstrating that the currency-exchange market for both EURUSD and GBPUSD can be effectively represented with our framework which relies on the Gaussian assumption.

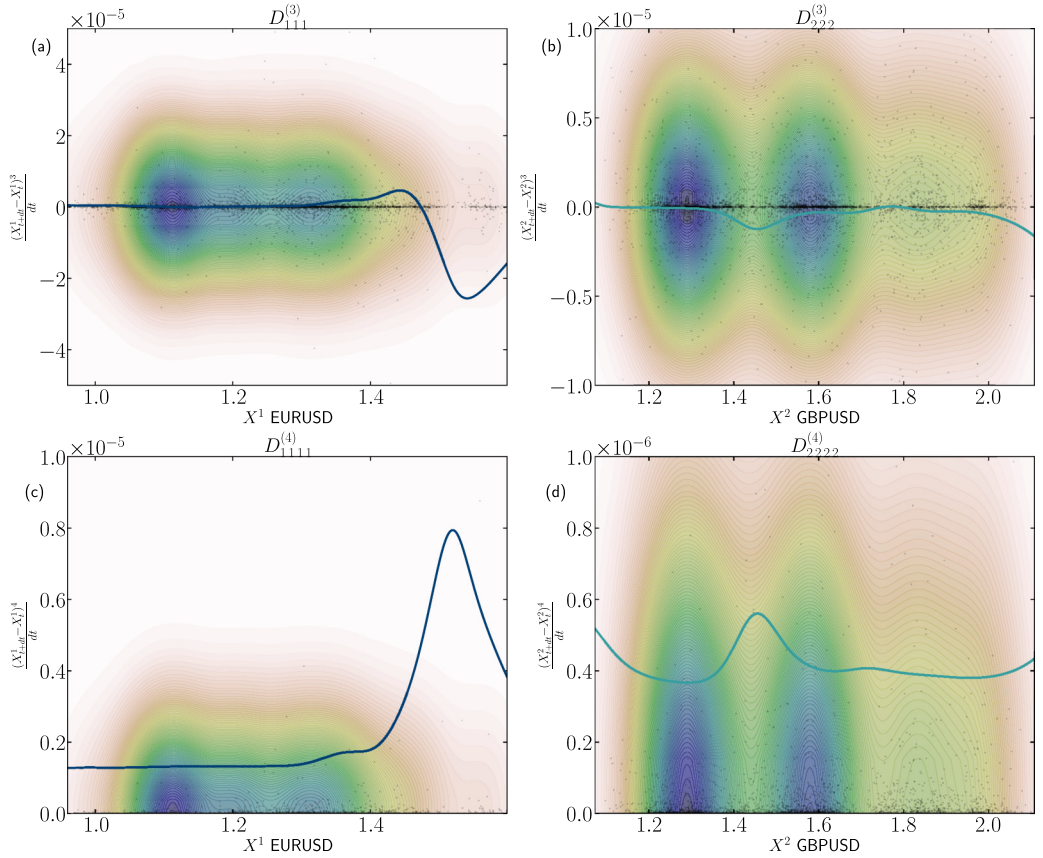

FIG. 11. Reconstruction of the elements in the main diagonal of the third and fourth KM coefficients tensor for the currency-exchange time series. Colored shadows represent the estimated PDF from KDE, ranging from red colors (low probability) to blue colors (higher probability). The blue lines in (a) and (c) indicate the mean value of the daily rate differences to the power of three and four for the EURUSD, X^1. The green lines in (b) and (d) represent the same values but for GBPUSD, X^2. These are related to $D^{(3)}_{iii}$ and $D^{(4)}_{iiii}$ by a factor of $1/n!$, $n = \{3, 4\}$, as stated in Eq. (3).

[1] S. Kalliadasis, S. Krumscheid, and G. A. Pavliotis, A new framework for extracting coarse-grained models from time series with multiscale structure, J. Comput. Phys. **296**, 314 (2015).

[2] H. Mori, Transport, collective motion, and Brownian motion, Prog. Theor. Phys. **33**, 423 (1965).

[3] R. Zwanzig, Nonlinear generalized Langevin equations, J. Stat. Phys. **9**, 215 (1973).

[4] A. Chorin and P. Stinis, Problem reduction, renormalization, and memory, Commun. Appl. Math. Comput. Sci. **1**, 1 (2006).

