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

S. Krumscheid[1,2], M. Pradas[2], G.A. Pavliotis[1], and S. Kalliadasis[2],
[1] *Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom*
[2] *Department of Chemical Engineering, Imperial College London, London SW7 2AZ, United Kingdom*

(Dated: September 5, 2014)

In many natural, technological, social and economic applications, one is commonly faced with the task of estimating statistical properties from empirical data (experimental observations), such as mean-first-passage times of a temporal continuous process. 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 observed rarely. We propose here a theoretical-computational framework which enables the systematic and rational estimation of statistical quantities of a given temporal process, such as waiting times between subsequent bursts of activity. Our framework is illustrated with applications from real-world data sets, ranging from marine biology to climate change.

Over the last few years, there has been an increasing demand for capturing generic statistical properties of complex systems based on available data only. The term complex systems here refers to a class of problems for which the number of variables is either sufficiently large and/or each of the variables has a behavior which is individually erratic or totally unknown, but in spite of this, the system as a whole can posses orderly and analyzable average properties.

Such systems are strongly influenced by random fluctuations which play a crucial role in the various intriguing phenomena emerging in temporal observations [20, 23], e.g. dynamic state transitions. Understanding the underlying complex processes of such phenomena is a common task in many disciplines, but often it is not even possible to estimate statistical properties directly from empirical data alone because e.g. the phenomenon of interest occurs rarely. In a purely reductionist approach, one could try to derive the governing equations of the process from first principles. This bottom-up approach, however, is often impossible and in the few cases where it can be done, the resulting mathematical models are fairly complicated due to the high-level of detail involved in their derivation which makes them computationally prohibitive.

An alternative approach is to identify a reduced (coarse grained) model on the basis of the experimental data which retains the fundamental aspects of the original system. This is in fact at the core of data-driven coarse-graining methodologies but despite their fundamental significance, to date there does not exist a rational and systematic framework for obtaining coarse-grained models from empirical observations. Relying exclusively on the observations and treating the corresponding reduced model as a "black box" (that is, in technical terms using nonparametric estimators [36], see also [16] for a review of such techniques) is, however, not reasonable since such an approach is typically rather crude and introduces errors in regions where only few observations exist, such as in rare phenomena (see also discussion in [44]), thus corrupting model-based predictions. A 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" modelling 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 modelling based on the above semiparametric approach with the ultimate aim of analysing complex phenomena arising in a wide spectrum of different systems. A schematic representation of our methodology is shown in Fig. 1 which consists of two main steps: A model selection (postulate - assess/validate) procedure, which allows to select a simple coarse grained model, and a second prediction step where the predictive capability of the selected model is tested, i.e. we use it to find and predict both analytically and numerically the behaviour of several underlying statistical quantities of interest which cannot be obtained from the original data. The key point of the proposed methodology is that it is a synergistic interdisciplinary approach that combines elements from critical phenomena physics, statistical physics and stochastic processes. To exemplify the methodology we use two representative examples of current interest, namely experimental observations of the foraging behaviour of marine predators [21], and the temperature record during the last glacial period [4].

Generic data-driven modelling framework

We are mainly interested in systems where the underlying noisy process is continuous with respect to time and so we focus on simple continuous-time models which are solutions of the following prototypical stochastic differential equation (SDE) (see e.g. [24, 30] for an introduction) that represents a diffusion process:

$$dX = f(X;\theta)\,dt + g(X;\theta)\,dW_t\ , \tag{1}$$

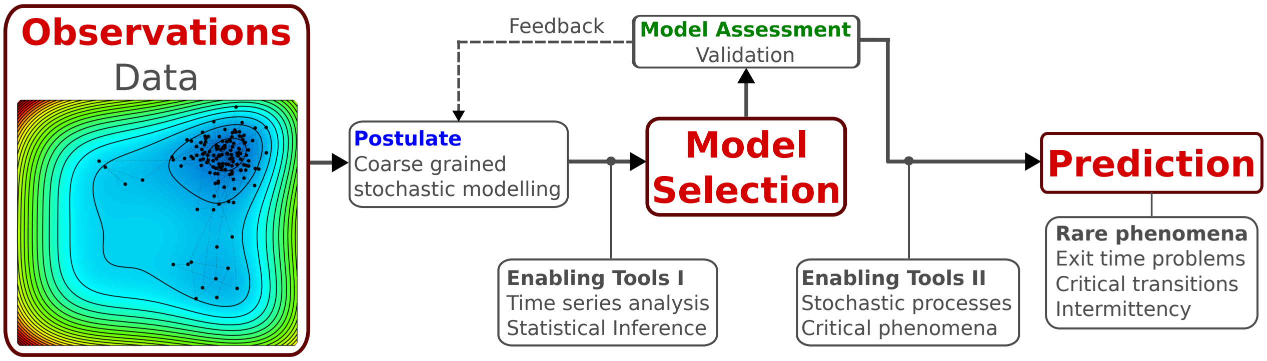

FIG. 1: **Flow chart of the data-driven modelling framework.** Given observations (data) we postulate a coarse-grained stochastic parametric model which is fitted (via statistical inference and time series analysis tools) to the data and refined via a model selection process. In particular, via an assessment/validation 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 of interest by making use of critical phenomena, statistical physics and stochastic processes tools. The far-left figure is a numerical example of Brownian motion in a two-dimensional potential.

