Estimation of drift and diffusion functions from unevenly sampled time-series data

William Davis and Bruce Buffett

Phys. Rev. E **106**, 014140 — Published 27 July 2022

DOI: 10.1103/PhysRevE.106.014140

Estimation of Drift and Diffusion Functions from Unevenly Sampled Time-Series Data

William Davis* and Bruce Buffett
Department of Earth and Planetary Science, University of California, Berkeley
(Dated: July 5, 2022)

Complex systems can often be modelled as stochastic processes. However, physical observations of such systems are often irregularly spaced in time, leading to difficulties in estimating appropriate models from data. Here we present extensions of two methods for estimating drift and diffusion functions from irregularly sampled time-series data. Our methods are flexible and applicable to a variety of stochastic systems, including non-Markov processes or systems contaminated with measurement noise. To demonstrate applicability, we use this approach to analyse an irregularly sampled paleoclimatological isotope record, giving insights into underlying physical processes.

I. INTRODUCTION

The time-dependent behavior of complex systems consisting of a large number of subsystems can often be described by low-dimensional order parameter equations [1]. In many cases, a separation between slow adjustments and fast fluctuations allows for a description of continuous observables X of such systems with a Langevin-type equation

$$\frac{d}{dt}X(t) = f(X,t) + g(X,t)\Gamma(t) \tag{1}$$

where $\Gamma(t)$ denotes the stochastic force, with $\langle\Gamma(t)\rangle = 0$ and $\langle\Gamma(t)\Gamma(t')\rangle = \delta(t-t')$ [2]. The same information is expressed in the Fokker-Planck equation,

$$\frac{\partial}{\partial t}p(x,t|x',t') = \Bigg[-\frac{\partial}{\partial x}D^{(1)}(x,t) + \frac{\partial^2}{\partial x^2}D^{(2)}(x,t)\Bigg]p(x,t|x',t') \tag{2}$$

which contains the Kramers-Moyal (KM) coefficients

$$D^{(n)}(x,t) = \lim_{\tau\to 0}\frac{1}{n!\tau}\int_{-\infty}^{\infty}\left[x'-x\right]^n p(x',t+\tau|x,t)\;dx', \tag{3}$$

where x and x' denote values that can be taken by X, and $p(\circ|\circ)$ is the transition probability. Here, the first two coefficients are the drift and diffusion, respectively, connecting to (1) under the Itô interpretation, with $f(x,t) = D^{(1)}(x,t)$ and $g(x,t) = \sqrt{2D^{(2)}(x,t)}$.

It has been shown that it is possible to estimate the forms of such processes directly from regularly sampled time-series data using a technique called "direct estimation" [3, 4]. This approach has been applied to various fields of science [5].

There are two main difficulties associated with applying this approach to "real-world" time-series data. The first occurs when observations are contaminated by another undesirable signal, or measurement noise. In this case, Böttcher *et al.* [6] introduced a method to parametrically estimate drift and diffusion functions as well as the amplitude of the measurement noise, an approach has been expanded in subsequent studies [7–9].

The other difficulty involves the discrete sampling of the time-series data. For low sampling frequencies, is can be difficult to perform or infer the limit $\tau \to 0$ required for direct estimation. In this case, Honisch and Friedrich [10] proposed a finite τ optimisation method that correctly recovers drift and diffusion functions even at large sampling. However a related impediment is the presence of irregular sampling. In this case, there is no obvious way to calculate averages in (3). This is commonly encountered in geoscientific measurements [e.g. 11, 12], but also is encountered in turbulence measurements [13–15], astrophysical observations [16–19], and biological systems [20]. Interpolation is sometimes used to side-step these difficulties, however this can introduce a significant and hard-to-quantify bias [12, 21–23]. This motivates a method for estimating drift and diffusion functions directly from unaltered time-series data.

In the next section we review current estimation techniques, and propose two extensions for irregular sampling. Section III shows numerical examples where we demonstrate the functionality of our new methods. In Section IV we apply this framework to an empirical dataset, namely a paleoclimatological isotope record [24]. Summaries are given in Section V, where further applications are proposed.

II. ESTIMATION OF CONDITIONAL MOMENTS

We consider a *stationary* scalar process $X(t)$ that is observed at a set of N increasing points in time, $\{t_1, t_2, \ldots, t_N\}$, with no guarantee of a regular sampling. Observations at these points are denoted as $\{X(t_1), X(t_2), \ldots, X(t_N)\}$. The finite-time KM coefficients of $X(t)$ are defined as [10]

* williamjsdavis@berkeley.edu

$$D_\tau^{(n)}(x) = \frac{1}{n!\tau} M^{(n)}(x,\tau), \tag{4}$$

which are calculated using the finite-time conditional moments

$$M^{(n)}(x,\tau) = \int_{-\infty}^{\infty} [x' - x]^n p(x', t+\tau | x, t)\ dx'. \tag{5}$$

The task is to make an estimate of these moments from data $X(t)$. These moments will subsequently be used as finite-time KM coefficients in an appropriate method in order to estimate drift and diffusion functions of the underlying process.

Conditional moment estimates are denoted as $\hat{M}^{(n)}(x_i, \tau_j)$, and are evaluated at a set of evaluation points in $x_i \in \{x_1, x_2, \ldots, x_{\max}\}$, and $\tau_j \in \{\tau_1, \tau_2, \ldots, \tau_{\max}\}$.