[5] A. Russo, M. A. Durán-Olivencia, I. G. Kevrekidis, and S. Kalliadasis, Machine learning memory kernels as closure for non-Markovian stochastic processes, IEEE Trans. Neural Netw. Learn. Syst. **35**, 6531 (2024).

[6] H. Grabert, *Projection Operator Techniques in Nonequilibrium Statistical Mechanics*, 1st ed. (Springer, Berlin, 1982).

[7] R. Friedrich, S. Siegert, J. Peinke, S. Lück, M. Siefert, M. Lindemann, J. Raethjen, G. Deuschl, and G. Pfister, Extracting model equations from experimental data, Phys. Lett. A **271**, 217 (2000).

[8] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar, Approaching complexity by stochastic methods: From biological systems to turbulence, Phys. Rep. **506**, 87 (2011).

[9] M. Anvari, M. R. R. Tabar, J. Peinke, and K. Lehnertz, Disentangling the stochastic behavior of complex time series, Sci. Rep. **6**, 35435 (2016).

[10] P. Xie, R. Car, and W. E, Ab initio generalized Langevin equation, Proc. Nat. Acad. Sci. USA **121**, e2308668121 (2024).

[11] M. R. R. Tabar, *Analysis and Data-Based Reconstruction of Complex Nonlinear Dynamical Systems*, 1st ed. (Springer, Cham, 2019).

[12] S. Siegert, R. Friedrich, and J. Peinke, Analysis of data sets of stochastic systems, Phys. Lett. A **243**, 275 (1998).

[13] A. M. van Mourik, A. Daffertshofer, and P. J. Beek, Estimating Kramers–Moyal coefficients in short and non-stationary data sets, Phys. Lett. A **351**, 13 (2006).

[14] D. Kleinhans and R. Friedrich, Maximum likelihood estimation of drift and diffusion functions, Phys. Lett. A **368**, 194 (2007).

[15] S. J. Lade, Finite sampling interval effects in Kramers–Moyal analysis, Phys. Lett. A **373**, 3705 (2009).

[16] D. Lamouroux and K. Lehnertz, Kernel-based regression of drift and diffusion coefficients of stochastic processes, Phys. Lett. A **373**, 3507 (2009).

[17] C. Honisch and R. Friedrich, Estimation of Kramers-Moyal coefficients at low sampling rates, Phys. Rev. E **83**, 066701 (2011).

[18] F. Nikakhtar, L. Parkavousi, M. Sahimi, M. R. R. Tabar, U. Feudel, and K. Lehnertz, Data-driven reconstruction of stochastic dynamical equations based on statistical moments, New J. Phys. **25**, 083025 (2023).

[19] M. R. Tabar, F. Nikakhtar, L. Parkavousi, A. Akhshi, U. Feudel, and K. Lehnertz, Revealing higher-order interactions in high-dimensional complex systems: A data-driven approach, Phys. Rev. X **14**, 011050 (2024).

[20] E. Anahua, S. Barth, and J. Peinke, Markovian power curves for wind turbines, Wind Energy **11**, 219 (2008).

[21] J. Gottschall and J. Peinke, How to improve the estimation of power curves for wind turbines, Environ. Res. Lett. **3**, 015005 (2008).

[22] R. Friedrich, J. Peinke, and C. Renner, How to quantify deterministic and random influences on the statistics of the foreign exchange market, Phys. Rev. Lett. **84**, 5224 (2000).

[23] M. Petelczyc, J. J. Żebrowski, and R. Baranowski, Kramers-Moyal coefficients in the analysis and modeling of heart rate variability, Phys. Rev. E **80**, 031127 (2009).

[24] C. Wiedemann, H. M. Bette, M. Wächter, J. Freund, T. Guhr, and J. Peinke, Extension of the Langevin power curve analysis by separation per operational state, arXiv:2305.15512.

[25] M. Verleysen and D. François, The curse of dimensionality in data mining and time series prediction, in *Computational Intelligence and Bioinspired Systems*, edited by J. Cabestany, A. Prieto, and F. Sandoval (Springer, Berlin, 2005), pp. 758–770.

[26] B. D. Goddard, A. Nold, N. Savva, G. A. Pavliotis, and S. Kalliadasis, General dynamical density functional theory for classical fluids, Phys. Rev. Lett. **109**, 120603 (2012).

