Data-driven coarse graining in action: Modeling and prediction of complex systems

Journal Item

Data-driven coarse graining in action: Modeling and prediction of complex systems

S. Krumscheid,[1,2,*] M. Pradas,[1,†] G.A. Pavliotis,[2] and S. Kalliadasis[1]

[1] *Department of Chemical Engineering, Imperial College London, London SW7 2AZ, United Kingdom*
[2] *Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom*

(Dated: October 18, 2015)

In many physical, technological, social, and economic applications, one is commonly faced with the task of estimating statistical properties, such as mean first passage times of a temporal continuous process, from empirical data (experimental observations). Typically, however, an accurate and reliable estimation of such properties directly from the data alone is not possible as the time series is often too short, or the particular phenomenon of interest is only rarely observed. We propose here a theoretical-computational framework which provides us with a systematic and rational estimation of statistical quantities of a given temporal process, such as waiting times between subsequent bursts of activity in intermittent signals. Our framework is illustrated with applications from real-world data sets, ranging from marine biology to paleoclimatic data.

PACS numbers: 02.50.-r, 02.50.Tt, 05.40.Fb, 05.45.Tp

I. INTRODUCTION

Over the last few years, there has been an increasing demand for capturing generic statistical properties of complex systems based on available data only. Such systems are often strongly influenced by random fluctuations which play a crucial role in the various intriguing phenomena emerging in temporal observations [1, 2]. Understanding the underlying complex processes of such phenomena is a common task in many disciplines, but often it is not possible to estimate statistical properties directly from empirical data alone because e.g. the phenomenon of interest occurs rarely. On the other hand, often also a purely reductionist/bottom-up approach is either impossible or results in computationally prohibitive mathematical models.

An alternative approach is to identify a reduced (coarse grained) model from the experimental data which retains the fundamental aspects of the original system. This is in fact at the core of data-driven coarse graining but despite its fundamental significance, to date there does not exist a systematic framework for this. Relying exclusively on the observations and treating the corresponding reduced model as a "black box" (that is, in technical terms using nonparametric estimators [3], see also [4] for a review of such techniques) is, however, not reasonable since such an approach introduces errors in regions where only few observations exist [5], e.g. rare phenomena, thus corrupting a model-based analysis. An accurate and more general procedure is to follow a semiparametric approach where we postulate a model, i.e. we introduce a parametric ansatz (in a "grey-box" modeling approach) which is consistent with the essential characteristics of the experimental data, such as for example dynamic state transitions.

In this study we outline a unified generic theoretical-computational framework for data-driven modeling based on the above semiparametric approach with the ultimate aim of analyzing complex phenomena arising in a wide spectrum of different systems. To exemplify the methodology we use two representative examples of current interest, namely experimental observations of the foraging behavior of marine predators [6], and the temperature record during the last glacial period [7].

The manuscript is organized as follows. Section II presents the data-driven modeling framework which is then applied to a test case in Sec. III. Section IV offers results of the analysis of two real data sets, corresponding to movement pattern of marine predators and climate transitions during the last glacial period. We conclude in Sec. V.

II. GENERIC DATA-DRIVEN MODELING

A schematic representation of our methodology is shown in Fig. 1 and consists of two main steps. The first one is a model selection (postulate - assess/validate) procedure, which allows to select a simple coarse grained model from experimental observations. This model is then assessed and validated and eventually used in a second step to predict different quantities of interest.

We are interested in systems where the underlying noisy process is continuous in time and thus consider the following prototypical Itô stochastic differential equation (SDE):

$$dX(t) = f(X(t);\theta)\,dt + g(X(t);\theta)\,dW(t)\,, \qquad (1)$$

$X(t) \in \mathbb{R}^d$ for $t \geq 0$, where f and g are the drift and diffusion coefficients, respectively, with the latter controlling the influence of the stochastic driving through a Wiener process, $W(t)$.

* Current address: Mathematics Institute of Computational Science and Engineering, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

† Current address: Department of Mathematics and Statistics, The Open University, Milton Keynes MK7 6AA, United Kingdom

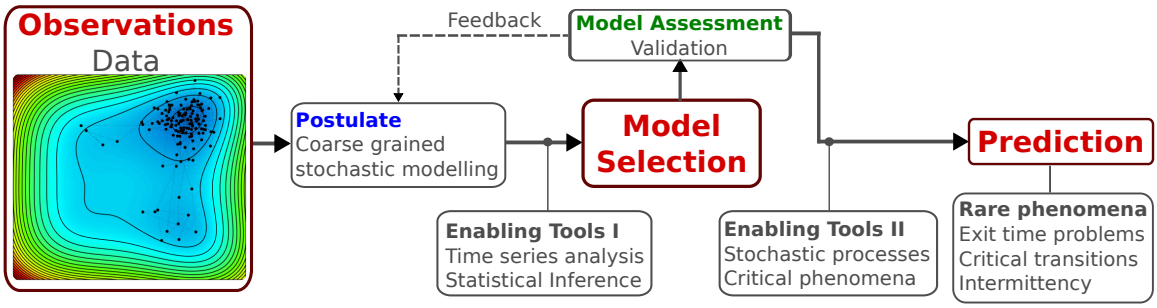

FIG. 1: (Color online) Flow chart of the data-driven modeling framework: Given observations (data) we postulate a coarse-grained stochastic parametric model which is fitted (via statistical inference and time series analysis tools, which we refer to as Enabling Tools I) to the data and refined via a model selection process (see main text). In particular, via an assessment/validation and fine-tuning procedure we determine the structure of the model and the minimum number of parameters needed. Once the model has been validated, it is used to predict underlying statistical properties by using critical phenomena and stochastic processes tools (Enabling Tools II). The far-left figure is a numerical example of Brownian motion in a two-dimensional potential.

We postulate first several model candidates, i.e. the functions f and g in (1), based on two criteria: (i) they must support features which are observed in the empirical data (e.g. state transitions) and (ii) they have to reproduce functional features observed in a preliminary nonparametric analysis in regions where most discrete-time observations of (1) are located. We note that the second criterion is primarily used as a data-driven modeling guidance. We reiterate, that a fully nonparametric modeling approach is typically not feasible in practice, as we will also illustrate in Sec. III. Due to the combination of phenomena-driven and data-driven aspects we can instead use a semiparametric approach here. That is, we consider expansions (e.g. Taylor or Fourier) of both drift and diffusion coefficients that support aspects observed in the time series. Different models can then be constructed by varying the number of unknown parameters in the these expansions. The postulated models are then compared and refined within the framework's model assessment/postulation feedback loop by combining statistical model selection criteria with further data-driven considerations (e.g. intermittency or shape of the probability density function).