A. Histogram Based Regression

The simplest way of estimating conditional moments is by means of regressogram, [e.g. 25], also known as histogram based regression (HBR). This estimator can be written as, [e.g. 26],

$$\hat{M}^{(n)}(x_i, \tau_j) = \frac{\sum_{k=1}^{N} I\big(X(t_k) \in B^{(x)}(x_i)\big) \big[X(t_k + \tau_j) - X(t_k)\big]^n}{\sum_{k=1}^{T} I\big(X(t_k) \in B^{(x)}(x_i)\big)}, \tag{6}$$

where $I(\circ)$ is the indicator function, and binning is indicated with the half closed interval $B^{(x)}(x_i) := [x_i - \frac{1}{2}b_x, x_i + \frac{1}{2}b_x)$, where b_x is the width of the bin.

B. Histogram-Time Based Regression

One simple way to extend HBR to account for uneven time-sampling is to average over all pairs of increasing times, and also bin data by time-step. We shall refer to this method as histogram-time based regression (HTBR). The estimator for conditional moments can be written as

$$\hat{M}^{(n)}(x_i, \tau_j) = \frac{\sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \overbrace{I\big(X(t_k) \in B^{(x)}(x_i)\big)}^{x\text{-conditioning}} \overbrace{I\big(\Delta t_{l,k} \in B^{(\tau)}(\tau_j)\big)}^{\tau\text{-conditioning}} \big[X(t_l) - X(t_k)\big]^n}{\sum_{k=1}^{T-1} \sum_{l=k+1}^{T} I\big(X(t_k) \in B^{(x)}(x_i)\big) I\big(\Delta t_{l,k} \in B^{(\tau)}(\tau_j)\big)} \tag{7}$$

where $\Delta t_{l,k} := t_l - t_k (> 0)$, and binning in τ is facilitated with a bounded half closed interval $B^{(\tau)}(\tau_j) := [\max(0, \tau_j - \frac{1}{2}b_\tau), \tau_j + \frac{1}{2}b_\tau)$.

Both HBR and HTBR provide simple methods of estimating moments, however the histogram based nature of both methods results in undesirable properties.

1. Histograms assign the same weight to every point inside each bin, resulting in sharp cut-offs between data across the edge of a bin.

2. The width of the bins sets the resolution length-scale. This length-scale dependence is not explicit, it is indirectly determined by the number and range of bins.

C. Kernel Based Regression

To address the deficiencies of the histogram based approach, Lamouroux and Lehnertz [26] introduced kernel based regression (KBR) method. For this, each estimate at x is assigned an estimate by averaging over all observations weighted by the distance of the observation $X(t)$ to x. Moments are then estimated with

$$\hat{M}^{(n)}(x_i,\tau_j) = \frac{\sum_{k=1}^{N} K_h(x_i - X(t_k))[X(t_k+\tau_j) - X(t_k)]^n}{\sum_{k=1}^{T} K_h(x_i - X(t_k))} \tag{8}$$

where $K_h(\circ) = K(\circ/h)/h$ is a scaled kernel, h is the bandwidth, and $K(\circ)$ is the kernel function. Here we use the Epanechnikov kernel [27]

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{if } x^2 < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

for its computationally desirable properties [28].

Kernel-based methods have a number of advantages over histogram based approached, including a higher convergence rate in the limit of a large number of data points [28, 29]. The introduction of a bandwidth gives an explicit indication of the length scale of averaging, although there is no optimal bandwidth. However, as points are indexed at set time-shifts τ_j in the future, this method is unsuitable for unevenly spaced data.

D. Kernel-Time Based Regression

To extend KBR to unevenly spaced data, kernel density estimation is applied to the τ component as well as the x component. We shall refer to this method as kernel-time based regression (KTBR). To enable this, bivariate kernel density estimation is employed

$$\hat{M}^{(n)}(x_i,\tau_j) = \frac{\sum_{k=1}^{T-1}\sum_{l=k+1}^{T} K_h^{(2)}(x_i - X(t_k), \tau_j - \Delta t_{l,k})\big[X(t_l) - X(t_k)\big]^n}{\sum_{k=1}^{T-1}\sum_{l=k+1}^{T} K_h^{(2)}(x_i - X(t_k), \tau_j - \Delta t_{l,k})} \tag{10}$$

where $K_h^{(2)}(\circ,\circ)$ is a bandwidth scaled, Euclidian distance 2D kernel

$$K_h^{(2)}(x,\tau) = \frac{C}{h_x h_\tau} K\left(\left((x/h_x)^2 + (\tau/h_\tau)^2\right)^{\frac{1}{2}}\right) \tag{11}$$

where h_x and h_τ and the bandwidths in x and τ, respectively [30]. The prefactor C is defined such that the kernel integrates to unity. We use the Epanechnikov kernel (9), therefore $C = 8/3\pi$.

As the domain in τ only has positive support, kernel estimations at $\tau < h_\tau$ can be biased. To account for this, we use a boundary correction method [31] that replaces the application of kernel (11) inside (10), with

$$K_h^{(2)}(x_i - X(t_k), \tau_j - \Delta t_{l,k}) \to \left[K_h^{(2)}(x_i - X(t_k), \tau_j - \Delta t_{l,k}) + K_h^{(2)}(x_i - X(t_k), \tau_j + \Delta t_{l,k})\right]. \tag{12}$$

III. NUMERICAL EXAMPLES

To validate the presented methods, we test them on a set of three synthetic data-sets.