[27] B. D. Goddard, A. Nold, N. Savva, P. Yatsyshin, and S. Kalliadasis, Unification of dynamic density functional theory for colloidal fluids to include inertia and hydrodynamic interactions: derivation and numerical experiments, J. Phys.: Condens. Matter **25**, 035101 (2013).

[28] M. A. Durán-Olivencia, P. Yatsyshin, B. D. Goddard, and S. Kalliadasis, General framework for fluctuating dynamic density functional theory, New J. Phys. **19**, 123022 (2017).

[29] A. Provenzale, Climate as a complex dynamical system, in *Mathematical Models and Methods for Planet Earth* (Springer International, Cham, 2014), pp. 135–142.

[30] M. Schmuck, M. Pradas, S. Kalliadasis, and G. A. Pavliotis, New stochastic mode reduction strategy for dissipative systems, Phys. Rev. Lett. **110**, 244101 (2013).

[31] P. Alaton, B. Djehiche, and D. Stillberger, On modelling and pricing weather derivatives, Appl. Math. Finance **9**, 1 (2002).

[32] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications*, 2nd ed. (Springer, Berlin, 1996).

[33] R. F. Pawula, Approximation of the linear Boltzmann equation by the Fokker-Planck equation, Phys. Rev. **162**, 186 (1967).

[34] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York, 2009).

[35] B. Silverman, *Density Estimation for Statistics and Data Analysis*, 1st ed. (Chapman & Hall, London, 1986).

[36] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, 1st ed. (John Wiley & Sons, New York, 1992).

[37] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, 1st ed. (Springer, Berlin, 1992).

[38] C. Harris, *Electricity Markets: Pricing, Structures and Economics*, 1st ed. (John Wiley & Sons, New York, 2006).

[39] S. Borenstein, The trouble with electricity markets: Understanding California's restructuring disaster, J. Econ. Perspect. **16**, 191 (2002).

[40] C. R. Knittel and M. R. Roberts, An empirical examination of restructured electricity prices, Energy Econ. **27**, 791 (2005).

[41] L. Tschora, E. Pierre, M. Plantevit, and C. Robardet, Electricity price forecasting on the day-ahead market using machine learning, Appl. Energy **313**, 118752 (2022).

[42] J. Lago, G. Marcjasz, B. De Schutter, and R. Weron, Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark, Appl. Energy **293**, 116983 (2021).

[43] J. J. Lucia and E. S. Schwartz, Electricity prices and power derivatives: Evidence from the Nordic power exchange, Rev. Deriv. Res. **5**, 5 (2002).

[44] B. Hambly, S. Howison, and T. Kluge, Modelling spikes and pricing swing options in electricity markets, Quant. Finance **9**, 937 (2009).

[45] R. Huisman and R. Mahieu, Regime jumps in electricity prices, Energy Econ. **25**, 425 (2003).

[46] R. Weron, M. Bierbrauer, and S. Trück, Modeling electricity prices: Jump diffusion and regime switching, Physica A **336**, 39 (2004).

[47] J. Collet, V. Duwig, and N. Oudjane, Some non-Gaussian models for electricity spot prices, in *Proceedings of the 9th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS'06)* (Stockholm, Sweden, 2006), pp. 1–7.

[48] F. E. Benth, The stochastic volatility model of Barndorff-Nielsen and Shephard in commodity markets, Math. Finance **21**, 595 (2011).

[49] S. Borovkova and M. D. Schmeck, Electricity price modeling with stochastic time change, Energy Econ. **63**, 51 (2017).

[50] R. Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future, Int. J. Forecast. **30**, 1030 (2014).

[51] T. Deschatre, O. Féron, and P. Gruet, A survey of electricity spot and futures price models for risk management applications, Energy Econ. **102**, 105504 (2021).

[52] B. Uniejewski, J. Nowotarski, and R. Weron, Automated variable selection and shrinkage for day-ahead electricity price forecasting, Energies **9**, 621 (2016).

[53] F. Ziel and R. Weron, Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs multivariate modeling frameworks, Energy Econ. **70**, 396 (2018).

[54] J. Portela González, A. Muñoz San Roque, and E. Alonso Pérez, Forecasting functional time series with a new Hilbertian ARMAX model: Application to electricity price forecasting, IEEE Trans. Power Syst. **33**, 545 (2018).

[55] S. Krumscheid, M. Pradas, G. A. Pavliotis, and S. Kalliadasis, Data-driven coarse graining in action: Modeling and prediction of complex systems, Phys. Rev. E **92**, 042139 (2015).