which is understood in Itô's convention (see e.g. [32]), with initial condition $X(0) = x \in \mathbb{R}^d$, where X describes a fluctuating variable of interest, and f and g are the so-called drift and diffusion coefficients, respectively, with the latter controlling the influence of the stochastic driving through a Wiener process, W_t. Given a series of experimental observations, we wish to fit the above model to the observations. The unknown model parameter vector θ is then estimated by using a maximum likelihood methodology (details are given in the Methods section). Using this data-driven modelling framework it is possible to identify simple (i.e. small number of unknown parameters) yet adequate, models which retain the essential statistical properties of the actual process. Once several possible model candidates are identified, we proceed with a model selection procedure which statistically compares and assesses the different models (more details are given in the Methods section). This allows us to narrow our search down to two possible models which are similar to each other and are good candidates to describe the observed complex phenomena.

We then validate these two models by computing quantities that can be directly compared with the experimental data, for example the probability distribution function (PDF). After this final step, we can study and predict different statistical properties and quantities of interest, such as exit times, or the mean-first-passage time (MFPT) of X solving (1) (details are given in the Methods section). In the following we apply our methodology to two seemingly unrelated examples of complex systems which are of fundamental significance in current topical research areas, such as ecology and climate change.

Representative Example I: Movement patterns of marine predators

The study of foraging behaviour 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 the prey is sparse, predators adopt an optimal search strategy based on Lévy flights [21, 43] – a special case of a random walk for which the movement displacements follow a PDF with a power-law tail. Understanding how such complex behaviour is linked to, e.g., the environment conditions and the available prey distribution [6] or the predator's physiological capabilities [46], and, more importantly, how to predict it in terms of simple statistical models, has become a major goal (see e.g. [38] and references therein).

Observations.- We consider the experimental observations of the movement pattern of an ocean sunfish (*Mola mola*) obtained by Humphries *et al.* [21] in a recent study to identify Lévy flights and Brownian movements in marine predators. Figure 2(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. The data set contains $n = 37800$ observations with temporal sampling rate $\Delta t = 10$ s. It is evident that the predator's behaviour is characterized by a complex intermittent dynamics which reveals at least two interesting patterns. Firstly, the individual seems to have two preferred habitats, i.e. depths where it spends most of its time, characterized by two main peaks in the histogram of the data [see Fig. 2(b)] located at depths of approximately 5.5 m and 26.5 m. The second interesting phenomenon is that the individual, on rare occasions, un-

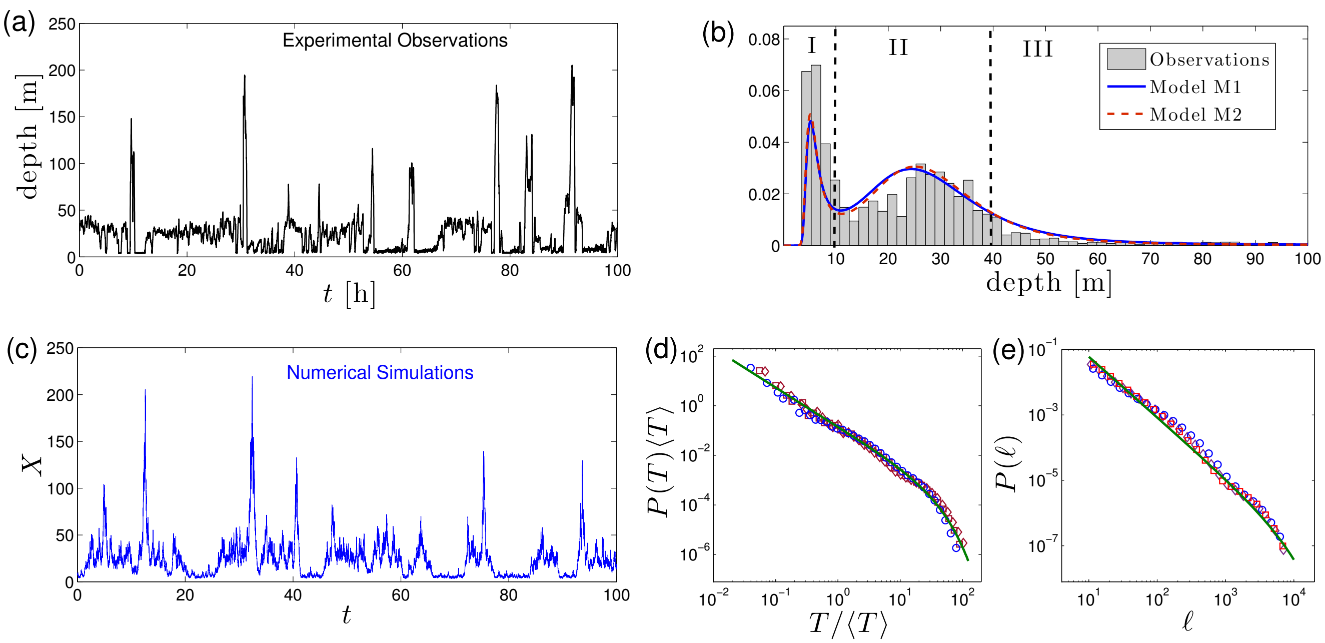

FIG. 2: **Analysis of marine predator movements**. (a) Diving depth time series of an ocean sunfish (Mola mola) (data is obtained from [21]). (b) PDF of the experimental observations (histogram in gray) and the numerical ones obtained from models `M1` and `M2`. The form of the PDF can be used to demarcate three habitats, I, II and III. (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 for the waiting times). The solid line corresponds to a fit with the function $P(T) = az^{-\gamma}\exp(-bz)$ with exponent $\gamma = 1.54 \pm 0.06$. (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 exponent $\mu = 1.83 \pm 0.09$. The different points in (d,e) correspond to different values of the threshold, namely $X_{th} = 25$ (○), 30 (□), and 35 (◇),