A. Parametric Inference for SDEs and Model Selection

Given a model candidate, we proceed then to estimate the parameter vector $\theta \in \Theta \subset \mathbb{R}^m$ using a maximum likelihood framework due to its favorable theoretical properties (see e.g. [8, 9]). Specifically, let $\mathcal{X}_n$ be the sample with true parameter θ^*, that is $\mathcal{X}_n := (X_i)_{0\le i\le n}$ at sampling times $0 = t_0 < t_1 < \cdots < t_n = \mathsf{T}$ with $X_i \equiv X(t_i)$. The maximum likelihood estimator for θ^* based on the observations $\mathcal{X}_n$ is defined as the maximizer of the likelihood function over Θ

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} L_n(\theta;\mathcal{X}_n)\ , \tag{2}$$

where $L_n(\theta;\mathcal{X}_n)$ denotes the likelihood function:

$$L_n(\theta;\mathcal{X}_n) = \prod_{i=0}^{n-1} p_\theta(t_{i+1}-t_i, X_{i+1}|X_i)p_\theta(X_0), \tag{3}$$

with $p_\theta(x)$ being the probability density function (PDF) of the initial condition and $p_\theta(\Delta t, x|y)$ denotes the conditional probability density function. The conditional probability density function is usually not known in closed-form and we approximate it by adopting the closed-form expansion due to Aït-Sahalia [10]. The main idea is to transform the problem into one with transition densities that can be approximated accurately by means of an expansion in terms of Hermite polynomials. Truncating this expansion and inverting the transformation, an approximation of $p_\theta(\cdot,\cdot|\cdot)$ can be obtained in closed form. The coefficients determining this expansion depend on the considered functional form of both drift and diffusion coefficient in (1) and can become rather involved. Using a careful combination of symbolic and numerical computations, it is possible nonetheless to evaluate these coefficients accurately and efficiently.

It is worth emphasizing that while the MLE approach works well for data sets (time series) with a single characteristic time scale, it becomes asymptotically biased when applied to data coming from multiscale stochastic systems. For such systems, statistical inference methodologies that take into account the multiscale nature of the data set have to be used [11, 12]. We emphasize though, that the general framework as illustrated in Fig. 1 remains unaltered even in that case.

Once we have obtained the parameter vector and the corresponding likelihood function for several model candidates, we proceed to select a few of them (typically

two) by making use of two model selection techniques, namely the sample size corrected Akaike Information Criterion (AICc) and the Bayesian Information Criterion (BIC), both of which provide measures of the relative quality of the SDE parametrization (1) based on the given set of data; see, e.g. [13]. These two techniques rely on the maximized likelihood function of the considered model and the available observations $\mathcal{X}_n$, that is they depend on $L_n\big(\hat{\theta}_n;\mathcal{X}_n\big)$, where $\hat{\theta}_n$ denotes the estimated m-dimensional parameter vector defining the SDE model (1). In particular, the finite sample size corrected AICc is given by

$$\mathrm{AICc} = 2m(n+1)/(n-m) - 2\ln\big(L_n(\hat{\theta}_n;\mathcal{X}_n)\big)\ ,$$

and the BIC is defined as

$$\mathrm{BIC} = m\ln(n+1) - 2\ln\big(L_n(\hat{\theta}_n;\mathcal{X}_n)\big)\ .$$

We note that both techniques are designed to penalize over-fitted models, that is a parametrization with many parameters is not as valuable as a parametrization with fewer parameters unless it significantly improves the goodness of the fit. The only difference between these two techniques is how this trade-off between complexity and goodness of the fit is realized: the AICc penalizes the number of parameters not as strongly as the BIC does. In both cases the preferred model is the one with a minimum value. Although the AICc has demonstrated to be both theoretically and practically advantageous in some applications (e.g. in regression problems) [13], we also monitor the BIC here.

B. Prediction

The second step is the model-based prediction step where we use the predictive capabilities of the selected model to estimate and predict the behavior of several underlying statistical quantities of interest which cannot be obtained from the original data, e.g. a mean-first-passage time (MFPT) of the process (see Appendix A for a detailed description of how to compute exit times). The key point of the proposed framework hence is that it is a synergistic interdisciplinary approach that combines elements from physics and mathematics, in particular statistical physics, theory of critical phenomena and stochastic processes. In the following we apply it first to a synthetic data set which is used as a test case, and second to two representative examples for which the underlying model is not known.

III. TEST CASE

To illustrate the estimation step of our data-driven coarse-graining framework, we perform a numerical experiment based on a computer-generated time series.

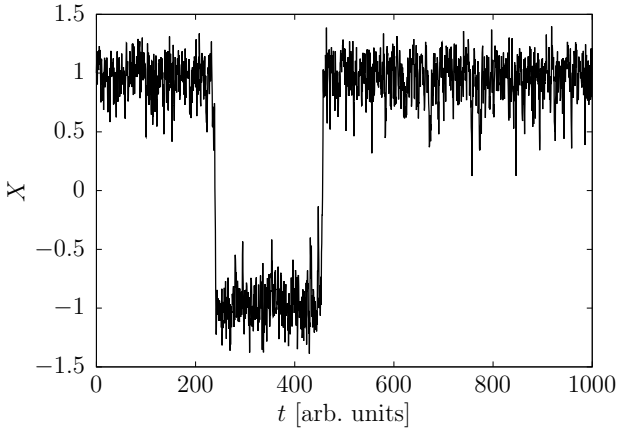

FIG. 2: Time series of SDE model (4) with $(\alpha,\beta,\gamma) = (1,-1,0.1)$ over the time interval $[0,1000]$ with sampling period $\Delta t = 0.5$.

Model	$f(x;\theta)$	$g(x;\theta)$
M1	$\sum_{j=0}^{1}\theta_j x^{2j+1}$	θ_2
M2	$\sum_{j=0}^{3}\theta_j x^{j}$	θ_4
M3	$\sum_{j=0}^{3}\theta_j x^{j}$	$\sqrt{\theta_4+\theta_5 x^2}$
M4	$\sum_{j=0}^{3}\theta_j x^{j}$	$\begin{cases}\theta_4 & ,\text{ if } x<\theta_6\\ \theta_5 & ,\text{ if } x\geq\theta_6\end{cases}$
M5	$\sum_{j=0}^{5}\theta_j x^{j}$	θ_6

TABLE I: A list of considered parametric SDE models for the test case.