A. Ornstein-Uhlenbeck process

First we examine an Ornstein-Uhlenbeck process given by the drift and diffusion functions

$$D^{(1)}(x) = -x, \tag{13a}$$
$$D^{(2)}(x) = 1. \tag{13b}$$

We consider a discrete time-series sampling of $X(t)$ consisting of 10^7 points with irregular time sampling, $\Delta t \sim \mathcal{N}(5\times10^{-3}, 3.2\times10^{-7})$. The solution is integrated [32] with an internal time-step of $\delta t \le 10^{-4}$, to ensure numerical accuracy.

To estimate the conditional moments of this data, we use three separate methods. First, the moments are estimated using HTBR (6). Sampling in x is performed by 11 evenly spaced bins in the range $[-2, 2]$. Sampling in τ is performed by a single bin, $[0, 0.01]$. Here τ is small enough that the drift and diffusion functions can be directly estimated from the moments

$$\hat{D}^{(n)}(x) \approx \frac{1}{n!\tau}\hat{M}^{(n)}(x,\tau). \tag{14}$$

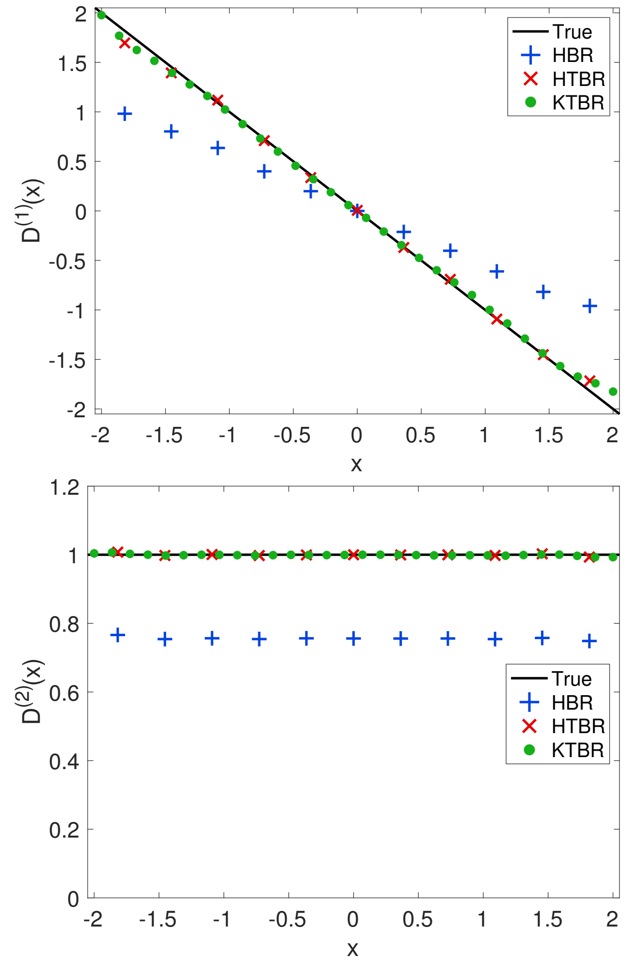

FIG. 1. Results for an Ornstein-Uhlenbeck process. Estimated functions $D^{(1)}(x)$ and $D^{(2)}(x)$ are shown in the top and bottom plots, respectively. Estimates from HBR are from interpolated data.

Second, the moments are estimated using KTBR (10). Evaluation points in x are 30 evenly spaced points in $[-2, 2]$, with a bandwidth of $h_x = 0.3$. Sampling in τ is performed with a single evaluation point at $\tau = 5\times10^{-3}$, with a bandwidth of $h_\tau = 2.5\times10^{-3}$. As with HTBR, the direct estimation method (14) is utilized. Finally, to compare with the two previous methods, naive resampling is performed on the time-series data. The data $X(t)$ is linearly interpolated to a regular sampling of $\Delta t = 5\times10^{-3}$, and then direct estimation is applied with the same bin sampling as the HTBR estimate. The drift and diffusion functions are shown in Fig. 1. In this example—and all the following examples—KBR performed similarly to HBR except with finer resolution, and hence will not shown for conciseness.

We find that the estimates of drift and diffusion functions are in good accordance with the true values for both HTBR and KTBR. These functions are systematically underestimated when using HBR with interpolated time-sampling.

B. Multiplicative process with measurement noise

Next we examine a multiplicative process with measurement noise. The drift and diffusion functions are set as

$$D^{(1)}(x) = -x, \tag{15a}$$
$$D^{(2)}(x) = 1 + x^2. \tag{15b}$$

Irregularly sampled data $X(t)$ is produced similarly to example III A, however we also add δ-correlated measurement noise

$$Y(t) = X(t) + \sigma\zeta(t), \tag{16}$$

where σ denotes the amplitude of the measurement noise, and $\zeta \sim \mathcal{N}(0, 1)$. We seek to estimate coefficients of parameterised drift and diffusion functions

$$\hat{D}^{(1)}(x) = p_1 + p_2 x, \tag{17a}$$
$$\hat{D}^{(2)}(x) = p_3 + p_4 x + p_5 x^2, \tag{17b}$$

using the method of Lind *et al.* [7]. The time-series $Y(t)$ is used to estimate noisy moments, $\hat{M}^{(n)}(y, \tau)$. These moments are separated with linear regression

$$\hat{M}^{(1)}(y_i, \tau_j) \approx \hat{m}_1(y_i)\tau_j + \hat{\gamma}_1(y_i), \tag{18a}$$
$$\hat{M}^{(2)}(y_i, \tau_j) \approx \hat{m}_2(y_i)\tau_j + \hat{\gamma}_2(y_i) + \sigma^2, \tag{18b}$$