dertakes dives into regions which are significantly deeper than its favourite habitats. To perform a statistical analysis of this behaviour we approximate it by means of a stochastic model.

Model Selection.- We first note that to postulate a simple model for the diving depth dynamics one needs to take into account the sea surface as a natural boundary in the problem so that the diving depth is always non-negative. To this end, we consider the change of variable $Y = \ln(X)$ so that Y solves an SDE of form (1) with drift function $\hat{f}$ and diffusion coefficient $\hat{g}$. These coefficients are expanded in terms of the new variable Y so that we have the two following models (see Supplementary Material for the full study comparing different models):

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

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

which differ from each other on the number of parameters: 7 for `M1` and 9 for `M2`. The dynamics of the diving depth is then given by $X = \exp(Y)$, and its corresponding SDE can be generically written as:

$$dX = f(X;\theta)\,dt + 2\sigma X\,dW_t, \tag{2}$$

which has a multiplicative noise and where 2σ equals either θ_6 or θ_8 in models `M1` or `M2`, respectively. Note that function $f(X;\theta)$ is obtained from $\hat{f}(Y;\theta)$ after changing back to the variable X. Figure 2(c) depicts an example of a time series generated from model `M1`, and 2(b) the theoretical PDFs associated with both estimated models superimposed on the experimental histogram. We observe that the numerically generated time series exhibits a similar behaviour as the one observed experimentally, and that there is a good match between the model PDFs and the time series. In fact, they nicely reproduce the bimodal nature of the empirical data in terms of both locations and values. The fact that the drift function 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. Both models are also very similar in view of the statistical model selection criteria (see Section 2 in Supplementary Material). It is important to emphasize that although this formulation is based on stochastic models, which can give rise to unreal local fluctuations at small time scales, it fully captures the macroscopic dynamics of the predator and the underlying quantities of interest.

Prediction.- 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, is defined as depths which are shorter than the local minimum at $X_{\mathrm{I}} = 10.5$ m between the two peaks of the PDF 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 $X \geq X_{\mathrm{II}}$ [see Fig. 2(b)].

We now 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 favourable 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, it is interesting to investigate the statistics of the rare events when the individual dives deeper into Region III. We first compute the transition time it takes for the predator to dive from Region II to deep water into Region III, specifically we consider dives to 150 m. We obtain that on average it takes approximately $\tau = 44.32$ h ($\tau = 48.18$ h) in view of model `M1` (model `M2`). We look next at the distribution of the waiting times between two consecutive deep depths. In particular, we define the waiting time T as the time the individual is in depths smaller than X_{II} (i.e. $X \leq X_{\mathrm{II}}$) before migrating from Region II to Region III. Figure 2(d) shows the results obtained with models `M1` and `M2` (both models give the same results) observing that the PDF of T (which is normalised to its mean value) follows a truncated power-law distribution, $P(T) = aT^{-\gamma}\exp(-T/T_0)$, with exponent $\gamma \simeq 1.54$ which does not depend on the chosen threshold value, denoted as X_{th}.

Interestingly, this particular type of power-law distribution (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 [42], electroconvection of nematic liquid crystals [22], fluid flow in porous media [29, 33] to colloidal quantum dots [15] and additive noise-induced transitions in dissipative systems [35]. Analytically, the power-law behavior can be understood by considering the first passage properties of the linearised SDE (2) around a small value X_0 below the threshold X_{II}, which to leading order corresponds to an underlying random walk process that follows the SDE: $dY = \alpha dt + 2\sigma dW_t$, where $\alpha = \partial_X f(X;\theta)|_{X=X_0}$. By looking then at the first-passage properties of the random walk in a semi-infinite domain one can show that the waiting times PDF follows a truncated power-law distribution with exponent 3/2 [34].

Finally, we can also analyse the statistics of the total diving length of the predator during a rare event, which we denote as $\ell(X)$, for a single trajectory X_i for $i = 0, \ldots, n$, where $n = \mathsf{T}/\Delta t$ with T being the final time. In particular, we define the total travelled length as $\ell \equiv \ell(X) = \sum_{i=0}^{n-1} |X_{i+1} - X_i| \cdot w(X_i)$, where $w(X_i)$ is a function which is zero for $X_i \leq X_{\mathrm{II}}$ and one otherwise. We compute the PDF of ℓ obtaining that for long distances it follows a truncated power law, $P(\ell) \sim \ell^{-\mu}\exp(-\ell/L_0)$ with an exponent $\mu = 1.83 \pm 0.09$ [see Fig. 2(e)]. It is noteworthy that the statistics of ℓ follows a similar behaviour with the statistics of the step length defined in [21] where an exponent of $\mu = 1.92$ is reported indicating the predator follows a Lévy searching description within a certain range step length.