Specifically, we consider the SDE

$$dX(t) = \Big(\alpha X(t) + \beta X(t)^3\Big)\,dt + \sqrt{\gamma}\,dW(t)\ , \qquad (4)$$

for which we set the true parameter vector θ^* to $\theta^* \equiv (\alpha,\beta,\gamma) = (1,-1,0.1)$. Figure 2 shows a time series on the time interval $[0,1000]$ with sampling period $\Delta t = 0.5$ (i.e. $n = 2001$ observations). The time series was obtained by integrating the SDE (4) numerically using the Euler–Maruyama scheme with a step size of $\delta t = 10^{-3}$; see e.g. [14] for details. The objective now is to fit an appropriate SDE model to this time series.

Since the time series shows a transition between two metastable states, we consider various candidate models that support metastability. Details of these parametric models are given in Table I. The outcome of the framework's model selection step is then summarized in Table II. Here the various models are compared with respect to the number of parameters m, the negative value of the log-likelihood function evaluated at the estimated parameter vector (i.e. $\hat{L} \equiv L_n(\hat{\theta}_n;\mathcal{X}_n)$), and the statistical model selection criteria. By comparing the selection criteria for the different models, we find that model M1 is clearly the preferred model among those considered. That is, the framework's model selection step does not only identify the underlying true SDE structure correctly,

Model	m	$-\ln(\hat{L})$	AICc	BIC
M1	3	943.56	−1881.11	−1864.31
M2	5	934.04	−1858.06	−1830.08
M3	6	838.72	−1665.41	−1631.84
M4	7	859.74	−1705.43	−1666.28
M5	7	853.03	−1692.00	−1652.85

TABLE II: Comparison of different estimated SDE models for the test case.

it also provides accurate estimates of the coefficients in Eq. (4) — the relative error of the estimated parameter vector being approximately 8% — despite the relatively high sampling period $\Delta t = 0.5$.

The results provided by our framework's selection are even more satisfactory when compared with the results obtained through a fully nonparametric (black-box) approach, something that has, e.g., been used in [4]. Here, both the drift f and the diffusion coefficient g are approximated using their infinitesimal definitions; cf. [15]. Specifically, for the drift function

$$f(x) = \lim_{t\to 0}\frac{1}{t}\mathbb{E}\Big(X(t) - X(0)\big|X(0) = x\Big) \approx \frac{\sum_{i=0}^{n-1} K\Big(\frac{x-X_i}{\kappa}\Big)(X_{i+1} - X_i)}{\Delta t \sum_{i=0}^{n-1} K\Big(\frac{x-X_i}{\kappa}\Big)} =: \hat{f}(x)\,,$$

is used, while the diffusion coefficient is approximated via

$$g(x)^2 = \lim_{t\to 0}\frac{1}{t}\mathbb{E}\Big(\big(X(t) - X(0)\big)^2\big|X(0) = x\Big) \approx \frac{\sum_{i=0}^{n-1} K\Big(\frac{x-X_i}{\kappa}\Big)(X_{i+1} - X_i)^2}{\Delta t \sum_{i=0}^{n-1} K\Big(\frac{x-X_i}{\kappa}\Big)} =: \hat{g}(x)^2\,,$$

where $K(z) = \exp(-z^2/2)/\sqrt{2\pi}$ and κ denotes a bandwidth parameter, which can be selected via least-squares cross validation techniques; see, e.g., [16, 17] for details. Clearly, both nonparametric estimators depend crucially on the size of the sampling period Δt and can only be expected to yield accurate estimates if $\Delta t \ll 1$. To illustrate that this intuitive statement is indeed correct, we depict in Fig. 3 the true drift and diffusion coefficients f and g together with the estimators obtained from our framework (i.e. model M1) and the ones obtained with a fully nonparametric approach (labeled by NP). While the results obtained from M1 provide also visually very accurate approximations of the true drift and diffusion coefficients used to generate the time series, the nonparametric counterparts deviate significantly from the true coefficients and show spurious and erratic effects.These effects are essentially due to the exclusively large sampling period Δt, as can be seen by repeating the same experiment with a very small sampling period (not shown here). As a consequence of these erratic effects, any further model-based analysis of the corresponding complex system using the fully nonparametric approach is bound to be ineffective since the artifacts associated with the sampling period would introduce nonphysical effects to the results.

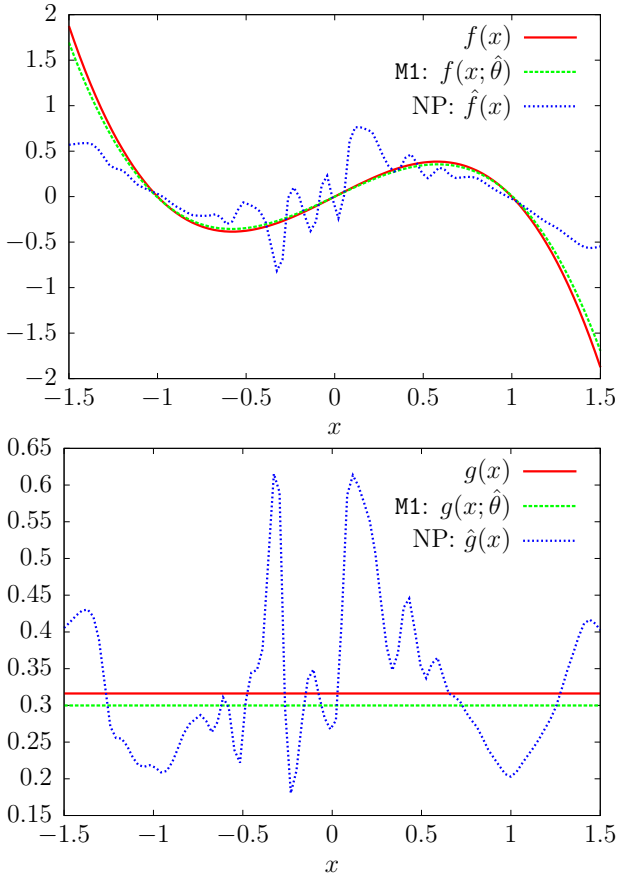

FIG. 3: (Color online) Comparison of estimated drift (top) and diffusion (bottom).

IV. REPRESENTATIVE EXAMPLES OF REAL-WORLD DATA SETS

A. Movement patterns of marine predators

The study of foraging behavior in marine life is an active research topic in ecology that has received considerable attention over the last few years. For example, analysis of the movement displacements of marine predators has suggested that, in certain cases, e.g. when prey is sparse, predators adopt an optimal search strategy based on Lévy flights [6, 18]. Understanding how such complex behavior is linked to, e.g., environmental conditions and the available prey distribution [19] or the predator's physiological capabilities [20], and, more importantly, how to predict it in terms of simple models, has become a major goal [21].