along with uncertainties $\sigma^2_{\hat{m}_1}(y_i), \sigma^2_{\hat{\gamma}_1}(y_i)$, etc... These estimates are compared with theoretical values of $m_1(y), \gamma_1, m_2(y)$, and γ_2, which depend solely on parameters $p_1, \ldots, p_5$, and σ, see Lind *et al.* [7] for more details. The parameters vary the fit function

$$F = \sum_{i=1}^{8} \left\{ \frac{[\hat{m}_1(y_i) - m_1(y_i)]^2}{\sigma^2_{\hat{m}_1}(y_i)} + \frac{[\hat{\gamma}_1(y_i) - \gamma_1(y_i)]^2}{\sigma^2_{\hat{\gamma}_1}(y_i)} + \frac{[\hat{m}_2(y_i) - m_2(y_i)]^2}{\sigma^2_{\hat{m}_2}(y_i)} + \frac{[\hat{\gamma}_2(y_i) - \gamma_2(y_i) - \sigma^2]^2}{\sigma^2_{\hat{\gamma}_2}(y_i)} \right\}, \tag{19}$$

which is minimised using simulated annealing [33].

For HTBR, sampling in y is performed with 50 equally spaced bins in the range $[-6, 6]$. Sampling in τ is performed by 8 equally spaced bins with centers from $\tau_1 = 5\times10^{-3}$ to $\tau_8 = 4\times10^{-2}$, with bin-widths $b_\tau = 5\times10^{-3}$. For KTBR, evaluation points in x are 50 equally spaced points in the range $[-6, 6]$, with $h_x = 0.18$. Sampling in time is performed with 8 equally spaced points from $\tau_1 = 5\times10^{-3}$ to $\tau_8 = 4\times10^{-2}$, with $h_\tau = 2.5\times10^{-3}$. Finally, the data $Y(t)$ is also linearly interpolated to a regular sampling of $\Delta t = 5\times10^{-3}$ and then processed in the same way as the HTBR example. The optimised parameters are shown in Table I.

The parameters of the drift and diffusion functions are very close to the true values for both HTBR and KTBR. For HBR with interpolated time-sampling, while some elements are estimated well, the absolute gradient of the drift, the constant diffusion term, and the quadratic term are all overestimated. Finally the measurement noise amplitude σ is underestimated.

C. Bistable system with correlated noise

Finally we examine a bistable process $X(t)$ driven by correlated noise $\eta(t)$ [34]. This system is defined as

$$\frac{d}{dt}X = D^{(1)}(X) + \sqrt{2D^{(2)}(X)}\eta(t), \tag{20a}$$
$$\frac{d}{dt}\eta = -\frac{1}{\theta}\eta + \frac{1}{\theta}\xi(t), \tag{20b}$$

where θ is the correlation time of the noise. The drift and diffusion functions are set as

$$D^{(1)}(x) = x - \frac{1}{2}x^3, \tag{21a}$$
$$D^{(2)}(x) = 1 + \frac{1}{20}\ln\cosh 2x, \tag{21b}$$

and the correlation time is $\theta = 0.01$. An unevenly spaced time-series is produced in the same way as example III A, however only $X(t)$ is observed.

We estimate the drift and diffusion functions using the non-parametric method of [34]. This involves comparing estimates of moments, $\hat{M}^{(n)}(x,\tau)$, with theoretical estimates

$$M^{(n)}(x,\tau) \approx \sum_{i=1}^{3}\lambda_i^{(n)}(x)r_i(\tau,\theta), \tag{22}$$

TABLE I. True and optimised parameter values for a multiplicative process with measurement noise. Parameters are rounded to either 2 significant figures or at least 2 decimal places. The HBR column represents results from interpolated $Y(t)$ data. We note that entering the true parameter values into function (19) with estimates gathered from interpolated HBR result in a value of F two orders of magnitude higher than the optimised minimum.

Parameter	True	HTBR	KTBR	HBR
p_1	0	-0.0050	-0.0040	-0.014
p_2	-1	-0.99	-1.00	-1.48
p_3	1	0.99	1.00	1.62
p_4	0	0.0062	0.013	0.0020
p_5	1	0.97	0.98	1.11
σ	1	1.00	1.00	0.76

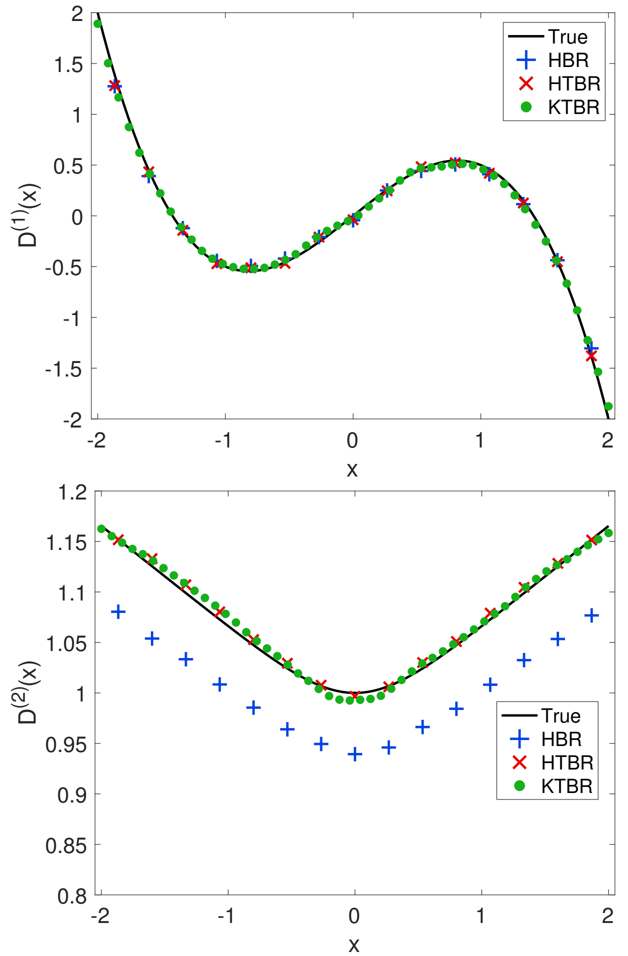

FIG. 2. Results for a bistable system with correlated noise. As Fig. 1.