Representative Example II: 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 [12], which are characterised by abrupt warmings. While the origin of these shifts is still actively debated [27], there seems to be the general consensus that DO events are transitions between two metastable climate states: a cold stadial and a warm interstadial state. In addition to identifying the underlying causes, it is also vital to understand how long it takes between DO events, as this would potentially yield indicators for the causes. Earlier research on this problem, based on previously obtained ice core records, reported a periodically occurrence of the DO events with period of approximately $\tau_{\mathrm{DO}} \approx 1.5$ ky [18], which has been subsequently refined to 1.47 ky [37, 41]. In a recent work, based on the newer full North Greenland Ice Core Project (NGRIP) record with its more accurate dating, it has been reported that there is not significant statistical evidence supporting the periodicity hypothesis of the DO events. Moreover, it is argued that these climate shifts are most likely due to stochastic events [13, 14]. Here we use our data-driven framework to investigate the DO events during the last glacial period without relying on the periodicity hypothesis.

Observations.- 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, Greenland's newest ice core [4], consisting of $n = 1000$ observations with temporal sampling rate of

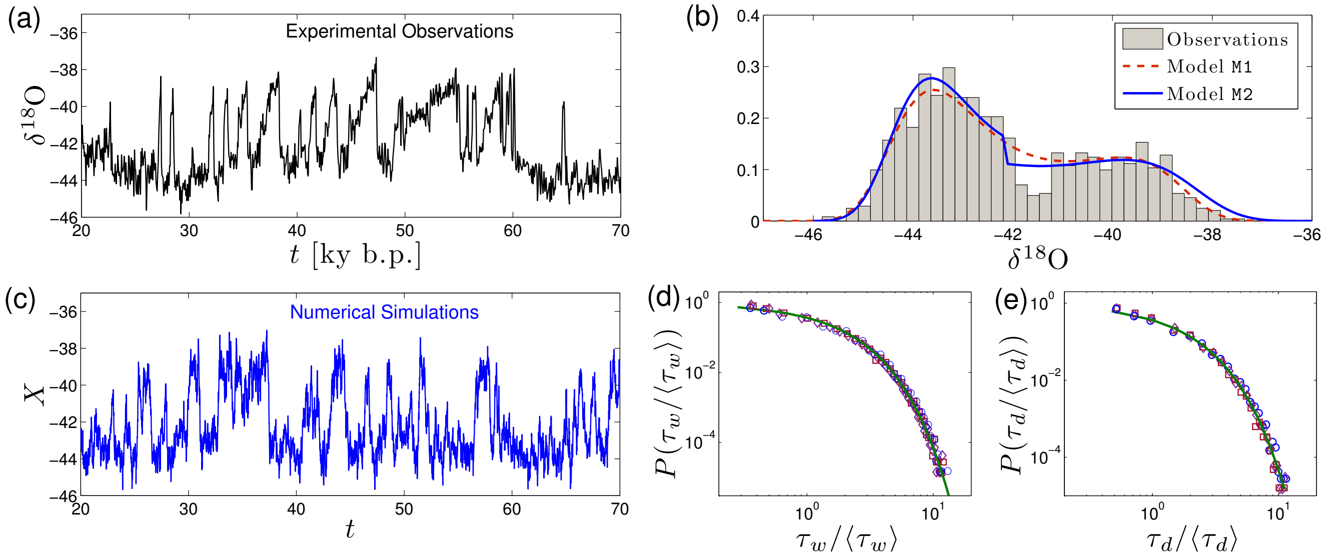

FIG. 3: **Analysis of paleoclimatic data during the last glacial period.** (a) Paleoclimatic record time series [4]. (b) PDF of the experimental observations (histogram in gray) and the numerical ones obtained from model M1 and M2. (c) Time series of the fitted coarse-grained process X computed by using model M2. (d) and (e) PDF of the residence times τ_w for which the solution is in the cooler state and PDF of the durations τ_d of the DO events, normalised to their corresponding mean values and for different values of the threshold, namely $X_{th} = -42$ (○), -42.2 (□), and -42.5 (◇). The solid lines correspond to the exponential function $P(z) = \exp(-z)$.

$\Delta t = 0.05$ ky [see Fig. 3(a)]. We observe a noisy temporal signal which is characterised by transitions between two states. The periods when the temperature increases up to the warm state until it abruptly goes down to the cold state corresponding to the DO events, and these two states give rise to a bimodal histogram, see Fig. 3(b).

Model Selection.- To account for transitions between two states, we consider two different parametrisations in the SDE model (1):

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

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

Note that model M2 generalises model M1 to a piecewise constant diffusion coefficient (again see Supplementary Material for the full study comparing different models). Figure 3(b) depicts the model-based PDFs in comparison with the histogram of the original time series, observing 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 $X = -42$ [see Fig. 3(c)]. It is noteworthy that although from a purely model selection criteria point of view model M1 appears to be marginally preferable (see Supplementary Material), 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 non-constant diffusion function also in other fields). M1 has been postulated before as a dynamical model for the NGRIP record [27, 28], 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.