We consider the experimental observations of the movement pattern of an ocean sunfish (*Mola mola*) ob-

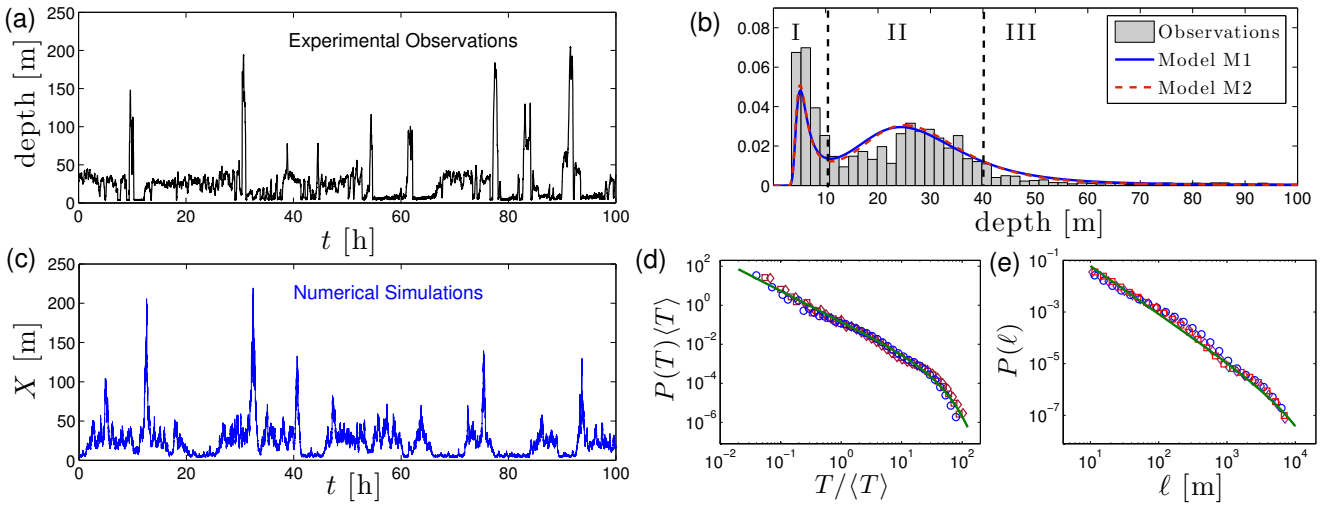

FIG. 4: (Color online) Analysis of marine predator movements: (a) Diving depth time series of an ocean sunfish (data from [6]). (b) PDF of the experimental observations (histogram in gray) and the numerical ones obtained from models M1 and M2. (c) Time series of the fitted coarse grained process X computed by using model M1. (d) PDF of the waiting times between large bursts of activity computed numerically by using model M1 and M2 (both models give the same results). The solid line corresponds to a fit with the function $P(z) = az^{-\gamma}\exp(-bz)$ with $\gamma = 1.54 \pm 0.06$. (e) PDF of the total diving length ℓ. The solid line corresponds to a truncated power law $P(\ell) \sim \ell^{-\mu}\exp(-\ell/L_0)$ with $\mu = 1.83 \pm 0.09$. Different points correspond to different values of the threshold, which is used to define the rare events (see text), in the range $X_{th} \in [35, 45]$.

tained by Humphries *et al.* [6] in a recent study to identify Lévy flights and Brownian movements in marine predators. Figure 4(a) shows the time series of the predator's diving depth (in positive value with respect to the sea surface) over a period of 4.5 days. It is evident that the predator's behavior is characterized by complex intermittent dynamics which we approximate by means of a stochastic model. To account for the sea surface as a natural boundary in the problem, and as the diving depth should be always non-negative, we change variables to $Y = \ln(X)$ so that Y solves an SDE of form (1) with drift and diffusion coefficients $\hat{f}$ and $\hat{g}$ which are expressed in terms of the new variable Y. As the result of the framework's model selection step we obtain the following two preferred models (see Appendix B for the full study comparing different models):

$$\texttt{M1}: \quad \hat{f}(Y;\theta) = \sum_{j=0}^{5}\theta_j Y^j; \quad \hat{g}(Y;\theta) = \theta_6,$$
$$\texttt{M2}: \quad \hat{f}(Y;\theta) = \sum_{j=0}^{7}\theta_j Y^j; \quad \hat{g}(Y;\theta) = \theta_8.$$

The dynamics of the diving depth is then given by $X = \exp(Y)$ with the following generic SDE:

$$dX(t) = f(X(t);\theta)\,dt + 2\sigma X(t)\,dW(t), \qquad (5)$$

which has multiplicative noise and where 2σ equals θ_6 or θ_8 in models M1 or M2, respectively. Figure 4(c) depicts an example of a time series generated from model M1, and Fig. 4(b) shows the theoretical PDFs associated with both models superimposed on the experimental histogram, observing a good match between them. The fact that the drift coefficient of model M1 is contained in the drift of model M2 together with the observation that the associated model PDFs are almost identical, indicates the robustness of the parametrization. It is important to emphasize that although our formulation is based on stochastic models, which can give rise to unrealistic local fluctuations at small time scales, it fully captures the macroscopic dynamics of the predator and the underlying quantities of interest.

We now use models M1 and M2 to accurately and confidently compute several quantities describing the dynamics of the predator. First, based on the bimodal PDF we define three regions of interest (habitats) as follows. Region I, which is the low-depth preferred habitat, corresponds to $X < X_{\rm I}$, where $X_{\rm I} = 10.5$ m is the local minimum between both peaks of the PDF. Region II, which is the deeper preferred habitat, is defined as $X_{\rm I} \le X < X_{\rm II}$, where $X_{\rm II} = 41.3$ m is the inflection point of the PDF for depths larger than the second maximum; and Region III, which consists of unlikely and rare events, is defined as depths $X \ge X_{\rm II}$ [see Fig. 4(b)]. Based on these definitions, we can compute several transition times, obtaining for example that, based on model M1 (model M2), the predator spends on average approximately $\tau = 1.24$ h ($\tau = 1.41$ h) in Region I before diving into II (see Appendix B). We look next at the PDF of the waiting times T between two consecutive deep dives, i.e. the periods for which $X \le X_{\rm th}$ at a stretch, where X_{th} is a chosen threshold (typically around $X_{\rm II}$). Figure 4(d) shows the results obtained with models M1 and M2 (both models give virtually identical results) observing that the PDF

of T (which is normalized to its mean value) follows a truncated power-law, $P(T) = aT^{-\gamma}\exp(-T/T_0)$, with exponent $\gamma \simeq 1.54$ which does not depend on X_{th}. Interestingly, this particular type of PDF (with exponent close to 3/2) has been observed ubiquitously in many different biological and physical systems exhibiting intermittent behavior (a signature usually of critical phenomena), from neuronal activity in the cortex [22], electroconvection of nematic liquid crystals [23], fluid flow in porous media [24] to colloidal quantum dots [25] and noise-induced transitions in infinite dimensional dissipative systems [26]. By studying the mean first passage time (MFPT) properties, the exponent 3/2 was obtained recently for SDEs of the form (5) with lineal multiplicative noise term [27].