where functions r_i are prescribed basis functions and $\lambda_i^{(n)}(x)$ are the corresponding coefficients. Coefficients are found through least squares, and then $\lambda_1^{(n)}(x)$ are directly related to estimates of the drift and diffusion functions at points in x. For a detailed description of the method, see Lehle and Peinke [34].

For HTBR, sampling in x is performed by 16 equally spaced bins in the range $[-2, 2]$. Sampling in τ is performed by 30 spaced bins with from $\tau_1 = 5 \times 10^{-3}$ to $\tau_{30} = 1.5 \times 10^{-1}$, with bin-widths $b_\tau = 5 \times 10^{-3}$. For KTBR, evaluation points in x are 50 equally spaced points in the range $[-2, 2]$, with $h_x = 0.24$. Sampling in time is performed with 30 equally spaced points from $\tau_1 = 5 \times 10^{-3}$ to $\tau_{30} = 1.5 \times 10^{-1}$, with $h_\tau = 2.5 \times 10^{-3}$. Finally, the data $X(t)$ is also linearly interpolated to a regular sampling of $\Delta t = 5 \times 10^{-3}$ and then processed in the same way as the HTBR example. For simplicity, we assume that the correlation time θ has been accurately estimated a priori [12, 18]. For all methods, the mean absolute error between estimated moments $\hat{M}^{(n)}(x,\tau)$ and fitted moments (22) is on the order of 10^{-5}. The drift and diffusion functions are shown in Fig. 2.

The estimates of the drift and diffusion functions compare well with the true values for both HTBR and KTBR. For the interpolated HBR the drift function is reproduced well, whilst the diffusion function is systematically underestimated.

IV. APPLICATION TO PALEOCLIMATOLOGICAL DATA

Paleoclimate proxies preserve a record of Earth's climate variability. This variability is commonly studied through carbon and oxygen isotopes records from benthic foraminifera [24, 35]. Of particular interest are large and rapid negative excursions in carbon isotope ratios, δ^{13}C, throughout the Cenozoic [36–40]. These excursions have been interpreted as "hyperthermal" warming events, and are speculated to be linked to the release of isotopically depleted organic carbon from permafrost or methane clathrates [41–43]. Such records offer insights to Earth's climate response to hyperthermal events, and provide an analogue to modern anthropogenic forcing [44–47]. Recently Arnscheidt and Rothman [48] suggested that the time-variability of these records can be modelled as stochastic processes, invoking a single-variable correlated additive-multiplicative (CAM) process

$$\frac{d}{dt}X = -\frac{1}{\tau_{\text{eff}}}X + v\,(X-c)\,\Gamma(t), \tag{23}$$

where τ_{eff}, v, and c are constants and $\Gamma(t)$ is white noise [49–53]. A non-parametric verification of this CAM hypothesis has been unreachable with previous estimation methods, as the δ^{13}C record is unevenly sampled in time. In this section, we apply KTBR to a section of this unevenly sampled paleoclimate record.

We choose a stationary section of the record, from 53 Ma to 46 Ma, containing a series of representative excursions but excluding the anomalous Paleocene–Eocene Thermal Maximum [48, 54]. The sampling in this timespan is approximately log-normally distributed, with $\log_{10}\Delta t \sim \mathcal{N}(-2.7, 0.2)$. To calculate moments, evaluation points in x are 50 equally spaced points in the range $[-0.8, 0.5]$, with $h_x = 0.4$. Sampling in time is performed with 30 equally spaced points from $\tau_1 = 3.5$ kyr to $\tau_{30} = 116$ kyr, with $h_\tau = 5$ kyr. The higher order moments in $M^{(4)}(x,\tau) \simeq 3\left(M^{(2)}(x,\tau)\right)^2$ are evaluated using (10) and are comparable, showing a small error of $\sim 5\times10^{-3}$, validating the continuity of the record [55, 56]. To estimate the drift and diffusion functions from these moments, we use the approach of Lehle and Peinke [34], while the correlation time is estimated through a grid search, $\theta \approx 0.4$ kyr. The moments are fit well, with an absolute error between estimated moments $\hat{M}^{(n)}(x,\tau)$ and fitted moments (22) on the order of 10^{-4}. The estimated drift and diffusion functions are shown in Fig. 4.

The drift function has a strongly linear form, and is well approximated by the CAM model (23) with $\tau_{\text{eff}} = 47$ kyr ($R^2 = 0.98$). For the diffusion function, while a CAM model (23) with the coefficients $v = -3.2$ and $c = -1.2$ falls within the confidence intervals ($R^2 = 0.67$), we cannot reject a likely piecewise diffusion of

$$D^{(2)}(x) = \begin{cases} p_1 + p_2(x - p_3) & \text{if } x \le p_3, \\ p_1 & \text{otherwise,} \end{cases} \tag{24}$$

with best fitting coefficients of $p_1 = 3.30, p_2 = -11.50$, and $p_3 = -0.36$ ($R^2 = 0.99$), although we note that this parameterization is not unique, and only meant to be suggestive.