Prediction.- Using the identified models, we compute the average time τ_{DO} between DO events during the last glacial period by using the techniques described in the Methods section. In particular, we calculate the time τ_{DO} as the average time to exit from a warm state plus the average time to exit from a cold state. For model M1 this approach results in $\tau_{\mathrm{DO}} \approx 1.602$, while for model M2 in $\tau_{\mathrm{DO}} \approx 1.511$. Both values, especially the one obtained with model M2 are in very good agreement with the values previously reported in the literature (as noted earlier, the most accurate value was 1.47 ky reported in [37, 41]). It is important to note, however, that this previously reported value was obtained by considering a deterministic periodic model, something that has been recently questioned [13, 14], whereas the value we obtain here is from a purely stochastic model which is derived via a data-driven framework.

We next look at the statistics of both the residence times in the cooler state, i.e. the waiting times between DO events which we denote as τ_w, and the durations of the DO events, i.e. the residence times in the warmer state, which are denoted as τ_d. To this end, we define a threshold X_{th} separating the two states to be at around -42.13 which corresponds to the mean value of the signal so that τ_w correspond to consecutive times for which $X \leq X_{th}$ and τ_d to consecutive times for which $X > X_{th}$. Figures 3(d,e) show the PDFs for both magnitudes (normalised to their corresponding mean value) observing that they follow an exponential behaviour, $P(z) = \exp(-z)$ for $z = \tau_w/\langle\tau_w\rangle$ or $\tau_d/\langle\tau_d\rangle$. Such exponential behaviour can be understood analytically as follows. First, we note that the waiting times τ_w are characterised by periods of time for which the solution is locally fluctuating around the stable cooler state before jumping to the warmer state. We can approximate such local dynamics as fluctuations of a particle around an effective harmonic potential which we express as $V_e = V_0 + (1/2)V_e''(X_0)(X - X_0)^2$, where X_0 corresponds to the cooler stable state and V_0 is its minimum value (see Supplementary Material for more details). Hence, we have that the local dynamics around X_0 follows the SDE, $dX = a(X_0 - X)dt + \theta_4 dW_t$ with $a = V_e''(X_0)$. After changing to $Y = X - X_0$, we obtain an underlying local process which is given by the well-known Ornstein-Uhlenbeck equation, a model for Brownian motion with friction: $dY = -aY dt + \theta_4 dW_t$, the first-passage properties of which are known to exhibit an exponential behaviour [3, 17]. It is noteworthy that this type of process appears in many other areas such as mathematical finance [45] and neuronal dynamics [10]. A similar argument for the local dynamics around the warmer state can also be applied so that the durations τ_d also follow an exponential behaviour.

Discussion

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. By combining tools from statistical inference and time series analysis we are able to assess a selection of different models which are fitted to the data. Once the best model is selected and validated we use it to trace the underlying statistical properties of the system under consideration which are not accessible from the experimental observations. 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 [39, 40]).

We have exemplified the methodology with two representative examples relevant in different areas, namely marine biology and climate prediction. We have first analysed the movements of a particular marine predator [21] which exhibits a complex intermittent behaviour. Our methodology has shown that the dynamics of the predator can be fitted into a reduced stochastic model with a multiplicative noise term from which we have been able to extract information about the statistics of the times spent in the different preferred habitats as well as of the rare events for which the predator dives into deep depths, observing that the waiting times PDF follow a truncated power-law with exponent 3/2, a behaviour which is ubiquitously observed in many other systems. As a second example, we have analysed the ice-core record during the last glacial period [4] which exhibits repeated rapid climate shifts, the so-called Dansgaard–Oeschger (DO) events. We have shown that such events can be described by a stochastic model with (piecewise) additive noise, obtaining that the average time between two consecutive DO events is 1.51 ky, which is in agreement with a previously reported value obtained using a periodic model [37, 41]. We have also analysed the PDF of both the waiting times between DO events and their durations showing that they both follow an exponential behaviour, a behavior which is observed in a wide spectrum of other systems.

The two examples analysed here thus 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. Moreover, in all cases diffusion is the underlying ubiquitous process for complex systems which can be described statistically, including the usual Brownian motion but also intermittent systems characterised by bursts of activity. The fact that fundamentally different phenomena can be described by the same model, Eq. (1), is a testimony of the wide applicability of the model. Of course as emphasised before Eq. (1), the noisy process is taken to be continuous with respect to time and hence it might not be able to reproduce all quantities of interest accurately from the outset (e.g. PDF exponents) for processes which are not continuous, such as pure Lévy flights, but it will capture at least part of the actual PDF behavior and as such it can be used as a first step in the analysis of a time series, a diagnostic tool to characterise the time series. For an accurate description of the quantities of interest then the noisy process in Eq. (1) would have to be viewed in a more general context and the Wiener process would have to be replaced with a discontinuous one, such as a Lévy process, but the general framework in Fig. 1 remains unaltered.

Our hope is that the outlined methodology can be applied to many other settings such as ranking processes [7] or cellular networks [47], to name but a few. The systems under consideration must be such that, either there isn't a macroscopic model, or it is difficult to obtain it, but due to the underlying multiscale structure of the systems [31], the global dynamics can be described by a coarse-grained formulation. Another key point is that the semiparametric approach we follow here is sufficiently flexible in that it allows other approaches, e.g. nonparametric which is a more restrictive approach, or even analytic if the governing model is known, to be easily adapted into the formulation. Understanding complex systems requires an arsenal of tools from different disciplines such as critical phenomena physics, statistical physics and stochastic processes and all these are brought together in our methodology. We expect that our results will improve the understanding of complex systems and will open a new systematic way for characterising their statistical properties.