Finally, we analyze the statistics of the total diving length during a rare event, which we denote as $\ell(\mathcal{X}_n)$, for a single time series $\mathcal{X}_n = (X_i)_{0\le i\le n}$. We define the total traveled length as $\ell \equiv \ell(\mathcal{X}_n) = \sum_{i=0}^{n-1} |X_{i+1} - X_i| \cdot w(X_i)$, where $w(z) = 0$ for $z \le X_{\rm II}$ and $w(z) = 1$ otherwise, and compute the PDF of ℓ. We conclude that for long distances it follows a truncated power law, $P(\ell) \sim \ell^{-\mu}\exp(-\ell/L_0)$ with exponent $\mu = 1.83 \pm 0.09$ [see Fig. 4(e)]. It is noteworthy that the statistics of ℓ follow a similar behavior with the statistics of the experimental step length defined in [6] where an exponent of $\mu = 1.92$ is reported indicating that the predator follows a Lévy flight description within a certain range step length.

B. Climate transitions during the last glacial period

Ice core records from Greenland reveal many intriguing phenomena of Earth's past climate and in particular records covering the last glacial period, approximately from 70 ky (1 ky = 1000 y) until 20 ky before present, are dominated by repeated rapid climate shifts, the so-called Dansgaard–Oeschger (DO) events [28], which are characterized by abrupt transitions from cold to warm periods. While the origin of these shifts is still actively debated [29], there seems to be a general consensus that DO events are transitions between two metastable climate states: a cold stadial and a warm interstadial state. Understanding how long it takes between DO events would potentially yield indicators for the causes, and earlier research based on previously obtained ice core records, reported a periodic occurrence of the DO events with period of approximately $\tau_{\rm DO} \approx 1.5$ ky [30], which has been subsequently refined to 1.47 ky [31, 32]. However, recent work based on the newer and more accurate North Greenland Ice Core Project (NGRIP) record has reported that there is not significant statistical evidence supporting the periodicity hypothesis and it is argued that DO shifts are most likely due to stochastic events [33, 34]. Here we use our data-driven framework to investigate the DO events without relying on the periodicity hypothesis.

We consider the δ^{18}O isotope record (as a proxy for Northern Hemisphere temperature) during the last glacial period which was obtained from the NGRIP [7], see Fig. 5(a). We observe a noisy temporal signal where the temperature increases up to a warm state until it abruptly goes down to a colder state (corresponding to an DO event), giving rise to a bimodal histogram, see Fig. 5(b). To account for transitions between two states, we consider two different parametrizations in the SDE model (1) (see Appendix C for full model candidate comparison):

$$\texttt{M1}:\ f(X;\theta) = \textstyle\sum_{j=0}^{3}\theta_j X^j;\quad g(X;\theta) = \theta_4,$$

$$\texttt{M2}:\ f(X;\theta) = \textstyle\sum_{j=0}^{3}\theta_j X^j; g(X;\theta) = \begin{cases}\theta_4 & \text{if } X < \theta_6\\ \theta_5 & \text{if } X \ge \theta_6\end{cases}.$$

Figure 5(b) depicts the model-based PDFs in comparison with the histogram of the original time series, illustrating a very good agreement between them. Due to its piecewise constant diffusion coefficient, the PDF associated with model M2 also captures the drop in the histogram around $X = -42$. It is noteworthy that although from a purely model selection criteria point of view model M1 appears to be marginally preferable (see Appendix C), M2 is a rather novel model in this field and shows strong statistical resemblance with the NGRIP data; something that should advocate the use of models with a state dependent diffusion coefficient also in other fields. Model M1 has been postulated before as a dynamical model for the NGRIP record [29, 35], however, in these studies, the accuracy of the model was not assessed and predictions were not made, as is done here. Moreover, the estimation procedure was ad hoc in that it made use of the same data set repeatedly several times.

Using the identified models, we compute the average time $\tau_{\rm DO}$ between DO events during the last glacial period by using the MFPT techniques described in Appendix A. In particular, we calculate the time $\tau_{\rm DO}$ as the average time to exit from a warm state plus the average time to exit from a cold state, obtaining $\tau_{\rm DO} \approx 1.602$ ky ($\tau_{\rm DO} \approx 1.511$ ky) from M1 (M2) which are in very good agreement with the values previously reported (the most accurate being 1.47 ky [31, 32]). Note, however, that these values were obtained by considering a deterministic periodic model, something recently questioned [33, 34], whereas the values reported here are from a purely stochastic model.

We next look at the statistics of both the residence times in the cooler state, i.e. the waiting times τ_w between DO events, and the duration τ_d of the DO events. To this end, we define a threshold X_{th} separating the two states to be at around the signal's mean value -42.13. Figures 5(d,e) show the PDFs for both magnitudes (normalized to their corresponding mean value) highlighting that they follow an exponential behavior, $P(z) = \exp(-z)$ for $z = \tau_w/\langle\tau_w\rangle$ or $\tau_d/\langle\tau_d\rangle$ which can be understood analytically as follows. We first note that the SDE for the variable X