[56] Y. Stepanov, P. Rinn, T. Guhr, J. Peinke, and R. Schäfer, Stability and hierarchy of quasi-stationary states: Financial markets as an example, J. Stat. Mech. (2015) P08011.

[57] J. Chen, *Essentials of Foreign Exchange Trading*, 1st ed. (John Wiley & Sons, New York, 2009).
[58] R. N. Mantegna and H. E. Stanley, *Introduction to Econophysics: Correlations and Complexity in Finance*, 1st ed. (Cambridge University Press, Cambridge, UK, 1999).
[59] M. O'Hara, *Market Microstructure Theory*, 1st ed. (John Wiley & Sons, New York, 1998).
[60] R. K. Lyons, *The Microstructure Approach to Exchange Rates*, 1st ed. (The MIT Press, Cambridge, MA, 2001).
[61] S. M. D. Queirós, Trading volume in financial markets: An introductory review, Chaos Solit. Fractals **88**, 24 (2016).
[62] L. Bachelier, Théorie de la spéculation, Ann. Sci. École Norm. Sup. **17**, 21 (1900).
[63] R. C. Merton, Theory of rational option pricing, Bell J. Econ. Manage. Sci. **4**, 141 (1973).
[64] V. Stojkoski, T. Sandev, L. Basnarkov, L. Kocarev, and R. Metzler, Generalised geometric Brownian motion: Theory and applications to option pricing, Entropy **22**, 1432 (2020).
[65] I. Halperin and M. Dixon, "Quantum Equilibrium-Disequilibrium": Asset price dynamics, symmetry breaking, and defaults as dissipative instantons, Physica A **537**, 122187 (2020).
[66] S. L. Heston, A closed-form solution for options with stochastic volatility with applications to bond and currency options, Rev. Financ. Stud. **6**, 327 (1993).
[67] T. Bollerslev, R. Y. Chou, and K. F. Kroner, ARCH modeling in finance: A review of the theory and empirical evidence, J. Econom. **52**, 5 (1992).
[68] S. Degiannakis and E. Xekalaki, Autoregressive conditional heteroscedasticity (ARCH) models: A review, Qual. Technol. Quant. Manage. **1**, 271 (2004).
[69] M. B. Garman and S. W. Kohlhagen, Foreign currency option values, J. Int. Money Finance **2**, 231 (1983).
[70] J. O. Grabbe, The pricing of call and put options on foreign exchange, J. Int. Money Finance **2**, 239 (1983).
[71] P. Doust, The stochastic intrinsic currency volatility model: A consistent framework for multiple FX rates and their volatilities, Appl. Math. Finance **19**, 381 (2012).
[72] M. Escobar and C. Gschnaidtner, A multivariate stochastic volatility model with applications in the foreign exchange market, Rev. Deriv. Res. **21**, 1 (2018).
[73] R. T. Baillie and T. Bollerslev, Common stochastic trends in a system of exchange rates, J. Finance **44**, 167 (1989).
[74] W. J. Crowder, Foreign exchange market efficiency and common stochastic trends, J. Int. Money Finance **13**, 551 (1994).
[75] J.-P. Bouchaud and R. Cont, A Langevin approach to stock market fluctuations and crashes, Eur. Phys. J. B **6**, 543 (1998).
[76] T. Wand, T. Wiedemann, J. Harren, and O. Kamps, Estimating stable fixed points and Langevin potentials for financial dynamics, Phys. Rev. E **109**, 024226 (2024).
[77] C. Renner, J. Peinke, and R. Friedrich, Evidence of Markov properties of high frequency exchange rate data, Physica A **298**, 499 (2001).
[78] F. Farahpour, Z. Eskandari, A. Bahraminasab, G. R. Jafari, F. Ghasemi, M. Sahimi, and M. R. R. Tabar, A Langevin equation for the rates of currency exchange based on the Markov analysis, Physica A **385**, 601 (2007).
[79] P. Rinn, Y. Stepanov, J. Peinke, T. Guhr, and R. Schäfer, Dynamics of quasi-stationary systems: Finance as an example, Europhys. Lett. **110**, 68003 (2015).
[80] B. Derrida, Non-equilibrium steady states: Fluctuations and large deviations of the density and of the current, J. Stat. Mech. (2007) P07023.
[81] R. Becker, S. Hurn, and V. Pavlov, Modelling spikes in electricity prices, Econ. Rec. **83**, 371 (2007).