METHODS

Parametric Inference for SDEs

To fit the parametrised model SDE (1) to available discrete-time observations, we estimate the parameter vector $\theta \in \Theta \subset \mathbb{R}^m$ using a maximum likelihood framework due to its favourable theoretical properties; see e.g. [5, 11] (refer to [26, 36] for an overview of alternative methods). Specifically, let $\mathcal{X}_n := \big(X(t_i)\big)_{0\leq i\leq n}$ be the sample of discrete-time observations of (1) with true parameter θ^*, at times $0 = t_0 < t_1 < \cdots < t_n = \mathsf{T}$. The maximum likelihood estimator (MLE) for θ^* based on $\mathcal{X}_n$ is then given as any (if not unique) element that maximises the so-called likelihood function over Θ. That is, $\hat{\theta}_n \in \arg\max_{\theta\in\Theta} L_n(\theta; \mathcal{X}_n)$, where $L_n(\theta; \mathcal{X}_n)$ denotes the likelihood function based on the observed data $\mathcal{X}_n$, given by $L_n(\theta; \mathcal{X}_n) = \prod_{i=0}^{n-1} p_\theta\big(t_{i+1} - t_i, X(t_{i+1})|X(t_i)\big)p_\theta\big(X(0)\big)$. Therein $p_\theta(x)$ denotes the probability density function of the initial condition and $p_\theta(\Delta t, x|y)$ denotes the conditional density function, i.e. transition density associated withe the SDE, of value x being reached in Δt time units when currently being at state y. As the transition density $p_\theta(\cdot,\cdot|\cdot)$ is only rarely known in closed-form, one has to approximate it to make this approach feasible in practice. In this work we adopt the closed-form expansion due to Aït-Sahalia [1, 2]. 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. After inverting the transformation, the expansion of $p_\theta(\cdot,\cdot|\cdot)$ is given in closed form. The coefficients determining this expansion depend on the considered functional form of both drift and diffusion in (1) and can become rather involved. Using a careful combination of symbolic and numerical computations, it is possible nonetheless to evaluate these coefficients.

It is noteworthy that, while the MLE worked well for the data sets used in this work, it becomes biased for examples with multiple time scales. In these circumstances specialist methods, such as those in [25], are more appropriate and should be used.

Model Selection

There is a wide range of model selection criteria available in the literature, which are used to statistically compare different model parameterisations against each other [8]. Here we use two techniques, both relying on the maximised likelihood function of the considered model and the available observations $\mathcal{X}_n$, i.e. they depend on $L_n\big(\hat{\theta}_n; \mathcal{X}_n\big)$. Let $\hat{\theta}_n$ be the estimated m-dimensional parameter vector in the SDE model (1). Then we use the finite sample size corrected Akaike Information Criterion (AICc), given by AICc $= 2m(n+1)/(n-m) - 2\ln\big(L_n(\hat{\theta}_n; \mathcal{X}_n)\big)$. Furthermore, we use the Bayesian Information Criterion (BIC), defined as BIC $= m\ln(n+1) - 2\ln\big(L_n(\hat{\theta}_n; \mathcal{X}_n)\big)$. Both criteria provide a measure of the relative quality of the SDE parametrisation (1) (i.e. of the statistical model), based on the given set of data. In particular, they are designed to penalise over-fitted models, i.e. a parametrisation with many parameters is not as valuable as a parametrisation with fewer parameters, unless it significantly improves the goodness of the fit. The only difference between these techniques is how this trade-off between complexity and goodness of the fit is realised: the AICc penalises the number of parameters not as strongly as does the BIC. In both cases the preferred criterion is the one with a minimum value. Although the AICc has sometimes theoretical advantages and can be practically advantageous [8], we also monitor the BIC.

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 $x \in D$:

$$\tau(x) := \mathbb{E}\big(\inf\{t \geq 0\colon X(t) \notin D\ ,\ \ X(0) = x\}\big)\ . \qquad (3)$$

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). For example, recently a multilevel Monte Carlo method has been introduced, which significantly reduces the computational cost over the costs for standard Monte Carlo approaches [19]. 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 PDE theory. In fact, τ solves the deterministic PDE

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

equipped with appropriate boundary conditions on ∂D. 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 that case, the PDE reduces to an ODE and can be solved analytically. In fact, 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. This explicit representation of τ is not only of practical interest, but also amenable to a simplified mathematical analysis of the MFPT, such as a sensitivity analysis with respect to the parametrisation [9].

[1] Y. Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. Econometrica, 70(1):223–262, 2002.

[2] Y. Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. Ann. Statist., 36(2):906–937, 2008.

[3] L. Alili, P. Patie, and J. L. Pedersen. Representations of the first hitting time density of an Ornstein-Uhlenbeck process 1. Stochastic Models, 21(4):967–980, 2005.