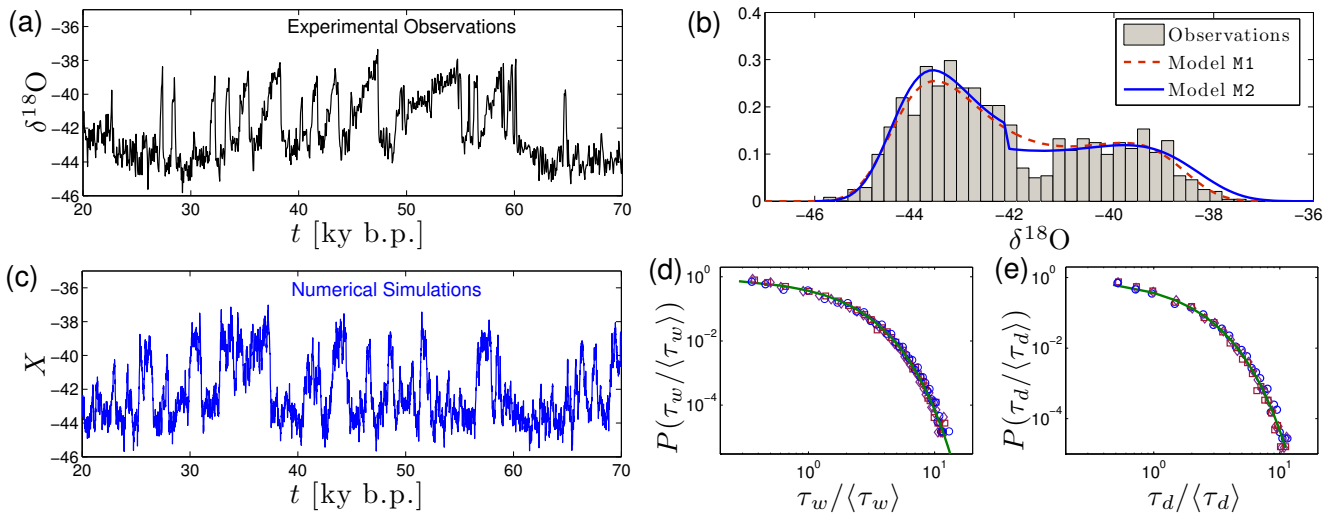

FIG. 5: (Color online) Analysis of paleoclimatic data during the last glacial period: (a) Paleoclimatic record time series [7]. (b) PDF of the experimental observations (histogram in gray) and the numerical ones obtained from model M1 and M2. (c) Time series of the fitted coarse-grained process X computed by using model M2. (d) and (e): PDF of the residence times τ_w at the cooler state and PDF of the duration τ_d of the DO events, normalized to their mean values, and for different values of the threshold ($X_{th} \in [-42.5, -42]$). The solid lines correspond to $P(z) = \exp(-z)$. As in Fig. 4 both models give very similar or almost identical results.

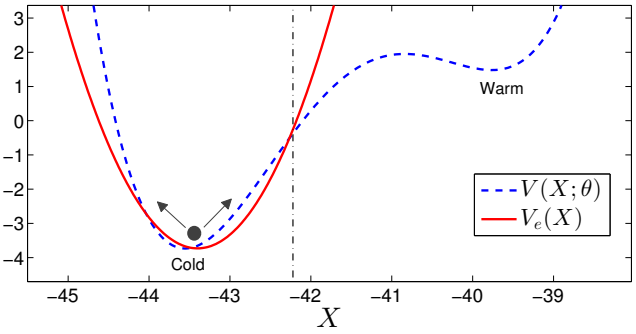

FIG. 6: Particle model for the climate transitions during the last glacial period: Schematic representation of the dynamics of a particle fluctuating around one of the local minima of the potential $V(X;\theta)$ obtained from the empirical data (dashed line). The solid line corresponds to an effective harmonic potential which is used to model the local dynamics before the particle jumps into the other equilibrium state. The dashed-dot line corresponds to the position of the threshold.

can be rewritten as:

$$dX = -\partial_X V(X;\theta)dt + g(X,\theta)dW(t), \qquad (6)$$

where $V(X;\theta)$ represents the potential function of the system so that $f(X;\theta) = -\partial_X V(X;\theta)$. From the empirical data we obtain that this potential is described by a function with two minima (see Fig. 6), one of them being stable (cold state) and the other one metastable (warm state). To study the statistics of the waiting times, consider the duration for which the variable X remains around one of the local minima until it jumps to the next state, the transition point of which is defined via the threshold. We can approximate this local dynamics with a simpler model corresponding to a particle fluctuating around a harmonic potential, i.e., we consider Eq. (6) with an effective potential given by:

$$V_e(X) = V_0 + \frac{a}{2}(X - X_0)^2, \qquad (7)$$

where V_0 corresponds to the minimum value of the potential and a and X_0 are effective parameters, respectively (Taylor expansion around the cold state), see Fig. 6. The dynamics of Eq. (6) can then be reduced to the well-known Ornstein–Uhlenbeck SDE $dX(t) = a\big(X_0 - X(t)\big)dt + \theta_4 dW(t)$, for which the first-passage properties are known to exhibit an exponential behavior [36, 37]. It is noteworthy that this type of process appears in many other areas such as mathematical finance [38] and neuronal dynamics [39], thus unifying seemingly unrelated complex systems. A similar argument can be used to explain the behavior of the duration time τ_d.

V. CONCLUSIONS

To conclude, we have presented a framework that allows to extract reliable statistical properties from a short set of available data (experimental observations) in a rational, systematic and efficient manner. Our approach aims to find a coarse grained (reduced) description of the full system which in turn necessitates the introduction of an appropriate stochastic process to account for

the unresolved degrees of freedom [40, 41]. The systems under consideration must be such that a coarse grained model exists, something which quite often can be rigorously proved by making use of multiscale techniques [42]. We have exemplified the methodology with two representative examples from the areas of marine biology and climate modeling. The two chosen examples belong to two generic classes of systems described by truncated power-law and exponential PDFs linked to the presence of multiplicative and additive noise, respectively, thus connecting complex systems whose dynamics is difficult if not impossible to model and subsequently understand, with well-studied stochastic processes. Moreover, the SDE models are versatile and enable us to consider both regular diffusion processes but also intermittent systems characterized by bursts of activity. The fact that fundamentally different phenomena can be described by the same type of model, Eq. (1), is a testimony of its wide applicability. Another key point is that the semiparametric approach we follow here is sufficiently flexible in that it allows other approaches, e.g. nonparametric which tend to be more restrictive, or even analytic if the governing model is known, to be easily adapted into our methodology. Our hope is that the outlined methodology can be applied to many other settings such as ranking processes [43] or cellular networks [44], to name but a few.

ACKNOWLEDGMENTS

We acknowledge financial support by the Engineering and Physical Sciences Research Council of the UK through Grants Nos. EP/H034587, EP/J009636, EP/K008595, EP/L020564, EP/L024926, EP/L025159, EP/L027186 and EP/N005465 as well as European Research Council through Advanced Grant No. 247031. We are grateful to Prof. David Sims (Marine Biological Association of the UK) for providing the data used in Example A.