To demonstrate that this linear drift and piecewise diffusion cannot be rejected by the data, we numerically integrate a sample path with these functions. The time-series and distributions of the original data and SDE simulation are shown in Fig. 3. The SDE matches the skewed distribution of the original record, and also displays characteristic excursions to low δ^{13}C values.

Beyond reproducing observations, the form of the estimated drift and diffusion functions can give insight into physical processes. The drift term indicates an average relaxation timescale of $\tau_{\text{eff}} = 47$ kyr, possibly reflecting the stabilizing feedback of weathering of carbonate and silicate rocks [e.g. 58]. The piecewise nature of the diffusion suggests a "tipping-point" beyond which fluctuations are amplified, indicating an imbalance in typical weathering feedback mechanisms [59–61]. Further work should investigate whether this behavior is reflected in related oxygen isotope records, as well as other epochs in the Cenozoic.

V. DISCUSSION AND CONCLUSION

We present two methods to estimate conditional moments from irregularly spaced time-series. These moments are used alongside parametric and non-parametric methods to facilitate the accurate estimation of drift and diffusion functions of stochastic differential equations. We demonstrate this for three numerical examples, in a number of settings. Even in the presence of measurement noise or non-Markovian processes, both HTBR and KTBR are able to produce moments that result in accurate estimates of the original drift and diffusion functions. Additionally, KTBR is applied to a series of irregularly spaced paleoclimatological measurements. The inferred model is able to produce similar time-dependent behavior and statistics, revealing underlying dynamics.

This study also illustrates the dangers of interpolation. While example III A shows that interpolation results in an absolute underestimate in the magnitude of estimated drift and diffusion functions, example III B shows the opposite bias (with an underestimated measurement noise amplitude). Interpolation in example III C has little effect on the estimated drift function, but not the diffusion

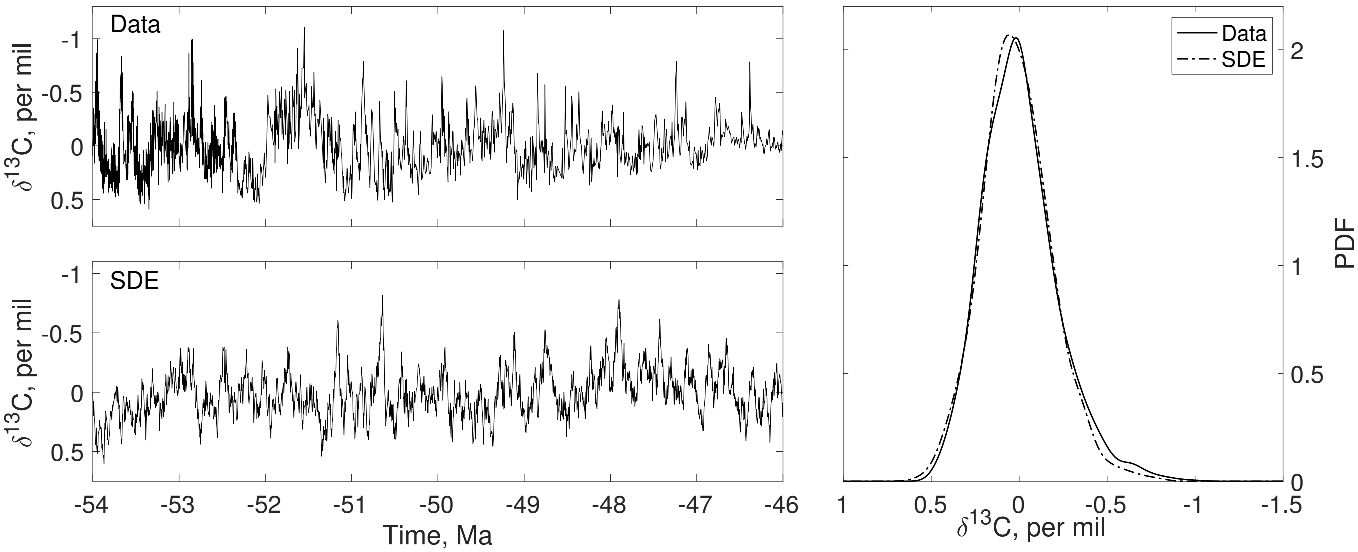

FIG. 3. Climate variations in the Early Eocene, recorded in benthic foraminiferal δ^{13}C data [24]. A running mean of 1-Ma has been subtracted to remove longer-scale climate effects. Time-series data and a simulated trajectory are shown in the top and bottom plots, respectively. Histograms are shown in the right plot. By convention, axes for δ^{13}C are reflected.

function. These smaller errors average out for the drift function, as is the case with weak measurement noise [62]. Overall, the bias may be small because longer time-scale information is included in the inversion, or the interpolation bias may be masked by the non-Markovian nature of the process.

In addition to being applicable to a wide class of stochastic systems, these methods could allow for the handling of other non-ideal sampling conditions. Data with inconvenient gaps, for example, can be approached by this outlook when framed as irregularly sampled processes. This method is also capable of estimating higher-order moments ($n > 2$ in (7) and (10)), which are useful for analysis of jump-diffusion processes [63]. On the effect of number of data points on the robustness of the estimated drift and noise functions, as HTBR and KTBR are inherently frequency based calculations we expect them to perform similarly to previous methods [26, 64, 65]. The methods here are demonstrated in one dimension, however extensions to higher dimensions is straightforward.