[4] K. K. Andersen, N. Azuma, J. M. Barnola, M. Bigler, P. Biscaye, N. Caillon, J. Chappellaz, H. B. Clausen, Dahl D. Jensen, H. Fischer, J. Flückiger, D. Fritzsche, Y. Fujii, Goto K. Azuma, K. Gronvold, N. S. Gundestrup, M. Hansson, C. Huber, C. S. Hvidberg, S. J. Johnsen, U. Jonsell, J. Jouzel, S. Kipfstuhl, A. Landais, M. Leuenberger, R. Lorrain, Masson V. Delmotte, H. Miller, H. Motoyama, H. Narita, T. Popp, S. O. Rasmussen, D. Raynaud, R. Rothlisberger, U. Ruth, D. Samyn, J. Schwander, H. Shoji, Siggard M. L. Andersen, J. P. Steffensen, T. Stocker, A. E. Sveinbjörnsdóttir, A. Svensson, M. Takata, J. L. Tison, Th Thorsteinsson, O. Watanabe, F. Wilhelms, and J. W. C. White. High-resolution record of Northern Hemisphere climate extending into the last interglacial period. Nature, 431:147–151, 2004.

[5] P. Billingsley. Statistical inference for Markov processes. Statistical Research Monographs, Vol. II. The University of Chicago Press, Chicago, Ill., 1961.

[6] B. A. Block, I. D. Jonsen, S. J. Jorgensen, A. J. Winship, S. A. Shaffer, S. J. Bograd, E. L. Hazen, D. G. Foley, G. A. Breed, A.-L. Harrison, J. E. Ganong, A. Swithenbank, M. Castleton, H. Dewar, B. R. Mate, G. L. Shillinger, K. M. Schaefer, S. R. Benson, M. J. Weise, R. W. Henry, and D. P. Costa. Tracking apex marine predator movements in a dynamic ocean. Nature, 475(7354):86–90, 2011.

[7] N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J.-P. Bouchaud, and A.-L. Barabási. Dynamics of ranking processes in complex systems. Phys. Rev. Lett., 109(12):128701, 2012.

[8] K. P. Burnham and D. R. Anderson. Model selection and multimodel inference. Springer-Verlag, second edition, 2002. A practical information-theoretic approach.

[9] D. G. Cacuci. Sensitivity and uncertainty analysis. Vol. I. Chapman & Hall/CRC, 2003. Theory.

[10] R. M. Capocelli and L. M. Ricciardi. Diffusion approximation and first passage time problem for a model neuron. Kybernetik, 8:214–223, 1971.

[11] D. Dacunha-Castelle and D. Florens-Zmirou. Estimation of the coefficients of a diffusion from discrete observations. Stochastics, 19(4):263–284, 1986.

[12] W. Dansgaard, S. J. Johnsen, H. B. Clausen, D. Dahl-Jensen, N. S. Gundestrup, C. U. Hammer, C. S. Hvidberg, J. P. Steffensen, A. E. Sveinbjornsdottir, J. Jouzel, and G. Bond. Evidence for general instability of past climate from a 250-kyr ice-core record. Nature, 364(6434):218–220, 1993.

[13] P. D. Ditlevsen, K. K. Andersen, and A. Svensson. The DO-climate events are probably noise induced: statistical investigation of the claimed 1470 years cycle. Clim. Past, 3(1):129–13, 2007.

[14] P. D. Ditlevsen and O. D. Ditlevsen. On the stochastic nature of the rapid climate shifts during the last ice age. J. Climate, 22(2):446–457, 2009.

[15] P. Frantsuzov, M. Kuno, B. Janko, and R. A. Marcus. Universal emission intermittency in quantum dots, nanorods and nanowires. Nat. Phys., 4(5):519–522, 2008.

[16] R. Friedrich, J. Peinke, M. Sahimi, and M. Reza Rahimi Tabar. Approaching complexity by stochastic methods: From biological systems to turbulence. Phys. Rep., 506:87–162, 2011.

[17] C. Gardiner. Stochastic Methods. Springer, 2009. A Handbook for the Natural and Social Sciences.

[18] P. M. Grootes and M. Stuiver. Oxygen 18/16 variability in Greenland snow and ice with 10^{-3}- to 10^5-year time resolution. J. Geophys. Res. Oceans, 102(C12):26455–26470, 1997.

[19] D. Higham, X. Mao, M. Roj, Q. Song, and G. Yin. Mean exit times and the multilevel monte carlo method. SIAM/ASA J. Uncertainty Quantification, 1(1):2–18, 2013.

[20] W. Horsthemke and R. Lefever. Noise-induced transitions. Springer, 1984.

[21] N. E. Humphries, N. Queiroz, J. R. M. Dyer, N. G. Pade, M. K. Musyl, K. M. Schaefer, D. W. Fuller, J. M. Brunnschweiler, T. K. Doyle, J. D.R. Houghton, G. C. Hays, C. S. Jones, L. R. Noble, V. J. Wearmouth, E. J. Southall, and D. W. Sims. Environmental context explains Lévy and brownian movement patterns of marine predators. Nature, 465(7301):1066–1069, 2010.

[22] T. John, R. Stannarius, and U. Behn. On-off intermittency in stochastically driven electrohydrodynamic convection in nematics. Phys. Rev. Lett., 83(4):749–752, 1999.

[23] M. Kac and J. Logan. Fluctuations. In E. W. Montroll and J. L. Lebowitz, editors, Fluctuation Phenomena, chapter 1, pages 1–60. Elsevier, 1979.