Appendix A: Exit from a domain

For a given SDE model such as (1), we wish to compute the mean first passage time (MFPT), which is defined as follows. For a domain $D \subset \mathbb{R}^d$ we wish to know how long it takes on average for the process X to leave the domain D for the first time when the process is initially started at $x \in D$:

$$\tau(x) := \mathbb{E}\big(\inf\{t \geq 0 \colon X(t) \notin D\,, \;\; X(0) = x\}\big)\,. \quad \text{(A1)}$$

Note that if $x \notin D$, then $\tau(x) = 0$ by definition. To approximate τ one typically resorts to Monte-Carlo techniques based on numerically solving the SDE (1) [45]. For small dimensions (i.e. $d \leq 3$), an alternative way of approximating τ is to exploit the relation between statistical properties of the solution to SDE (1) and partial differential equations (PDEs). In fact, τ solves the deterministic PDE

$$f \cdot \nabla\tau + \frac{1}{2} g g^T \colon \nabla\nabla\tau = -1 \quad \text{in } D\,,$$

equipped with appropriate boundary conditions on ∂D; see, e.g., [15, 37]. The boundary conditions (e.g. reflection or absorption on ∂D) depend on the problem at hand, i.e. on the statistical property one is interested in. The fact that τ solves a PDE is particularly useful in one-dimension ($d = 1$). In this case, the two point boundary value problem can be solved analytically. Let $D := (l, r)$, then the MFPT $\tau(x)$, $x \in D$, can be written as

$$\begin{aligned}\tau(x) = &-2\int_l^x \int_l^y \frac{\exp\left(\psi(z) - \psi(y)\right)}{g(z)^2}\, dz\, dy \\ &+ c_1 \int_l^x \exp\left(-\psi(y)\right) dy + c_0\,,\end{aligned}$$

where $\psi(x) = 2\int_l^x g(z)^{-2} f(z)\, dz$ and the constants c_0, c_1 are determined from the boundary conditions. The accuracy of the approximation of τ obtained from this integral form is then given by the tolerance of the numerical quadrature rule, which is typically 10^{-8} [46].

Model	$\hat{f}(y;\theta)$	$\hat{g}(y;\theta)$
M1	$\sum_{j=0}^{5} \theta_j y^j$	θ_6
M2	$\sum_{j=0}^{7} \theta_j y^j$	θ_8
M3	$\sum_{j=0}^{3} \theta_j y^j$	θ_4
M4	$\hat{g}(y;\theta)^2 \sum_{j=0}^{3} \theta_j y^j$	$\sqrt{\theta_4 \exp(-\theta_5 y)}$
M5	$\hat{g}(y;\theta)^2 \sum_{j=0}^{5} \theta_j y^j$	$\sqrt{\theta_6 \exp\left(-\theta_7 y\right)}$

TABLE III: Specifications of the considered SDE models for the transformed process $Y = \ln(X)$.

Appendix B: Model Selection for Representative Example A

For the transformed (auxiliary) process $Y = \ln(X)$, we considered the parametric models shown in Table III. For these models, Table IV then summarizes the outcome of the framework's model selection step (cf. Fig. 1). Here the different models for the auxiliary process Y are compared with respect to the number of parameters m, the negative value of the log-likelihood function evaluated at the estimated parameter vector (i.e. $\hat{L} \equiv L_n(\hat{\theta}_n; \mathcal{X}_n)$), the statistical model selection criteria, and whether or not the estimated model provides a normalizable probability density function (PDF) for the original process X. The symbol "✓" in the last column of Table IV means that the estimated model provides a normalizable PDF, while the "×" means that it does not. As explained in the Sec. II, the preferred SDE parametrization is the one which has the smallest value with respect to a model

Model	m	$-\ln(\hat{L})$	AICc	BIC	PDF of X
M1	7	−59331.0	−118648.0	−118588.2	✓
M2	9	−59335.2	−118652.4	−118575.5	✓
M3	5	−59094.9	−118179.8	−118137.1	✓
M4	6	−74000.1	−147988.2	−147937.0	×
M5	8	−74366.3	−148716.6	−148648.3	×

TABLE IV: Comparison of different SDE models for the transformed process $Y = \ln(X)$ related to the foraging data of a marine predator.

selection criteria, i.e. with respect to AICc or BIC. Selecting a model based only upon the values of these statistical model selection criteria could result in the choice of an SDE model with unrealistic properties, in the sense that the model might not provide a normalizable PDF contrary to the empirical observations. Combining both aspects, Table IV thus reveals that models M1 and M2 are the two preferable models among those providing a normalizable PDF. In fact, the AICc selects model M2 as the preferred model, while the BIC favors model M1 (an interpretation of the magnitude of these differences is given in [47]).

1. Numerically computed transition times

Once we select the SDE model we can use it, of course, to study additional transition times between the different habitats of the marine predator compared the one presented in the main text. We recall the definitions of the marine predator's habitats. Based on the bimodal PDF three regions of interest (habitats) can be defined as follows (see Fig. 5b and associated text in the manuscript). Region I, which is the low-depth preferred habitat, is defined as depths which are shorter than the local minimum between peaks of the PDF that is located at $X_{\mathrm{I}} = 10.5$ m and so Region I corresponds to $X < X_{\mathrm{I}}$. Region II, which is the deeper preferred habitat, is defined as $X_{\mathrm{I}} \leq X < X_{\mathrm{II}}$, where $X_{\mathrm{II}} = 41.3$ m is defined as the inflection point of the PDF for depths larger than the second maximum. Finally, Region III, which consists of unlikely and rare events, is defined as the depths $X \geq X_{\mathrm{II}}$.

Based on these definitions, we look at how long it takes on average to make the transition from Region I to II. Specifically, based on model M1 (model M2), the individual spends on average approximately $\tau = 1.24$ h ($\tau = 1.41$ h) in lower depths corresponding to Region I before diving to deeper depths of Region II. Conversely, when situated in its deeper favorable habitat II, it takes on average approximately $\tau = 4.48$ h ($\tau = 4.87$ h) before ascending to Region I according to model M1 (model M2). On the other hand, we also look at the statistics of the rare events when the individual dives deeper into Region III. We compute the transition time that it takes for the individual to dive from Region II deep into Region III, specifically we consider dives to 150 m or deeper. We obtain that it takes on average approximately $\tau = 44.32$ h ($\tau = 48.18$ h) in view of model M1 (model M2).

Model	$f(x;\theta)$	$g(x;\theta)$
M1	$\sum_{j=0}^{3}\theta_j x^j$	θ_4
M2	$\sum_{j=0}^{3}\theta_j x^j$	$\begin{cases}\theta_4 & \text{, if } x < \theta_6\\ \theta_5 & \text{, if } x \geq \theta_6\end{cases}$
M3	$\sum_{j=0}^{5}\theta_j x^j$	θ_6
M4	$g(x;\theta)^2\sum_{j=0}^{3}\theta_j x^j$	$\sqrt{\theta_4\exp(-\theta_5 x)}$
M5	$g(x;\theta)^2\sum_{j=0}^{5}\theta_j x^j$	$\sqrt{\theta_6\exp(-\theta_7 x)}$
M6	$\sum_{j=0}^{5}\theta_j x^j$	$\sqrt{\theta_6+\theta_7(x-\theta_8)^2}$

TABLE V: Specifications of the considered SDE models for X describing the δ^{18}O isotope record.