In the broader picture for stochastic process estimation, the methods presented here extend time-shift conditioning from index-based to histogram and kernel based methods. This reflects similar work regarding sample autocorrelation function estimators [12, 18, 66]. We note that it is not strictly required to match similar conditioning on x and τ. In theory hybrid methods could be used, for example, kernel conditioning in x combined with histogram conditioning in τ, however it is not clear if such an approach would have significant advantages.

ACKNOWLEDGMENTS

We thank Drew Gorin for insightful discussions, and two anonymous reviewers for their helpful comments and suggestions. This work is supported by the National Science Foundation (EAR-1644644).

[1] H. Haken, *Synergetics: Intoduction and advanced topics* (Spinger-Verlag, 2004).
[2] H. Risken, in *The Fokker-Planck Equation* (Springer, 1996) pp. 63–95.
[3] S. Siegert, R. Friedrich, and J. Peinke, Physics Letters A **243**, 275 (1998).
[4] J. Gottschall and J. Peinke, New Journal of Physics **10**, 083034 (2008).
[5] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar, Physics Reports **506**, 87 (2011).
[6] F. Böttcher, J. Peinke, D. Kleinhans, R. Friedrich, P. G. Lind, and M. Haase, Physical Review Letters **97**, 090603 (2006).
[7] P. G. Lind, M. Haase, F. Böttcher, J. Peinke, D. Kleinhans, and R. Friedrich, Physical Review E **81**, 041125 (2010).
[8] B. Lehle, Physical Review E **83**, 021113 (2011).
[9] T. Scholz, F. Raischel, V. V. Lopes, B. Lehle, M. Wächter, J. Peinke, and P. G. Lind, Physics Letters A **381**, 194 (2017).
[10] C. Honisch and R. Friedrich, Physical Review E **83**, 066701 (2011).
[11] M. Schulz and K. Stattegger, Computers & Geosciences **23**, 929 (1997).
[12] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths, Nonlinear Processes in Geophysics **18**, 389 (2011).

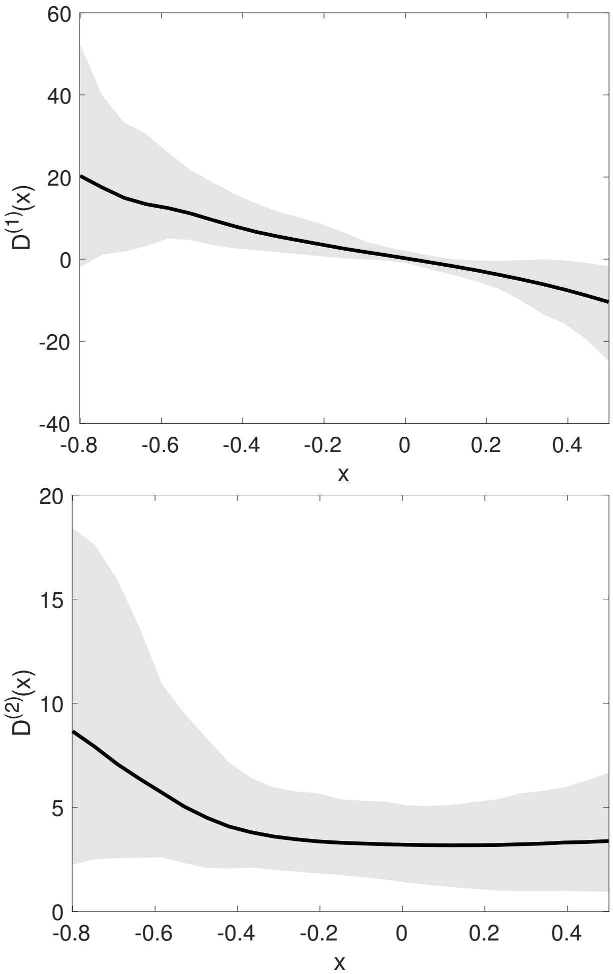

FIG. 4. Results for early Eocene δ^{13}C record. Estimated drift and diffusion functions $D^{(1)}(x)$ and $D^{(2)}(x)$ are shown in the top and bottom plots, respectively. Best estimates are plotted as black lines, and bootstrapped 95% confidence intervals are shown as grey regions [57].

[13] C. Tropea, Measurement Science and Technology **6**, 605 (1995).

[14] P. Broersen, S. De Waele, and R. Bos, in *Proceedings of the 10th International Symposium on Applications of Laser Techniques to Fluid Dynamics, Lisbon, Portugal* (2000).

[15] W. Harteveld, R. Mudde, and H. Van den Akker, Chemical engineering science **60**, 6160 (2005).

[16] J. D. Scargle, The Astrophysical Journal Supplement Series **45**, 1 (1981).

[17] J. D. Scargle, The Astrophysical Journal **263**, 835 (1982).

[18] R. Edelson and J. Krolik, The Astrophysical Journal **333**, 646 (1988).

[19] J. D. Scargle, The Astrophysical Journal **343**, 874 (1989).

[20] A. W.-C. Liew, J. Xian, S. Wu, D. Smith, and H. Yan, BMC bioinformatics **8**, 1 (2007).

[21] M. Scholes and J. Williams, Journal of financial economics **5**, 309 (1977).

[22] T. Hayashi and N. Yoshida, Bernoulli **11**, 359 (2005).

[23] A. Eckner, Preprint. Available at: `http://www.eckner.com/papers/unevenly_spaced_time_series_analysis.pdf` (2014).