[24] I. Karatzas and S. E. Shreve. Brownian motion and stochastic calculus. Springer, second edition, 1991.

[25] S. Krumscheid, G. A. Pavliotis, and S. Kalliadasis. Semi-

parametric drift and diffusion estimation for multiscale diffusions. Multiscale Model. Simul., 11(2):442–473, 2013.

[26] Y. A. Kutoyants. Statistical inference for ergodic diffusion processes. Springer, 2004.

[27] F. Kwasniok. Analysis and modelling of glacial climate transitions using simple dynamical systems. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 371(1991):20110472, 22, 2013.

[28] F. Kwasniok and G. Lohmann. Deriving dynamical models from paleoclimatic records: Application to glacial millennial-scale climate variability. Phys. Rev. E, 80:066104–1, 2009.

[29] J. M. López, M. Pradas, and A. Hernández-Machado. Activity statistics, avalanche kinetics, and velocity correlations in surface growth. Phys. Rev. E, 82:031127, 2010.

[30] B. K. Øksendal. Stochastic differential equations. Springer, 2003. An introduction with applications.

[31] G. A. Pavliotis and A. M. Stuart. Multiscale Methods: Averaging and Homogenization. Springer, first edition, 2007.

[32] G. Pesce, A. McDaniel, S. Hottovy, J. Wehr, and G. Volpe. Stratonovich-to-Itô transition in noisy systems with multiplicative feedback. Nat. Commun., 4:2733, 2013.

[33] M. Pradas, J. M. López, and A. Hernández-Machado. Avalanche dynamics in fluid imbibition near the depinning transition. Phys. Rev. E, 80(5):050101, 2009.

[34] M. Pradas, G. A. Pavliotis, S. Kalliadasis, D. T. Papageorgiou, and D. Tseluiko. Additive noise effects in active nonlinear spatially extended systems. Eur. J. Appl. Math., 23(5):563–591, 2012.

[35] M. Pradas, D. Tseluiko, S. Kalliadasis, D. T. Papageorgiou, and G. A. Pavliotis. Noise induced state transitions, intermittency, and universality in the noisy Kuramoto-Sivashinksy equation. Phys. Rev. Lett., 106(6):060602, 2011.

[36] B. L. S. Prakasa Rao. Statistical inference for diffusion type processes, volume 8 of Kendall's Library of Statistics. Arnold, London, 1999.

[37] S. Rahmstorf. Timing of abrupt climate change: A precise clock. Geophys. Res. Lett., 30(10):1510, 2003.

[38] A. M. Reynolds and C. J. Rhodes. The Lévy flight paradigm: random search patterns and mechanisms. Ecology, 90(4):877–887, 2009.

[39] M. Schmuck, M. Pradas, S. Kalliadasis, and G. A. Pavliotis. New stochastic mode reduction strategy for dissipative systems. Phys. Rev. Lett., 110(24):244101, 2013.

[40] M. Schmuck, M. Pradas, G.A. Pavliotis, and S. Kalliadasis. A new mode reduction strategy for the generalized kuramotosivashinsky equation. IMA J. Appl. Math., 2013. doi: 10.1093/imamat/hxt041.

[41] M. Schulz. On the 1470-year pacing of Dansgaard-Oeschger warm events. Paleoceanography, 17(2):4–1–4–9, 2002.

[42] W. L. Shew, H.Yang, T.Petermann, R. Roy, and D. Plenz. Neuronal avalanches imply maximum dynamic range in cortical networks at criticality. J. Neurosci., 29(49):15595–15600, 2009.

[43] D. W. Sims, E. J. Southall, N. E. Humphries, G. C. Hays, C. J. A. Bradshaw, J. W. Pitchford, A. James, M. Z. Ahmed, A. S. Brierley, M. A. Hindell, D. Morritt, M. K. Musyl, D. Righton, E. L. C. Shepard, V. J. Wearmouth, R. P. Wilson, M. J. Witt, and J. D. Metcalfe. Scaling laws of marine predator search behaviour. Nature, 451(7182):1098–1102, 2008.

[44] P. Sura and J. Barsugli. A note on estimating drift and diffusion parameters from time series. Phys. Lett. A, 305(5):304–311, 2002.

[45] Cheng Yong Tang and Song Xi Chen. Parameter estimation and bias correction for diffusion processes. Journal of Econometrics, 149(1):65 – 81, 2009.

[46] S. R. Thorrold, P. Afonso, J. Fontes, C. D. Braun, R. S. Santos, G. B. Skomal, and M. L. Berumen. Extreme diving behaviour in devil rays links surface waters and the deep ocean. Nat. Commun., 5:4274, 2014.

[47] I. Y. Wong, M. L. Gardel, D. R. Reichman, Eric R. Weeks, M. T. Valentine, A. R. Bausch, and D. A. Weitz. Anomalous diffusion probes microstructure dynamics of entangled F-actin networks. Phys. Rev. Lett., 92(17):178101, 2004.

ACKNOWLEDGEMENTS

This work is supported by the Engineering and Physical Sciences Research Council of the UK through Grant No. EP/H034587 and 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 I.

AUTHOR CONTRIBUTIONS

All authors contributed to all aspects of this work.