Model	m	$-\ln\hat{L}$	AICc	BIC	PDF of X
M1	5	1123.4	2256.8	2281.3	✓
M2	7	1188.1	2390.4	2424.6	✓
M3	7	1111.3	2236.8	2271.0	×
M4	6	5180.0	10372.1	10401.4	✓
M5	8	4969.6	9955.3	9994.4	✓
M6	9	1193.7	2405.7	2449.6	✓

TABLE VI: Comparison of different SDE models for the climate data.

Appendix C: Model Selection for Representative Example B

For Example B we postulated the SDE models shown in Table V for the process X. The results of the model selection step are then summarized in Table VI, where these different SDE models for the process X are again compared with respect to the number of parameters m, the negative value of the log-likelihood function evaluated at the estimated parameter vector, the statistical model selection criteria, and whether or not the estimated model provides a normalizable PDF.

Comparing both model selection criteria in Table VI (i.e. AICc and BIC) for parametrizations that provide a normalizable PDF (indicated by the symbol "✓" in the last column), we find that model M1 offers the preferred parametrization with respect to these criteria. Due to its excellent agreement with the features shown by the histogram of the empirical data (see main text), we also selected model M2 for the subsequent analysis.

[1] M. Kac and J. Logan, in *Fluctuation Phenomena* (Elsevier, 1979) pp. 1–60.
[2] W. Horsthemke and R. Lefever, *Noise-induced transitions* (Springer, 1984).
[3] B. L. S. Prakasa Rao, *Statistical inference for diffusion type processes* (Arnold, 1999).
[4] R. Friedrich *et al.*, Phys. Rep. **506**, 87 (2011).
[5] P. Sura and J. Barsugli, Phys. Lett. A **305**, 304 (2002).
[6] N. E. Humphries *et al.*, Nature **465**, 1066 (2010).
[7] K. K. Andersen *et al.*, Nature **431**, 147 (2004).
[8] P. Billingsley, *Statistical inference for Markov processes* (Statistical Research Monographs, Vol. II. The University of Chicago Press, Chicago, Ill., 1961).
[9] D. Dacunha-Castelle and D. Florens-Zmirou, Stochastics **19**, 263 (1986).
[10] Y. Aït-Sahalia, Econometrica **70**, 223 (2002); Ann. Statist. **36**, 906 (2008).
[11] S. Krumscheid, G. A. Pavliotis, and S. Kalliadasis, Multiscale Model. Simul. **11**, 442 (2013).
[12] S. Kalliadasis, S. Krumscheid, and G. A. Pavliotis, J. Comput. Phys. **296**, 314 (2015).
[13] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference*, 2nd ed. (Springer-Verlag, 2002).
[14] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Applications of Mathematics, Vol. 23 (Springer, 1992).
[15] G. A. Pavliotis, *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations* (Springer, 2014).
[16] S. M. Iacus, *Simulation and Inference for Stochastic Differential Equations: With R examples* (Springer, 2008).
[17] D. P. Kroese, T. Taimre, and Z. I. Botev, *Handbook of Monte Carlo Methods* (John Wiley and Sons, 2011).
[18] D. W. Sims *et al.*, Nature **451**, 1098 (2008).
[19] B. A. Block *et al.*, Nature **475**, 86 (2011).
[20] S. R. Thorrold *et al.*, Nat. Commun. **5**, 4274 (2014).
[21] A. M. Reynolds and C. J. Rhodes, Ecology **90**, 877 (2009).
[22] W. L. Shew *et al.*, J. Neurosci. **29**, 15595 (2009).
[23] T. John, R. Stannarius, and U. Behn, Phys. Rev. Lett. **83**, 749 (1999).
[24] M. Pradas, J. M. López, and A. Hernández-Machado, Phys. Rev. E **80**, 050101 (2009); J. M. López, M. Pradas, and A. Hernández-Machado, **82**, 031127 (2010).
[25] P. Frantsuzov *et al.*, Nat. Phys. **4**, 519 (2008).
[26] M. Pradas *et al.*, Phys. Rev. Lett. **106**, 060602 (2011).
[27] M. Pradas *et al.*, Eur. J. Appl. Math. **23**, 563 (2012).
[28] W. Dansgaard *et al.*, Nature **364**, 218 (1993).
[29] F. Kwasniok, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **371**, 20110472, 22 (2013).
[30] P. M. Grootes and M. Stuiver, J. Geophys. Res. Oceans **102**, 26455 (1997).
[31] M. Schulz, Paleoceanography **17**, 4 (2002).
[32] S. Rahmstorf, Geophys. Res. Lett. **30**, 1510 (2003).
[33] P. D. Ditlevsen, K. K. Andersen, and A. Svensson, Clim. Past **3**, 129 (2007).
[34] P. D. Ditlevsen and O. D. Ditlevsen, J. Climate **22**, 446 (2009).
[35] F. Kwasniok and G. Lohmann, Phys. Rev. E **80**, 066104 (2009).
[36] L. Alili, P. Patie, and J. L. Pedersen, Stochastic Models **21**, 967 (2005).
[37] C. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences* (Springer, 2009).
[38] C. Y. Tang and S. X. Chen, Journal of Econometrics **149**, 65 (2009).
[39] R. M. Capocelli and L. M. Ricciardi, Kybernetik **8**, 214 (1971).
[40] M. Schmuck *et al.*, Phys. Rev. Lett. **110**, 244101 (2013).
[41] M. Schmuck *et al.*, IMA J. Appl. Math. (2013).
[42] G. A. Pavliotis and A. M. Stuart, *Multiscale Methods: Averaging and Homogenization* (Springer, 2008).
[43] N. Blumm *et al.*, Phys. Rev. Lett. **109**, 128701 (2012).
[44] I. Y. Wong *et al.*, Phys. Rev. Lett. **92**, 178101 (2004).
[45] D. Higham *et al.*, SIAM/ASA J. Uncertainty Quantification **1**, 2 (2013).
[46] L. F. Shampine, J. Comput. Appl. Math. **211**, 131 (2008).
[47] R. E. Kass and A. E. Raftery, J. Amer. Statist. Assoc. **90**, 773 (1995).