[24] T. Westerhold, N. Marwan, A. J. Drury, D. Liebrand, C. Agnini, E. Anagnostou, J. S. Barnet, S. M. Bohaty, D. De Vleeschouwer, F. Florindo, *et al.*, Science **369**, 1383 (2020).

[25] J. W. Tukey, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, Vol. 4 (University of California Press, 1961) pp. 681–695.

[26] D. Lamouroux and K. Lehnertz, Physics Letters A **373**, 3507 (2009).

[27] V. A. Epanechnikov, Theory of Probability & Its Applications **14**, 153 (1969).

[28] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and semiparametric models*, Vol. 1 (Springer, 2004).

[29] L. R. Gorjão and F. Meirinhos, Journal of Open Source Software **4**, 1693 (2019).

[30] M. P. Wand and M. C. Jones, Journal of the American Statistical Association **88**, 520 (1993).

[31] M. C. Jones, Statistics and computing **3**, 135 (1993).

[32] G. Mil'shtejn, Theory of Probability &; Its Applications **19**, 557 (1975).

[33] J. Carvalho, F. Raischel, M. Haase, and P. Lind, in *Journal of Physics: Conference Series*, Vol. 285 (IOP Publishing, 2011) p. 012007.

[34] B. Lehle and J. Peinke, Physical Review E **97**, 012113 (2018).

[35] J. Zachos, M. Pagani, L. Sloan, E. Thomas, and K. Billups, Science **292**, 686 (2001).

[36] B. S. Cramer, J. D. Wright, D. V. Kent, and M.-P. Aubry, Paleoceanography **18** (2003).

[37] L. J. Lourens, A. Sluijs, D. Kroon, J. C. Zachos, E. Thomas, U. Röhl, J. Bowles, and I. Raffi, Nature **435**, 1083 (2005).

[38] M. J. Nicolo, G. R. Dickens, C. J. Hollis, and J. C. Zachos, Geology **35**, 699 (2007).

[39] P. F. Sexton, R. D. Norris, P. A. Wilson, H. Pälike, T. Westerhold, U. Röhl, C. T. Bolton, and S. Gibbs, Nature **471**, 349 (2011).

[40] V. Lauretano, K. Littler, M. Polling, J. C. Zachos, and L. J. Lourens, Climate of the Past **11**, 1313 (2015).

[41] G. R. Dickens, Earth and Planetary Science Letters **213**, 169 (2003).

[42] D. J. Lunt, A. Ridgwell, A. Sluijs, J. Zachos, S. Hunter, and A. Haywood, Nature Geoscience **4**, 775 (2011).

[43] R. M. DeConto, S. Galeotti, M. Pagani, D. Tracy, K. Schaefer, T. Zhang, D. Pollard, and D. J. Beerling, Nature **484**, 87 (2012).

[44] G. J. Bowen, T. J. Bralower, M. L. Delaney, G. R. Dickens, D. C. Kelly, P. L. Koch, L. R. Kump, J. Meng, L. C. Sloan, E. Thomas, *et al.*, Eos, Transactions American Geophysical Union **87**, 165 (2006).

[45] J. C. Zachos, G. R. Dickens, and R. E. Zeebe, Nature **451**, 279 (2008).

[46] T. Dunkley Jones, A. Ridgwell, D. Lunt, M. Maslin, D. Schmidt, and P. Valdes, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **368**, 2395 (2010).

[47] R. E. Zeebe and J. C. Zachos, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **371**, 20120006 (2013).

[48] C. W. Arnscheidt and D. H. Rothman, Science Advances **7**, eabg6864 (2021).

[49] P. Sura and P. D. Sardeshmukh, Journal of Physical Oceanography **38**, 639 (2008).

[50] P. D. Sardeshmukh and P. Sura, Journal of Climate **22**, 1193 (2009).

[51] P. Sura, Atmospheric Research **101**, 1 (2011).

[52] C. Penland and P. D. Sardeshmukh, Chaos: An Interdisciplinary Journal of Nonlinear Science **22**, 023119 (2012).

[53] P. D. Sardeshmukh and C. Penland, Chaos: An Interdisciplinary Journal of Nonlinear Science **25**, 036410 (2015).

[54] F. A. McInerney and S. L. Wing, Annual Review of Earth and Planetary Sciences **39**, 489 (2011).

[55] K. Lehnertz, L. Zabawa, and M. R. R. Tabar, New Journal of Physics **20**, 113043 (2018).

[56] R. Tabar, *Analysis and data-based reconstruction of complex nonlinear dynamical systems*, Vol. 730 (Springer, 2019).

[57] H. R. Kunsch, The Annals of Statistics , 1217 (1989).

[58] J. C. Walker, P. Hays, and J. F. Kasting, Journal of Geophysical Research: Oceans **86**, 9776 (1981).

[59] T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber, Proceedings of the national Academy of Sciences **105**, 1786 (2008).

[60] D. H. Rothman, Science Advances **3**, e1700906 (2017).

[61] D. H. Rothman, Proceedings of the National Academy of Sciences **116**, 14813 (2019).

[62] M. Siefert, A. Kittel, R. Friedrich, and J. Peinke, EPL (Europhysics Letters) **61**, 466 (2003).

[63] M. Anvari, M. Tabar, J. Peinke, and K. Lehnertz, Scientific reports **6**, 1 (2016).

[64] A. M. van Mourik, A. Daffertshofer, and P. J. Beek, Physics Letters A **351**, 13 (2006).

[65] D. Kleinhans and R. Friedrich, in *Wind Energy* (Springer, 2007) pp. 129–133.

[66] P. Babu and P. Stoica, Digital Signal Processing **20**, 359 (2010).